UNIVERSITE PARIS-SORBONNE

ECOLE DOCTORALE V Concepts et Langages

Laboratoire de recherche STIH

T H E S E

pour obtenir le grade de
DOCTEUR DE L'UNIVERSITE PARIS-SORBONNE

Discipline/ Spécialité : Informatique

Présentée et soutenue par :

Milan Stankovic

le **7 décembre 2012**

Convergence entre Web Social et Web Sémantique

Application à l'innovation à l'aide du Web

JURY:

M Jean-Pierre Desclés	Professeur, Université Paris-Sorbonne (directeur)
M Philippe Laublet	MCF, Université Paris-Sorbonne (co-directeur)
M Fabien Gandon	Dr HDR, INRIA Sophia-Antipolis (rapporteur)
Mme Chantal Reynaud	Professeur, Université Paris-Sud XI, Orsay (rapporteur)
M Aldo Gangemi	Professeur, Université Paris XIII (examinateur)
M Alexandre Passant	Chercheur, DERI NUI Galway, Ireland (examinateur)

Resumé

Cette thèse s'inscrit dans le cadre de travaux relatifs au Web Social Sémantique, dans la perspective de la complémentarité et de la coévolution de deux aspects du Web, l'aspect social et sémantique. Le développement du Web au cours de ces dernières années a fait émerger un énorme graphe de données structurées, sémantiques résultant en partie de l'activité des utilisateurs, le LOD. Nous nous intéressons particulièrement à l'utilisation de ce graphe afin de faciliter l'accès à l'information présente sur le Web, et ce de manière utile, informative et enrichissante pour l'utilisateur. Cette problématique est notamment étudiée dans les scénarios de l'innovation sur le Web – pratiques visant à utiliser des technologies du Web pour contribuer à l'émergence de l'innovation. Une spécificité de ce contexte, assez peu abordé dans la littérature existante, est sans doute le besoin d'inciter les découvertes inattendues et fortuites. Au delà de la simple pertinence sollicitée dans toute situation de recherche et de recommandation sur le Web, le contexte d'innovation impose une certaine ouverture d'esprit pour permettre à l'utilisateur d'accéder aux informations inattendues mais néanmoins pertinentes, et permet par la même occasion de s'inspirer et de transposer des idées d'un domaine à l'autre.

Les travaux présentés dans cette thèse ont donc pour objectif d'aider, de manière directe ou indirecte, les acteurs de l'innovation en ligne (e.g., les entreprises qui cherchent à innover, les experts et les porteurs d'idées) de faire des découvertes. Nous abordons chacun de ces défis dans les différentes parties de nos travaux de thèse.

Nous avons d'abord soulevé la question de la recherche d'experts, très importante pour toute activité de l'innovation via le Web. Dans cette partie nous exploitons la richesse du Web Social Sémantique, richesse qui repose notamment sur la diversité de types de liens et de ressources. Celles-ci représentent les traces de l'activité des utilisateurs, reliant un utilisateur à des concepts, e.g des sujets d'intérêt. C'est sur cette richesse que nous avons fondé le développement d'une approche de recherche d'experts très souple, qui peut s'adapter à la nature de la compétence recherchée et à la spécificité de communication au sein d'une communauté de pratique. Cette approche a été implémenté dans un système de recherche d'experts, hy.SemEx, qui permet à la fois de trouver des experts sur le graphe du Web Social Sémantique et d'orienter cette recherche vers des types de traces les plus révélateurs par rapport à une thématique recherchée.

Le deuxième axe de notre travail concerne la découverte de concepts pertinents et inattendus,, notamment utiles pour élargir une recherche d'experts. Dans le cas de la résolution de problèmes en ligne, il est particulièrement intéressant de pouvoir élargir l'ensemble des concepts par lesquelles un problème est initialement décrit, en y incluant par exemple des sujets de compétence divers qui sortent parfois du cadre conceptuel du problème, permettant ainsi de solliciter les compétences en dehors du « domaine » initial. Dans nos travaux nous montrons comment la richesse du graphe Web peut être utilisée pour permettre la découverte d'une telle diversité de concepts. L'outil hyProximity, a été développé pour mettre en œuvre notre approche.

Comme souligné précédemment, la découverte de concepts pertinents, inattendus, et la recherche d'experts permettent de faciliter l'émergence de l'innovation sur le Web. Dans ce contexte, innover signifie très souvent collaborer. La pluridisciplinarité et le travail collaboratif étant des piliers de l'innovation moderne, nous avons abordé la thématique de recommandation de collaborateurs. Dans nos travaux nous montrons comment les liens sociaux et les liens entre concepts dans le graphe Web peuvent être utilisés ensemble, dans le but de trouver de meilleurs collaborateurs pour résoudre un problème donné. Ce travail combine à la fois les approches mentionnées précédemment et hyProximity pour assurer une diversité de candidats collaborateurs. Il s'appuie sur leur complémentarité avec l'utilisateur cherchant à collaborer ainsi que sur leur compatibilité personnelle.

Finalement, nous nous sommes intéressés à la question de la représentation de la connaissance, notamment dans le cadre de l'innovation sur le Web. Cette thèse présente donc également les différentes manières de décrire sémantiquement les acteurs et autres éléments importants dans un processus d'innovation. A cette réflexion sur la représentation de connaissance (les ontologies nécessaires) s'ajoutent les représentations d'outils et applications Web qui sont susceptibles de soutenir l'être humain dans l'innovation. Ces dernières permettraient la réutilisation d'un historique de fonctionnement des applications Web afin de faciliter l'émergence de futures innovations.

Convergence between the Socaial and the Semantic Web

Application to Innovation on the Web

Abstract

This thesis builds upon the work on the Social Semantic Web, a research perspective on the complementarity and coevolution of two aspects of the Web, the social and semantic one. Web development in recent years has given rise to a huge graph of semantically structured data, partly resulting from user activity. We are particularly interested in the use of this graph in order to facilitate access to information found on the Web, in a useful, informative manner. This problem is particularly studied in scenarios related to innovation on the Web - practices to use Web technologies to contribute to the emergence of innovation. A notable specificity of this context, so far little discussed in literature, is the need to encourage serendipity and discovery. Beyond the simple relevance sought in any search and recommendation situation on the Web, the context of innovation requires a certain openness to allow the user to access information relevant yet unexpected, and should also open opportunities to learn and translate ideas from one domain to another.

The work presented in this thesis therefore aims to assist, directly or indirectly, the innovators online (eg, companies seeking to innovate, experts and carriers of ideas) to make discoveries. We address each of these challenges in different parts of the thesis.

We first raised the issue of expert finding, very important for any activity of innovation via the Web. In the corresponding section we explore the richness of the Social Semantic Web, in terms of diversity of types of links and resources useful in expert finding. The Social Semantic Web graph is mostly composed of traces of user activity, connecting a description of a user with concepts, e.g. topics of interest. Our flexible expert search approach is based on the wealth of traces and can adapt to the nature of competence sought and specificity of communication within particular communities of practice. This approach has been implemented in a system, hy.SemEx, which allows both to find experts on the Social Semantic Web graph and guide the search towards the use of the most revealing types of traces with regards to the chosen subject of expertise.

The second focus of our work is the discovery of relevant and unexpected concepts, especially useful to broaden an expert search. In the case of solving problems online, it is particularly interesting to expand the set of concepts by which a problem is initially described by including such diverse topics of expertise that sometimes come out of the conceptual framework of the problem, allowing to solicit skills outside of the problem's initial "field". In our work we show how the richness of the Web graph can be used to enable the discovery of such diverse concepts. A tool, HyProximity, has been developed to implement our approach.

As multidisciplinary and collaborative work are the pillars of modern innovation, we also discussed the topic of recommendation of collaborators as one of our focal points in this thesis. In our work we show how social ties and relationships between concepts in the graph may be used together in order to form the best collaborator teams to solve a given problem. This work partially relies on hyProximity to ensure the diversity of candidates employees. The collaborator recommendations are based on complementarity with the user seeking to collaborate as well as on their personal compatibility, which ensures their ability to work together.

Finally, we are interested in the question of knowledge representation, particularly in the context of innovation on the Web. This thesis therefore also explores different ways to semantically describe important actors and artefacts in the innovation process. We provide ontologies and guidelines needed to describe those elements, with the aim of reusing them in further search and recommendation processes.

Mots-Clés

Web Sémantique
Web Social
Innovation sur le Web
Innovation Ouverte
Plate-forme de résolution de problèmes
Traces utilisateurs
Données Liées

Keywords

Semantic Web
Social Web
Innovation via the Web
Open Innovation
Problem-sovling Platform
User Traces
Linked Data

Remerciements

Si ces trois dernières années de thèse m'ont amené à faire des nombreux sacrifices elles m'ont apporté beaucoup, d'un point de vue aussi bien personnel que scientifique. Ce travail étant achevé, je souhaite remercier un certain nombre de personnes sans qui je n'aurai sans doute pas pu franchir ce cap et qui m'ont supporté dans mes changements d'humeur et dans mes plongées en obscurité dans la réflexion scientifique, en m'excusant par avance auprès de celles et ceux que j'oublie.

Tout d'abord, je tiens à remercier Philippe Laublet pour l'encadrement exceptionnel de cette thèse. C'est grâce à lui que cette thèse n'était pas seulement un projet scientifique, mais elle a donné une impulsion à une métamorphose intellectuelle que j'ai vécue en changeant mes manières de penser, d'analyser, de réagir et d'anticiper. Philippe a su faire un travail pédagogique extraordinaire, qui a porté fruits malgré toute ma résistance au changement – un changement indispensable pour devenir docteur. Philippe a aussi su me rassurer lors de moments durs que j'ai connus face aux différentes incertitudes, à la fois scientifiques et existentielles, en rappelant les objectifs fondamentaux. D'être dirigé sans se sentir contrôlé, c'était un luxe que Philippe m'a offert et pour lequel je lui serai éternellement reconnaissant.

Merci également à Jean-Pierre Desclés d'avoir permis cette thèse en m'ayant accueilli dans son équipe et merci pour nos échanges enrichissants qui m'ont ouvert à de nouvelles perspectives. Merci à Alexandre Passant qui a fourni un soutien énorme à mes travaux, à la fois par l'exemple de son succès dans la communauté scientifique et par son soutien pratique dans nos travaux communs. Merci à Fabien Gandon pour les commentaires apportés à la lecture de ce mémoire et les différentes discussions que nous avons pu avoir dès mes premiers pas dans le monde de la recherche. Ses travaux m'ont initié au domaine du Web Social Sémantique et pour cela je lui serais toujours reconnaissant. Merci à Chantal Reynaud d'avoir accepté le rôle de rapporteur, ainsi que pour l'intérêt qu'elle a toujours porté à mes recherches. La formation et l'orientation que j'ai reçues de sa part lors de mes études de Master ont ou un impact important pour mes futurs travaux. Merci également à Aldo Gangemi pour son soutien et ses commentaires donnés dans les différentes phases de mes travaux.

Merci à Simone De Liberato d'avoir proposé cette thèse au sein d'Hypios. Je tiens également à remercier à l'ensemble de mes collaborateurs, personnes brillantes, chez Hypios. Merci à Oussama Amar qui a créé la vision innovante de l'entreprise et qui à développé une ambiance de travail créative et stimulante. Merci à Louis-Daniel de Segovia Gross qui a aidé à faire vivre l'entreprise dans un moment de difficulté ainsi préservant la continuité de ma thèse. Merci à Olivier Petros et à Alain Risbourg qui ont assuré cette même continuité à un autre moment dur. Grace aux circonstances bouleversantes et au mélange de personnes hétérogènes, créatives, motivées, brillantes avec qui j'ai eu l'occasion de travailler chez Hypios, cette entreprise qui est constamment en train de vivre l'impossible, représente pour moi une grande mystère – un concept unique dans l'espace et dans le temps. Les expériences tirées de mon travail chez hypios ont eu un énorme impact sur ma capacité de prendre des risques dans mes travaux de recherches, et ces risques ont été à l'origine de mes découvertes les plus intéressantes.

Merci à l'ensemble des personnes avec qui j'ai pu échangé et travaillé durant cette thèse : Matthew Rowe, Jelena Jovanovic, Claudia Wagner, Danica Damljanovic, Valentina Presutti, Nikola Milikic, Darko Jesic, Jovana Kovacevic et bien d'autres encore. C'était souvent grâce à la synergie, émergente de notre travail collaboratif, que nous avons pu être contents des résultats de nos travaux. Merci à Lora Aroyo pour ses conseils utiles lors du PhD Symposium à Shanghai qui m'ont permis de mieux maîtriser le conflit entre mes objectifs scientifiques et le temps, toujours insuffisant.

Merci à mes relecteurs qui ont transformé mon texte initial en un vrai ouvrage en français: Denis Wodrascka, Anna Shurtz, Sandrine Dang, Maria Biasutti, et Chantal Lafargue.

Merci à ma mère et à ma famille pour leur soutien et surtout à ma tante, Roska, qui a fait pour moi des énormes services et sacrifices me permettant de me consacrer à ma thèse.

Enfin, merci à Vaja qui a toujours été à mes côtés pour me soutenir et me mettre au défi durant cette longue étape et à qui je dois tant. Sans elle, ma vie et mes travaux n'auraient sans doute pas pris un chemin si intéressant et si passionnant.

Milan Stankovic, Paris, Septembre 2012

Table des matières

1. Introduction 7

1.1. Contexte des travaux 7

1.1.1. Innovation Ouverte et innovation sur le Web 8

1.1.2. Cadre institutionnel et expérimental 11

1.2. Problématique scientifique 13

1.2.1. Traitement du graphe Web en vue des découvertes pertinentes dans un processus d'innovation 13

1.2.2. Sérendipité et l'aspect inattendu de la navigation dans le graphe de ressources Web 14

1.2.3. Modélisation des entités Web liées à la résolution de problèmes et l'innovation sur le Web et l'échange des données entre des systèmes hétérogènes 15

1.3. Notion d'expertise et de compétence dans le cadre de nos travaux 16

1.3.1. Définition composée de la compétence 16

1.3.2. Perspective de la formation en ligne 18

1.3.3. Perspective de la gestion des ressources humaines 18

1.3.4. Perspective de la résolution de problèmes et de l'innovation 18

1.4. Organisation du mémoire et vision synthétique des résultats 20

1.4.1. Guide de lecture 20

1.4.2. Résultats et contributions par chapitre 21

Chapitre 1 – Introduction 21

Chapitre 2 - Web Social Sémantique 21

Chapitre 3 - Cadre conceptuel: Rapprochement des éléments Web liés à l'innovation 22

Chapitre 4 - Rapprochement des experts et concepts de leur expertise 23

Chapitre 5 - Rapprochement de concepts 24

Chapitre 6 - Rapprochement de traces utilisateurs 25

Chapitre 7 - Rapprochement de collaborateurs dans les scénarios d'innovation sur le Web 25

Chapitre 8 - Ontologies pour faciliter les rapprochements 26

Chapitre 9 - Conclusion 26

2. Web Social Sémantique 29

2.1. Web 29

2.1.1. Contexte historique 29

2.1.2. Le paradigme 32

2.1.3. La réalisation technique 33

2.2. Web Social 34

2.3. Web Sémantique et Données Liées 36

2.3.1. Vision du Web Sémantique 36

2.3.1.1. Représentation des données 38

2.3.1.2. Ontologies, sens de données et raisonnement 40

2.3.1.3. Interrogation du Web Sémantique 41

2.3.2. Données liées – Le Web Sémantique d'aujourd'hui 41
2.3.3. La désambigüisation des sujets d'intérêt sur le Web de données 43
2.4. Web Social Sémantique 44
2.4.1. Vers des données structurées à partir du contenu créé par les utilisateurs 45
2.4.1.1. Extraction de ressources représentant des sujets d'intérêt, définies dans le LOD 46
2.5. Une vision partielle du Web d'avenir 47
2.6. Conclusions 49
3. Cadre conceptuel : Rapprochement des éléments Web liés à l'innovation 53
3.1. Accès à l'information sur le Web 54
3.2. Notion du rapprochement 56
3.2.1. Définition du rapprochement des éléments du graphe Web 56
3.2.2. Exemples de la pratique de rapprochement dans le contexte du Web de documents 57
3.3. Les éléments du graphe Web dont le rapprochement est utile dans les scénarios liés à l'innovation 58
3.3.1. Traces utilisateurs 58
3.3.1.1. Abstractions : profil social et profil conceptuel 60
3.3.1.2. La genèse de traces - les activités 61
3.4. Rapprochement dans le contexte du Web de données 63
3.4.1. Les avantages de Web de données pour la pratique des rapprochements 63
3.4.2. Rapprochements abordés dans cette thèse 64
3.4.2.1. R1 : Rapprochement des concepts et des experts 67
3.4.2.2. R4 : Rapprochement des concepts entre eux 68
3.4.2.3. R7 : Rapprochement visant à enrichir des traces utilisateur 69
3.4.2.4. R10 : Rapprochement des experts entre eux 70
3.4.3. Quelques caractéristiques du rapprochement des éléments Web 71
3.4.3.1. Effets des rapprochements dans les scénarios liés à l'innovation 71
3.4.3.2. Rapprochements permanents et temporaires 72
3.5. Travaux Liés 73
3.5.1. Systèmes de recommandation et de recherche qui utilisent le Web de données 73
3.5.2. Approches pour la découverte de liens dans le Web de données 74
3.5.3. Interdépendance du graphe social et graphe conceptuel 75
3.6. Conclusions 77
4. Rapprochement des experts et des concepts de leur expertise 81
4.1. Rapprochement des personnes et concepts dans le cadre de la recherche d'experts 81
4.1.1. Scénario d'usage et les objectifs de rapprochement 81
4.1.2. Processus générique de rapprochement 82
4.2. Hypothèses d'expertise et le graphe Web 83
4.2.1. Hypothèses d'expertise par le type de trace concerné 83
4.2.1.1. Quelques exemples d'hypothèses d'expertise et leur utilisation dans la littérature 85
4.2.1.1.1. Les hypothèses d'expertise liées au contenu créé par les utilisateurs 85

4.2.1.1.2. Les hypothèses d'expertise liées aux activités des utilisateurs 86
4.2.1.1.3. Les hypothèses d'expertise liées à la réputation des utilisateurs 88
4.2.2. Richesse du graphe Web en traces de différents types 89
4.2.3. Problèmes techniques actuels et remèdes possibles 94
4.2.3.1. Données à l'accessibilité réduite ; 94
4.2.3.2. Manque de données 95
4.2.3.3. Manque de détails dans les données 96
4.2.3.4. Manque de liens entre les données 96
4.2.3.5. Bonne pratique de la publication de données liées utiles pour la recherche d'experts 97
4.3. Choix d'hypothèse d'expertise 97
4.3.1. Principe général 98
4.3.2. Conception de métriques 99
4.3.3. Expérimentation 100
4.3.3.1. Jeu de données expérimental 101
4.3.3.2. Mesures de performance 101
4.3.3.3. Création de l'étalon-or 102
4.3.3.4. Test de corrélation 103
4.3.4. Niveaux d'applicabilité des métriques 104
4.3.4.1. Calcul des métriques sur un seul EDS 104
4.3.4.2. Calcul de métriques sur un ensemble d'EDS 105
4.3.4.3. Calcul de métriques sur le LOD entier 105
4.4. Hy.SemEx – Système de recherche d'experts sur les EDS du LOD 105
4.4.1. HySem.Ex dans le contexte de son usage 106
4.4.2. Architecture du logiciel 108
4.4.3. Retour d'utilisateurs 111
4.5. Conclusions 112
5. Rapprochement de concepts 117
5.1. Rapprochement des concepts au service de la découverte et de la latéralité dans la résolution de problèmes 118
5.1.1. Contexte de rapprochement 118
5.1.2. Qualités désirées d'une approche de découverte de concepts 119
5.2. Etat de l'art 120
5.2.1. Types d'approches 120
5.2.2. Approches basées sur la cooccurrence 121
5.2.3. Approches basées sur l'exploitation du graphe 122
5.2.3.1. Approches basées sur les données liées 123
5.2.4. Ce qui motive le développement d'une approche basée sur de données liées 123
5.3. Conception d'une approche basée sur des données liées 124
5.3.1. Approche générale 125

5.3.1.1. Traitement des liens hiérarchiques 127
5.3.1.1.1. Comparaison des versions alternatives de l'approche 130
5.3.1.2. Traitement des liens transversaux 134
5.3.1.2.1. Adaptation de l'approche par rapport au cas d'utilisation 135
5.3.1.2.2. Valeur ajoutée par l'utilisation de la fonction de pondération 136
5.3.1.3. Approche mixte 137
5.3.1.4. Comparaison des approches 137
5.4. Evaluation 139
5.4.1. Evaluation à l'aide de l'étalon-or 139
5.4.1.1. Comparaison 140
5.4.2. Etude utilisateurs 141
5.4.2.1. Contexte de l'étude 142
5.4.2.2. Comparaison avec Google AdWords 144
5.4.2.3. Comparaison avec Random Indexing 144
5.5. Le logiciel hyProximity 145
5.5.1. L'interface Web de hyProximity 147
5.6. Utilisation du système dans la plateforme Hypios 149
5.6.1. Cas d'utilisation 1 149
5.6.2. Cas d'utilisation 2 151
5.6.3. Exemples de suggestions 152
5.7. Conclusions 153
6. Rapprochement de traces utilisateurs 157
6.1. Introduction 157
6.1.1. Motivation générale 157
6.1.2. Le cas de tweets et d'articles scientifiques 158
6.2. Travaux liés 159
6.2.1. Annotation de tweets par des concepts 159
6.2.2. Différents travaux de mise en correspondance de tweets et autres ressources 161
6.3. L'approche RATED 161
6.3.1. Transformation de tweets en RDF 162
6.3.2. Enrichissement par concepts 163
6.3.3. Alignements tweet-article 164
6.3.3.1. Alignement manuel 164
6.3.3.2. Automatisation des alignements 166
6.4. Exemple d'usage des alignements 167
6.5. Conclusions 168
7. Rapprochement de collaborateurs dans les scénarios d'innovation sur le Web 171
7.1. Rapprochement de collaborateurs 172
7.1.1. Qualités désirées d'une approche de découverte de collaborateurs 173
7.2. Etat de l'art 174

7.2.1. Recommandation de personnes 174
7.3. Vers une approche de recommandation des collaborateurs dans le cadre de l'innovation à travers le Web 176
7.3.1. L'approche générale 177
7.3.2. Création des profils 178
7.3.3. Extension des profils 179
7.3.4. Mesures de similarité et calcul 180
7.3.5. Classement de candidats collaborateurs 182
7.4. Evaluation 182
7.4.1. Evaluation 1 184
7.4.1.1. Contexte de l'étude 184
7.4.1.2. Résultats 185
7.4.2. Evaluation 2 186
7.4.2.1. Création de « l'étalon-or » 186
7.4.2.2. Résultats 186
7.5. Un exemple de collaborateurs suggérés 188
7.6. Conclusions 189
8. Ontologies pour faciliter les rapprochements 193
8.1. Cadre général des ontologies 194
8.1.1. Introduction 194
8.1.2. Les éléments modélisés 194
8.1.3. Principes de modélisation 196
8.2. Modélisation des éléments fondamentaux 196
8.2.1. Modélisation d'utilisateurs 197
8.2.2. Modélisation de concepts 198
8.2.3. Modélisation de traces 199
8.2.3.1. Ontologies existantes pour des traces 199
8.2.3.1.1. Contenu généré par les utilisateurs 199
8.2.3.1.2. Activités et accomplissement des utilisateurs 202
8.2.3.1.3. Réputation des utilisateurs 202
8.2.4. Modélisation de problèmes 202
8.2.4.1. Scénarios d'usage de données 203
8.2.4.2. Ontologie PCO 203
8.3. Modélisation des hypothèses 205
8.3.1. Hypothèses d'expertise 207
8.3.2. Hypothèses de proximité conceptuelle 208
8.4. Modélisation des proximités 209
8.4.1. Proximité expert-concept 209
8.4.2. Proximité concept-concept 210
8.5. Un scénario d'usage des données 210

8.6. Conclusions 211
9. Conclusions 213
9.1. Conclusions par rapport aux hypothèses de départ 213
9.2. D'autres champs d'applicabilité de nos travaux et les travaux futurs 215
9.2.1. La recherche d'experts au sein d'entreprises 215
9.2.2. Suggestion de concepts pour des campagnes publicitaires en ligne 216
9.2.3. Prédiction de liens sociaux et recommandation de personnes à suivre 216
9.3. Problématiques émergentes dans le domaine de l'innovation sur le Web 216
9.3.1. Soutien aux consultants et métiers intermédiaires 216
9.3.2. Veille dans le cadre de processus d'innovation 217
9.4. Au-delà du Web Social Sémantique 217
9.4.1. Vers un Web véritablement intelligent 217
9.4.2. Un Web qui « comprend » les émotions et les traits de personnalité 218
9.4.3. Vers un Web de subconscience 218
Annexe 221
Annexe 1 : Exemple d'appel aux solutions en RDF 221
Annexe 2 : Hypothèses de proximité sociale 222
Annexe 3 : Hypothèses de proximité des traces 223
Annexe 4 : Proximité collaborateur-collaborateur 224

Table des Figures

Figure 1 Le paradigme de l'innovation ouverte 10
Figure 2 L'innovation ouverte à travers une plate-forme d'innovation sur le Web 11
Figure 3 Définition composée de la compétence 17
Figure 4 Organisation du mémoire 20
Figure 5 Articulation de chapitres clés 20
Figure 6 Evolution du Web. Repris de Nova Spivack [Spivack, 2009] 31
Figure 7 « Semantic Web Layer Cake », repris de Tim Berners-Lee 37
Figure 8 Trois couches du Web Sémantique 38
Figure 9 Exemple de graphe (décrivant la couche de données représentée sur Figure 8) en format Turtle 39
Figure 10 Nombre d'ensembles de données liées ouvertes, D'après les statistiques de Richard Cyganiak 42
Figure 11 Extrait de l'information concernant Paris sur DBPedia 43
Figure 12 Un extrait du graphe de données liées ouvertes, en Septembre 2011. D'après Richard Cyganiak 44
Figure 13 Convergence du Web Sémantique et Web Social. Repris de [Breslin et al., 2009] 45
Figure 14 La vision d'un Web d'avenir 48
Figure 15 Propagation de l'information à travers le Web et sa conséquence dans le graphe Web 54
Figure 16 Changement de la portée immédiate d'accès à l'information du graphe Web lors de l'interaction itérative 55
Figure 17 Exemple de trace (triplet utilisateur-trace-sujet) sous forme graphique 59
Figure 18 Interdépendance entre le graphe social et le graphe conceptuel à travers les activités des utilisateurs 62
Figure 19 Interdépendance des traces et des connaissances 63
Figure 20 Rapprochements théoriquement possibles entre des éléments du graphe Web liés à l'innovation 65
Figure 21 Chapitres dans lesquelles les rapprochements sont présentés 66
Figure 22 Rapprochement basé sur des types de traces 68
Figure 23 Rapprochement des concepts utilisant différents types de liens 68
Figure 24 Rapprochement basé sur des mises en correspondance des traces 69
Figure 25 Rapprochement en plusieurs dimensions 70
Figure 26 Rapport entre la pertinence et l'aspect inattendu des recommandations, diagramme inspiré par [Akiyama & Obara, 2010] 72
Figure 27 Le choix et l'application des hypothèses d'expertise dans les plateformes d'innovation sur le Web 83

Figure 28 Représentation graphique de tests 89
Figure 29 Exemple de description d'EDS en VoID et SCOVO 99
Figure 30 Mode d'interaction avec le système Hy.SemEx 106
Figure 31 Capture d'écran de Hy.SemEx - choix d'hypothèse d'expertise 107
Figure 32 Capture d'écran de Hy.SemEx - Liste d'experts 107
Figure 33 Principaux composants du logiciel Hy.SemEx et du logiciel complémentaire pour la collecte et la normalisation de données 108
Figure 34 Enrichissement de l'ensemble de concepts initiaux 119
Figure 35 Structure d'un exemple de graphe de données liées 125
Figure 36 Distribution des distances dans DBPedia repris de [Lehamnn et al., 2007] .. 128
Figure 37 Précision pour des concepts obtenus par $hyP^{hyérarchique}$ en utilisant différentes fonctions de distance et p1 en tant que fonction de pondération 132
Figure 38 Rappel pour des concepts obtenus par $hyP^{hyérarchique}$ en utilisant différentes fonctions de distance et p1 en tant que fonction de pondération 132
Figure 39 Précision obtenue en utilisant des fonctions de pondération p1 et p2, toujours avec la même fonction de distance d2 133
Figure 40 Rappel obtenu en utilisant des fonctions de pondération p1 et p2, toujours avec la même fonction de distance d2 134
Figure 41 Distances transversales 134
Figure 42 Précision de la fonction de distance transversale sans et avec la fonction de pondération 136
Figure 43 Rappel de la fonction de distance transversale sans et avec la fonction de pondération 137
Figure 44 Précision de $hyP^{hiérarchique}$, $hyP^{tranversale}$, et hyP^{mixte} 138
Figure 45 Rappel de $hyP^{hiérarchique}$, $hyP^{tranversale}$, et hyP^{mixte} 138
Figure 46 Mesure F de $hyP^{hiérarchique}$, $hyP^{tranversale}$, et hyP^{mixte} 139
Figure 47 Précision $hyP^{hiérarchique}$, $hyP^{tranversale}$, adWords et Random Indexing 140
Figure 48 Rappel de $hyP^{hiérarchique}$, $hyP^{tranversale}$, adWords et Random Indexing 141
Figure 49 Mesure F de $hyP^{hiérarchique}$, $hyP^{tranversale}$, adWords et Random Indexing 141
Figure 50 Vue d'haut niveau des composants principaux du logiciel HyProximity 146
Figure 51 Interface hyProximity - page 1 147
Figure 52 Interface hyProximity - page 2 148
Figure 53 Interface hyProximity - page 3 149
Figure 54 Intersections d'ensemble d'experts 150
Figure 55 Rétention de concepts suggérés par hyProximity dans un processus itératif de rédaction de problème 151
Figure 56 Production de données liées à partir de tweets et articles 161
Figure 57 Structure du système pour la transformation des tweets en RDF 162
Figure 58 Un exemple de tweet dans le format RDF 163

Figure 59 Structure du système pour l'enrichissement du contenu par concepts 164
Figure 60 Exemple de sous-graphe Web pertinent pour le rapprochement de collaborateurs potentiels 173
Figure 61 Quatre phases de la découverte de collaborateurs possibles 177
Figure 62 Vectorisation de deux profils conceptuels 180
Figure 63 Calcul de la différence conceptuelle et vectorisation des profils. Sous l'hypothèse que toutes les valeurs d'importance soient = 1 pour des concepts présents dans le profil 182
Figure 64 Résultats de la métrique DCG appliquée sur le calcul de la similarité SDC, avec l'extension de profils de problème par différentes méthodes 185
Figure 65 Résultats de la métrique DCG appliquée sur le calcul de la similarité SCI, avec l'extension de profils de problème par des méthodes différentes 186
Figure 66 mesure AMP calculée pour différentes positions dans le classement (axe x) 187
Figure 67 DCG calculé pour les 10 premières suggestions de chaque approche 188
Figure 68 Eléments modélisés et leurs relations 196
Figure 69 Description d'un utilisateur à l'aide de l'ontologie FOAF, en RDF (syntaxe XML) 197
Figure 70 Exemple d'usage du SKOS pour décrire les relations entre des concepts 199
Figure 71 Exemple d'un billet représenté par SIOC, repris du site officiel. 200
Figure 72 Exemple du tag représenté par MOAT, repris du site officiel. 201
Figure 73 Classes principales de l'ontologie PCO 204
Figure 74 Classes principales de l'ontologie d'hypothèses de pertinence 206
Figure 75 Un exemple d'hypothèse d'expertise représentée par EHO, format Turtle ... 208
Figure 76 Un exemple d'hypothèse de la proximité conceptuelle représentée par RHO 208
Figure 77 Un exemple de résultats de rapprochement de concepts et experts décrits par l'ontologie RHO 209
Figure 78 Un exemple de résultats de rapprochement de concepts décrit par l'ontologie RHO 210
Figure 79 Scénario de réutilisation des proximités 211
Figure 80 Exemple de problème en format RDF. L'extrait de code est obtenu à partir d'un problème réel sur le site hypios.com, en traitant sa page Web par le distillateur RDFa pour en extraire du code RDF. 222
Figure 81 Un exemple d'hypothèse de proximité sociale représenté par EHO 223
Figure 82 Un exemple d'hypothèse d'expertise représentée par EHO 223
Figure 83 Un exemple d'hypothèse de la proximité des traces représentée par RHO ... 223
Figure 84 Un exemple d'hypothèse d'expertise représentée par EHO 224

Liste des Tableaux

Tableau 1 Résultats de tests T1-T4 pour les hypothèses d'expertise H1-H18 91
Tableau 2 Matrice de confusion pour le calcul de la métrique kappa..... 103
Tableau 3 Le taux d'accord des évaluateurs par sujet, dans les deux itérations..... 103
Tableau 4 Valeurs des métriques et la performance de la recherche d'experts 104
Tableau 5 Fonctions de distance hiérarchiques..... 129
Tableau 6 Propriétés pertinentes..... 136
Tableau 7 Note moyenne ± déviation standard dans l'étude 1 144
Tableau 8 Note moyenne ± déviation standard dans l'étude 2 145
Tableau 9 Résultats de la campagne de diffusion du problème 150
Tableau 10 Résultats de la diffusion des 3 problèmes de février 2011 151
Tableau 11 Exemple de suggestions hyProximity 1 152
Tableau 12 Exemple de suggestions hyProximity 1 153
Tableau 13. Les concepts les plus populaires dans les tweets correspondant à ESWC2010..... 165
Tableau 14. Les concepts les plus populaires ESWC2010 dans les articles liés aux tweets 165
Tableau 15 Informations à modéliser dans le cas de rapprochements différents 194

Glossaire

cocnept – Un concept est une ressource, définie dans le graphe de données liées, qui identifie un sujet de la connaissance humaine. Par exemple, la ressource définie sur http://dbpedia.org/resource/Nanotechnologie est un concept qui identifie la nanotechnologie. Tous les sujets qui constituent le savoir humain et surtout ceux qui peuvent faire objet de la recherche scientifique peuvent être représentés par une ressource –concept.

données sémantiquement structurées (abrégé par « données structurées ») – Des données sémantiquement structurées sont des données mises à disposition aux formats structurés et typés, tels que RDF. Au delà de la structure riche, ces données sont auto-descriptives et forment un graph de nœuds typés, liés entre eux par des liens typés. Leur richesse permette aux machines de mieux les traiter.

élément Web – le Web est composé de ressources informationnelles potentiellement utiles pour un utilisateur dans une situation particulière. Par exemple, une image sur le Web est un élément Web ainsi qu'un ensemble d'informations décrivant la France, consultables dans une page Web. Nous appelons ces composants informationnels « les éléments Web » quelle que soit leur forme.

EDS (ensemble de données structurées) – Un EDS est un ensemble de données structurées. Ces ensembles peuvent être mis à disposition sous forme de fichiers ou bien consultables via un point d'accès permettant de les interroger en utilisant un langage de requêtes.

graphe de données liées (abrégé par « données liées ») – Le graphe de données liées est le graphe qui emerge de la création de données structurées et de la pratique de la création de liens entre données. Certaines parties de ce graphe peuvent être accessibles à tout le monde, tandis que l'accès à certaines autres parties peut être restreint.

graphe Web – Toutes les ressources du Web (les documents, les données, etc.) et leur liens forment le graphe Web.

hypothèse d'expertise – L'hypothèse d'expertise est une règle qui définit des conditions nécessaires pour considérer quelqu'un comme expert d'un sujet précis dans une situation spécifique. Cette notion permet de mieux distinguer et comparer des différentes approches de recherche d'experts sur le Web.

latéralité – la mise en connection des idées, solutions, personnes provenant des spheres intellectuelles et concepteulles differentes mais neanmoins pertinentes. La latéralité est souvent vue comme nécessaire pour faire émerger la sérendipité.

LOD (en. *Linked Open Data Cloud*) – Le graphe de données liées publique ou simplement *LOD* est la partie publique du graphe de données liées.

rapprochement des éléments Web (abrégé par « rapprochement ») – Un rapprochement des éléments Web est une opération, ayant des effets temporaires ou permanentes, effectuée au niveau du graphe Web, et qui permet de considérer deux (ou plusieurs) éléments utiles l'un pour l'autre dans une situation particulière comme étant plus proches dans le sens de l'accessibilité mutuelle, ainsi augmentée.

sérendipité – La sérendipité est le fait de réaliser une découverte inattendue, au cours d'une recherche dirigée initialement vers un objet différent de cette découverte

trace utilisateur – Une trace utilisateur est un élément Web, ou bien un ensemble d'éléments Web, qui relient un utilisateur et un concept - sujet de connaissance humaine. Une trace qui vient sous forme de ressources qui font partie du graphe de données structurées est dite « trace structurée ».

tweet – messages très courtes echangés entre des utilisateurs sur le Web. Le mot « tweet » viens du nom d'un service Twitter.com qui a popularisé la partique de publication massages d'une taille limitée à 140 caractères.

Web de données – Le Web des données est un Web dont le contenu n'est pas limité aux pages Web interliées, mais qui comprends aussi des données structurées liées entre elles pour constituer un réseau global d'informations.

Web Sémantique – Le Web Sémantique est une vision du Web qui, par la richesse de sa structure des nœuds et des liens typés, favorise la collaboration entre les utilisateurs humains et les machines. Celle-ci est rendue possible par la plus grande capacité de machines de « comprendre » et traiter le contenu du Web.

Web Social – Le Web Social est une facette du Web qui concerne les echanges sociales entre les utilisateurs. C'est un ensemble de caracteristiques des applications Web qui favorisent ces echanges et rendent le Web un véritable cannal de communication.

1. Introduction

> *Ce premier chapitre présente le contexte général de la thèse (§1.1). Il permet aux lecteurs de comprendre les raisons qui nous ont motivé à aborder le sujet choisi et la problématique qui se trouve au cœur de nos recherches. Nous décrivons notamment les avantages et inconvénients du milieu industriel dans lequel la thèse fut réalisée. Nous présentons enfin l'organisation générale du mémoire (§1.3)*

1.1. Contexte des travaux

Les travaux présentés dans cet ouvrage ont été effectués dans le cadre d'une convention CIFRE[1] entre le laboratoire LaLIC (Language, Logiques, Informatique, Cognition)[2], maintenant composante du laboratoire STIH (Sens, Texte, Informatique, Histoire)[3] de l'Université Paris-Sorbonne IV et Hypios SAS, une jeune entreprise qui propose une manière alternative de réaliser de la recherche et développement (R&D par la suite) en industrie. Par conséquent, les résultats obtenus reflètent un conflit générateur de créativité entre ces deux mondes, celui de la recherche et celui de l'industrie, qui leur donne une note de dynamisme. Ces deux plans, théorique et appliqué, se sont mutuellement influencés au cours de la progression de la thèse. Les percées théoriques ont été sollicitées pour satisfaire des besoins découverts dans le milieu industriel. Les réalisations pratiques, à leur tour, ont également enrichit rétroactivement le plan théorique. Ces cycles d'influence sont à l'origine des contributions issues de la thèse.

---

[1] Convention Industrielle de Formation par la Recherche, http://www.anrt.asso.fr/fr/espace_cifre/accueil.jsp

[2] http://lalic.paris-sorbonne.fr/

[3] http://www.stih.paris-sorbonne.fr/

Si le contexte industriel d'une start-up qui a changé son modèle d'affaire et sa vision de sa place sur le marché plusieurs fois depuis le début de nos travaux nous a confronté à une certaine instabilité et incertitude, il nous a aussi amené à chercher une stabilité du côté de la recherche et à construire des théories et des solutions qui seront applicables dans une variété de situations industrielles. L'instabilité du financement et l'incertitude de l'avenir de l'entreprise, naturellement présentes dans une start-up, nous ont mis sous une pression constructive en nous incitant à travailler chaque jour comme si c'était le dernier et à produire peut-être plus et plus rapidement que ce que nous aurions fait dans une ambiance de stabilité et de certitude.

Un grand nombre de nos choix de direction de recherche s'explique également par l'époque dans laquelle nous avons entamé ces travaux au cours de laquelle le domaine du Web Sémantique est entré dans une phase de légère maturation. Cette maturité nous a permis d'aller au delà des propositions théoriques et d'expérimenter avec des composantes du Web Sémantique dans des conditions réelles. A ce stade dans l'évolution de la technologie, une orientation vers la recherche empirique a été nécessaire. Ce choix se reflète dans les nombreuses évaluations que nous avons conduites avec la plus grande rigueur possible.

La nature pluridisciplinaire du domaine de la recherche de la Science du Web dans lequel cette thèse s'inscrit partiellement a incité à développer un certain nombre de collaborations, souvent internationales, qui ont permis d'explorer des intersections possibles de nos travaux de recherche avec des travaux issus d'autres domaines tels que la recherche d'information, la fouille de données, la gestion de l'innovation, etc. Nous nous sommes servis de ces échanges enrichissants pour étendre nos travaux et renforcer notre argumentation.

Il est également important de souligner la vision fondatrice qui nous a motivé à entreprendre cette thèse, et qui a d'ailleurs été à l'origine du choix de notre parcours en informatique. La vision des hommes et des machines qui travaillent ensemble, induisant l'émergence d'un Homme dont les capacités sont augmentées par la technologie informationnelle a été d'abord proposée par Vanevar Bush dans son article *As We May Think* [Bush, 1945]. Elle se poursuit dans les travaux de Ted Nelson sur hypertext [Nelson, 1965], et également, de manière plus développée, dans les travaux de Douglas Engelbart, plus précisément dans son système NLS Augment [Englebart, 1962] qui vise à augmenter l'intelligence humaine par des outils de collaboration en ligne et d'extension de la mémoire humaine. Le Web, créé par Tim Berners-Lee en 1989 [Bernes-Lee, 2000], représente un des plus grands chantiers de réalisation de cette vision. Nous avons donc fondé nos travaux sur le Web et ses différentes variétés telles le Web Social et le Web Sémantique, dans la poursuite d'une solution qui augmentera la capacité intellectuelle de l'homme par l'accès à l'information qui lui est utile dans une variété des situations, notamment celles liées à l'activité inhérente à l'innovation.

1.1.1. Innovation Ouverte et innovation sur le Web

La capacité d'innover est d'une importance critique dans la réussite des entreprises dans l'économie moderne. Dans la recherche de compétitivité, les entreprises de tous les secteurs cherchent à concevoir des produits et des services innovants avant leurs concurrents. Dans le passé, ce travail créatif était uniquement confié aux départements de R&D internes aux entreprises, mais nous sommes maintenant témoins d'une évolution du paradigme de l'innovation industrielle vers un modèle d'innovation de plus

en plus ouvert. Le nouveau paradigme d'ouverture, baptisé « l'innovation ouverte », à été notamment repéré et étudié par Chesbrough [Chesbrough, 2003]. Ce paradigme comprend la recherche des innovateurs et des idées à l'extérieur de l'entreprise, principalement en établissant des relations avec des universités et des experts indépendants, souvent grâce à des consultants et médiateurs (Figure 1). L'émergence du nouveau paradigme est dû principalement à une mobilité des cerveaux et des capitaux de plus en plus élevée, ainsi qu'à l'accessibilité du savoir faire par l'intermédiaire des médias tels que le Web. Les principaux avantages associés à la notion d'innovation ouverte sont :

- **La réduction du risque,** qui est inhérente aux activités de R&D**.** S'investir dans une seule direction de recherche, ce qui est le cas dans certains des modèles de R&D internes, augmente le risque d'échec par rapport aux modèles qui exploitent plusieurs pistes diverses en même temps ;
- **La réduction des frais de R&D**. La mise en place des activités de R&D internes implique des frais importants engagés dans l'exploration de pistes, dont seulement certaines portent des fruits utiles pour l'entreprise. Contrairement aux acteurs industriels, les organismes et les individus spécialisés en R&D peuvent potentiellement valoriser leurs travaux auprès de plusieurs clients dans des domaines différents ce qui réduit le coût du savoir faire qu'ils produisent ;
- **L'exploitation des idées provenant de sources inattendues et la découverte de nouvelles pistes de recherche**. Une ouverture vers l'extérieur augmente naturellement les chances des découvertes fortuites et inattendues qui sont souvent à l'origine des innovations les plus révolutionnaires.

Si ce nouveau paradigme a ouvert l'accès à plus de sources d'innovation, il a aussi imposé des efforts supplémentaires dans la gestion des relations entre les nouveaux acteurs de cet écosystème d'innovation. La multiplication des liens dans ce nouvel espace a rendu évident le besoin d'une plate-forme commune de gestion des échanges dans le processus d'acquisition de savoir faire extérieur. Les plates-formes d'innovation sur le Web, telles que NineSigma[4], InnoCentive[5], et Hypios[6] ont vu le jour justement pour répondre à ce besoin, en attirant beaucoup de grandes entreprises à utiliser leurs services. A titre d'exemple, Innocentive, fondée en 2005, a effectuée, à elle seule des transferts de propriété intellectuelle d'un valeur totale supérieure à 34 million de dollars[7]. Ces plates-formes proposent aux entreprises de mettre en ligne sur le Web leurs besoins de savoir faire sous forme de « problèmes » d'innovation, et de les diffuser à un grand nombre d'experts, de personnes expérimentées et de porteurs d'idées. La relation entre les porteurs des solutions innovantes et l'entreprise prend la forme d'un appel aux solutions qui est géré en intégralité par la plate-forme qui assure le transfert de la propriété intellectuelle en échange d'une prime.

La valeur ajoutée qu'apportent les plates-formes d'innovation ouvertes est supérieure à celle générée par la notion d'innovation ouverte dont elles sont issues. En premier lieu, le processus d'interaction entre les différents acteurs que proposent les plates-formes, par leur nature centralisée réduit le nombre d'interactions élémentaires et par là même

[4] http://www.ninesigma.com/

[5] http://www.innocentive.com/

[6] http://www.Hypios.com/

[7] http://www.innocentive.com/about-innocentive/facts-stats

réduit le temps passé dans l'acquisition d'un savoir faire, par rapport aux modèles non basés sur le Web (Figure 2).

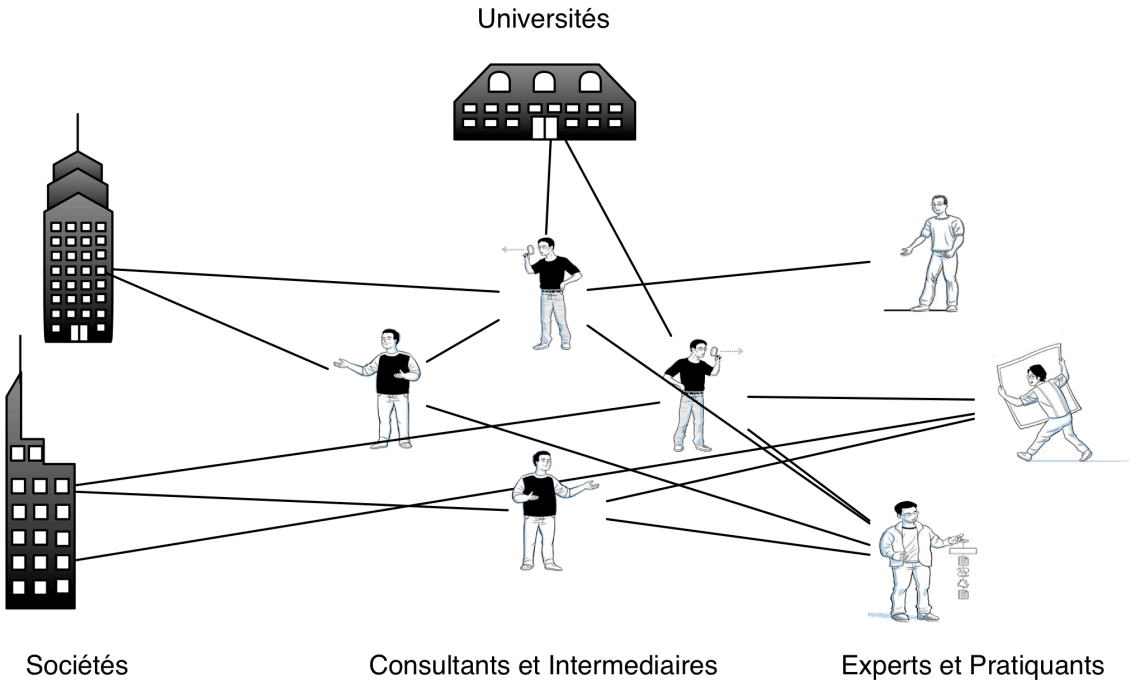


Figure 1 Le paradigme de l'innovation ouverte

La démocratisation de processus d'innovation est probablement parmi les avantages les plus importants de cette approche. Par la démocratisation nous entendons la possibilité donnée à chaque individu de participer au processus d'innovation, quel que soit son niveau de compétence, d'éducation, et quelle que soit sa réputation dans le domaine d'expertise du problème. Une étude menée par Jeppesen et Lakhani [Jeppesen & Lakhani, 2009] montre que cette démocratisation est fortement souhaitable de par ses effets favorables au processus d'innovation. Cette étude démontre notamment que les personnes qui se considèrent marginalisées par rapport à un problème d'innovation, sont souvent très motivés pour entrer en compétition et sont souvent à l'origine des solutions les plus innovantes. Si ce sentiment de marginalité est parfois dû à un statut démographiquement minoritaire de l'individu, comme par exemple cela peut être le cas des femmes dans un domaine dominé par des hommes e.g. l'informatique, il peut également provenir du fait que le domaine d'expertise d'innovateur est à cheval sur plusieurs domaines et qu'il n'est pas principalement centré autour du domaine d'activité duquel est issu le problème.

Enfin, le modèle d'acquisition de savoir faire par concours, permet aux entreprises de considérer **la multiplicité des perspectives** sur un même problème, et de s'assurer d'avoir recherché un nombre plus élevé des champs de solutions possibles d'une manière beaucoup plus efficace que ce que les interactions individuelles avec des prestataires pré-choisis permettraient.

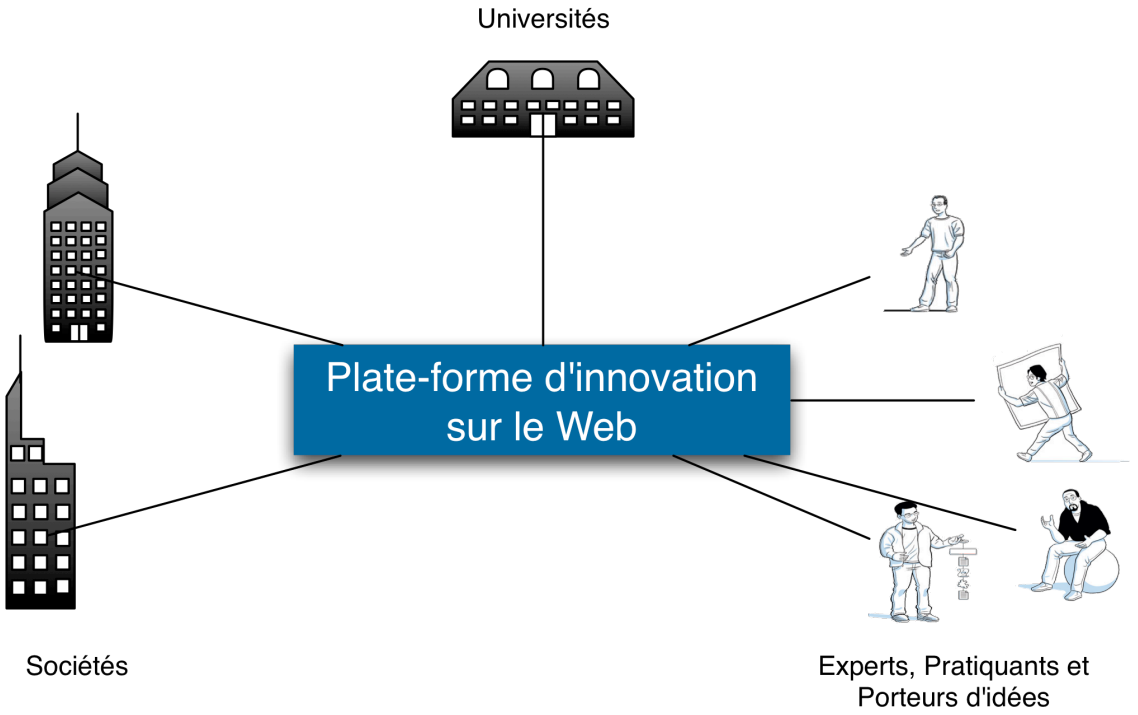


Figure 2 L'innovation ouverte à travers une plate-forme d'innovation sur le Web

Avec l'émergence du Web Social (appelé également le Web 2.0), nous sommes témoins de l'amplification de l'interactivité sur les sites Web modernes qui incite les utilisateurs à créer du plus en plus de contenus, en dévoilant de cette manière de plus en plus de leur intérêts et compétences. De cette nouvelle masse de données naissent de nouvelles opportunités pour chaque utilisateur de valoriser ses connaissances dans un processus de résolution de problème en ligne.

1.1.2. Cadre institutionnel et expérimental

Hypios[8] est une jeune entreprise créée en 2008, basée à Paris et qui met à disposition une plate-forme d'innovation sur le Web[9]. Tous les types de problèmes peuvent être soumis sur le site mais la plupart sont d'ordre technologique (la mesure fiable de l'épaisseur de la glace de la banquise) ou reliés à des problématiques de motivation et de science de gestion (e.g. trouver des motivations non financières pour inciter les ménages à réduire leur consommation d'électricité). Chaque internaute, quelque soit son domaine et son niveau de compétence peut s'inscrire sur la plate-forme et devenir auteur de solutions (en. *solver*). Au moment de la rédaction de ce mémoire Hypios compte plus de 6000 membres inscrites au site en tant qu'auteurs de solutions, dont plus de 300 ont participé de manière active dans la résolution des 36 problèmes que la plateforme a hébergés depuis sa création en 2009 jusqu'à juillet 2012.

Le processus de résolution de problème d'innovation sur Hypios commence par une formulation textuelle d'un problème qu'un client souhaite faire résoudre. Pour effectuer cette étape très importante, Hypios s'appuie sur des cabinets de conseil partenaires, spécialisés dans différents domaines, qui assistent les clients en les aidant à isoler les

[8] Le nom Hypios est un acronyme de "here your problem is our solution" (ici votre problème est notre solution, en anglais).

[9] http://www.Hypios.com

problèmes de recherche pertinents et à les reformuler de manière à les rendre accessibles à des experts issus de domaines variés. Cette description de problème liste également les contraintes que toute solution proposée doit satisfaire, les essais et expériences antérieures, le contexte industriel dans lequel la solution sera utilisée, etc. Elle est suffisamment détaillée pour permettre d'y répondre efficacement sans dévoiler ni l'identité ni l'intention commerciale de l'entreprise cliente. La description textuelle est parfois accompagnée par des images et des croquis illustrant le contexte du problème.

Ensuite, Hypios propose les services suivants qui visent à augmenter la probabilité de résolution d'un problème donné :

- mise en ligne de la description de problème et ouverture d'un concours récompensant par une prime la meilleure solution ;
- diffusion du problème par e-mail à ses membres actifs, membres provenant de domaines différents et variés ;
- identification d'internautes, non inscrits au site Hypios.com, susceptibles d'avoir des solutions et des idées pertinentes sur le problème, provenant soit du domaine du problème soit d'un domaine connexe. Une fois identifiés, ces internautes sont contactés par l'équipe d'Hypios et invités à participer au processus de la résolution du problème ; et
- activités publicitaires sur les moteurs de recherche et les réseaux sociaux, incitant des internautes à consulter la description de problème et de proposer des solutions.

Hypios envisage également de proposer des services aux auteurs de solutions, leur permettant de découvrir des collaborateurs potentiels ou des travaux déjà existants.

Pendant la durée du concours, les internautes peuvent proposer leurs contributions. Selon le type de concours, la forme de ces contributions peut varier entre des simples idées ou bien des solutions complètes, des preuves de concept, et même des prototypes.

Une fois le concours arrivé à son terme, le client revoit les solutions récoltées et en choisit une à laquelle il accorde une prime financière en échange des droits d'usage de la propriété intellectuelle associée à la solution. L'interaction avec l'auteur de la solution gagnante peut se poursuivre dans le cadre de l'implémentation de la solution au sein de l'entreprise. Dans ce cas les identités des deux parties seront finalement révélées.

Nos travaux de recherche effectués pendant les trois dernières années ont, d'un point de vue applicatif, principalement visé à améliorer la capacité d'Hypios à faire la jonction entre les problèmes et les internautes susceptibles d'apporter des solutions innovantes. Si ce cadre de recherche industrielle nous a imposé certains objectifs, il n'a en aucun cas limité l'applicabilité de nos résultats, théoriques aussi bien que pratiques, à d'autres champs d'application. Nous avons cherché à comprendre la relation entre la navigation dans le Web dans le cas général, avec ce qui se produit dans un scénario lié à l'innovation sur le Web. Cette compréhension nous à permis de réfléchir à la transposabilité de nos travaux aux autres scénarios, et souvent d'effectuer des expériences afin d'évaluer le caractère général de nos travaux et d'en définir les limites d'utilisation.

Notre présence à Hypios, presque dès le démarrage de son activité, nous a permis de nous familiariser avec les clients et les auteurs de solutions et de mieux comprendre

leurs vraies attentes. En même temps cela nous a donné un cadre d'évaluation de nos approches, en nous mettant en relation avec des auteurs des solutions, des clients et des collaborateurs d'Hypios qui ont souvent accepté de participer à nos études en tant qu'évaluateurs. Nous avons largement exploité cet avantage pour donner un caractère empirique à notre travail.

Le suivi des différentes phases de développement de la prestation d'Hypios nous a également rendu conscients de la volatilité des prestations et des processus, naturellement présents dans les systèmes dépendants d'utilisateurs humains. Cette volatilité nous à incité à chercher des solutions généralement applicables dans le processus d'innovation sur le Web, quelle que soit la forme de la prestation et de l'interaction des acteurs avec des objets Web. Nous avons même fortement envisagé la création d'une approche indépendante des plates-formes d'innovation qui pourrait être utile dans le but de rapprocher des problèmes et des auteurs de solutions, des auteurs de solutions entre eux, et des solutions et d'autres entités Web liées à l'innovation, quel que soit le cadre de ce rapprochement.

1.2. Problématique scientifique

Si la nécessité de faciliter l'innovation sur le Web a donné un cadre d'application à notre travail, le cadre scientifique fut inspiré par l'évolution du Web vers la convergence du Web Social et du Web Sémantique (voire Chapitre 2), devenue très évidente à l'époque de démarrage de nos travaux. Cette convergence impose la perception du Web comme un immense graphe comprenant des entités (e.g. des représentations des individus, des types différents de contenu, des évènements) liées entre elles à la fois par des liens explicites et par les interactions sociales ayant lieu autour de ces entités. Dans cette représentation du Web en tant qu'immense graphe, nous nous sommes intéressés aux distances qui séparent les entités entre elles (e.g. les participants dans le processus d'innovation, tels que les auteurs de solutions, les problèmes, et les solutions), qui empêchent des rapprochements pertinents qui pourraient mener à la découverte de solutions innovantes.

C'est en cherchant à effectuer dans le graphe les raccourcis qui pourraient faciliter l'émergence de l'innovation sur le Web que l'on pourra proposer des applications utiles et innovantes pour les différents acteurs du processus d'innovation. Cette démarche permet également de proposer des améliorations des outils déjà utilisés dans les différentes phases de ce processus.

1.2.1. Traitement du graphe Web en vue des découvertes pertinentes dans un processus d'innovation

Si l'intensité de l'activité sociale des utilisateurs du Web a suscité une abondance de données relatives à ces utilisateurs et si les nouveaux formats de représentation de données issus des travaux sur le Web Sémantique ont rendu le traitement de ces données plus facile, la navigation dans cet espace d'information souffre toujours d'une grande complexité. Un internaute, dans sa navigation quotidienne, peut toujours avoir l'impression d'avoir parcouru la plupart de chemins pertinents, d'être bien informé et pourtant rater une occasion de soumettre ses propres recherches comme une solution potentiellement gagnante à un problème d'innovation.

Dans nos travaux nous essayions de traiter le graphe du Web en « raccourcissant » ces chemins pour faciliter l'accès à l'information pertinente. Chaque type de chemin dans le graphe exige un traitement particulier selon sa nature et selon le sens qu'il représente. Ainsi, il est par exemple possible de déduire, dans un graphe où un utilisateur est connecté à un objet Web qu'il a créé, et dans la mesure où cet objet est lui-même en relation directe avec les centres d'intérêt de l'utilisateur, qu'il existe un lien implicite entre l'utilisateur et le sujet sur lequel s'est porté son intérêt en créant cet objet. Deux sujets d'intérêt peuvent, de la même manière, être rapprochés, sur la base des liens qu'ils partagent, avec d'autres sujets, catégories ou objets Web. En partant de ce principe, il est possible d'envisager de rapprocher deux utilisateurs en fonctions de leur graphe d'intérêts et/ou de leur graphe social. Nous nous sommes notamment intéressés :

- aux différentes manières permettant d'introduire les raccourcis utiles lors du traitement du graphe Web dans les différentes phases du processus d'innovation sur le Web ;
- à l'inclusion des éléments issus de la pratique sociale de l'utilisation du Web dans le graphe et leur usage dans le rapprochement des acteurs et éléments du graphe dans des scénarios liés à l'innovation ;
- à l'utilité des bases de connaissance, déjà présentes sur le Web dans des formats riches en sémantique (dans le sens du Web Sémantique), dans le cadre de traitement du graphe et notamment lors du raccourcissement de chemins.

1.2.2. Sérendipité et l'aspect inattendu de la navigation dans le graphe de ressources Web

Plus le Web grandit, plus il est difficile d'extraire des informations utiles du volume croissant de données qui alimentent le Web. La surabondance d'information (en. *information overload*) peut rendre l'utilisateur mal informé. La lutte contre ce phénomène est aujourd'hui au cœur des efforts des créateurs des applications Web, comme les moteurs de recherche et les réseaux sociaux, qui visent à satisfaire les besoins informationnels de leurs utilisateurs et à rendre leur navigation dans le Web facile, agréable et efficace. Ces applications mettent en place des filtres, des systèmes de recommandation et de recherche pour séparer les informations utiles pour un utilisateur, dans sa démarche à un instant donné, de l'ensemble des informations disponibles. Ces approches visent majoritairement à mettre en avant des informations que l'utilisateur s'attend à trouver, en les classant comme les plus utiles. Néanmoins, le besoin d'un utilisateur d'informations inattendues, qui peuvent l'enrichir et lui permettre de découvrir et explorer des pistes inhabituelles, reste incontestable et pourtant peu satisfait.

Si l'aspect inattendu de la navigation dans le graphe de ressources du Web est utile de manière générale, c'est dans les scénarios liés à l'innovation où il devient indispensable. L'histoire connaît beaucoup de découvertes scientifiques qui ont émergé de manière inattendue, parfois même sans intention explicite de faire cette découverte, et en exploitant des pistes à priori tenues pour être non pertinentes pour la découverte faite. Pour ne citer que quelques exemples très connus: l'Amérique fut découverte par Christophe Colomb en cherchant un nouveau chemin vers l'Asie, les vertus de la pénicilline dans la lute contre les bactéries furent découvertes par hasard dans le

laboratoire d'Alexandre Fleming où les champignons avait attaqué ses cultures bactériennes abandonnées. Ce phénomène fut distingué sous le nom « sérendipité » par Horace Walpole en 1754 [Andel & Bourcier, 2009]. Afin de soutenir la recherche et l'innovation sur le Web, nous devons donc favoriser une telle navigation qui favorise la découverte fortuite et une ouverture d'esprit vers des informations dont la pertinence, plutôt latérale, est peu évidente. Par latéralement pertinentes nous comprenons des idées qui ne sont pas directement liées à un chemin de pensé de départ, mais qui sont quand même pertinentes.

Nos raccourcis dans le graphe Web nécessitent donc également ce caractère fortuit. A titre d'exemple, un internaute cherchant un collaborateur, à priori informaticien, pour résoudre un problème lié à la conception d'un outil d'aide à la décision peut collaborer utilement avec un juge dont les modes de prise de décisions peuvent être transposés à la problématique en question. Même si collaborer avec quelqu'un qui n'est pas expert en informatique où logique ne paraît pas évident au départ, une telle collaboration pourrait offrir une nouvelle perspective sur le problème car le processus de prise de décision juridique et en effet très lié aux mécanismes de raisonnement.

La question clé de cet axe de recherche peut être résumé par : comment, dans le cadre d'un démarche liée à l'innovation, intégrer un caractère sérendipique (à la fois inattendu et pertinent) à la navigation d'un utilisateur dans un espace conceptuel du graphe Web, qui favoriserait des découvertes et la latéralité de la pensée ?

1.2.3. Modélisation des entités Web liées à la résolution de problèmes et l'innovation sur le Web et l'échange des données entre des systèmes hétérogènes

Enfin, si une importante partie de nos travaux utilise le graphe sémantique et social du Web, une autre partie permet de introduire des résultats de nos transformations et rapprochements dans le Web Sémantique. De cette manière l'interaction avec le graphe Web porte un caractère enrichissant à la fois pour l'utilisateur et pour le Web.

Les modélisations des différents paradigmes essentiels à l'innovation sur le Web nous ont permis d'envisager la création des nouveaux liens dans le graphe du Web en permettant l'échange d'information de manière précise, structurée et riche de sens. Par modélisation, nous entendons à la fois l'établissement des modèles conceptuels des notions introduites par nos travaux, et leurs alignements avec des modèles déjà utilisés dans le Web Sémantique, pour transmettre la connaissance relative aux entités liées à l'innovation. Cette pratique est souvent complétée par la mise à disposition des données sociales, issues de la pratique quotidienne des utilisateurs, dans un format riche de sémantique leur permettant d'être traitées de manière efficace et distribuée.

Si nos efforts de modélisation sont principalement motivés par la nécessité d'enrichir le Web, ils sont inspirés également par la nécessité de permettre une interopérabilité entre des applications Web. Un processus d'innovation qui se déroule sur le Web est, par sa nature, forcement distribué. Les différentes phases de ce processus ne sont pas seulement effectuées par les différents acteurs, mais ont lieu également dans les différentes applications. Ces applications doivent être élaborées de manière à échanger les informations sans perte de sens afin de permettre une continuité du processus et la réutilisation des résultats de chaque phase.

* * *

Si nous devions résumer la problématique scientifique abordée dans cette thèse en une seule phrase, on pourrait formuler : *Comment effectuer des traitements et des transformations du graphe de ressources du Web afin de stimuler la convergence de l'utilisateur vers des éléments informationnels du Web qui lui sont utiles et qui représentent une découverte valable, tout en s'appuyant sur la nature sémantique et sociale de ce graphe et en permettant la collaborations des machines et des humains dans la mise en place de telles convergences*. Cette problématique est notamment étudiée dans le contexte de la pratique émergeante de l'innovation sur le Web, qui emploie des technologies du Web pour accélérer des échanges entre les principaux acteurs d'un processus d'innovation.

1.3. Notion d'expertise et de compétence dans le cadre de nos travaux

La notion d'expertise est très importante dans le contexte d'innovation dans lequel s'inscrivent nos travaux. Il est donc nécessaire de bien la définir dès le départ, notamment en raison de différents usages que ce terme a pu avoir dans les travaux issus de différentes disciplines et époques, qui peuvent être sources de malentendu.

Beaucoup de travaux existants issus des domaines tels que l'apprentissage en ligne ou la gestion des ressources humaines, font un effort pour définir la notion de l'expertise et de la compétence dans les systèmes informatiques et notamment pour chercher une définition adaptée à leurs travaux. La question d'une définition adaptée à notre cadre de travail s'impose également à nous. Dans cette section nous présentons une analyse des définitions existantes proposées dans les travaux liées. Nous présentons d'abord une définition générale de la compétence. Ensuite nous identifions des parties de cette définition générale pertinentes pour différents domaines d'utilisation et enfin, nous présentons une définition, adaptée au cadre de l'innovation ouverte sur le Web, que nous utiliserons par la suite dans ce mémoire.

1.3.1. Définition composée de la compétence

La définition composée est donnée sous forme d'un graphe (Figure 3). D'après cette définition nous entendons par compétence une qualité (connaissance, capacité, aptitude, comportement, savoir-faire, expérience ou correspondant à un objectif d'apprentissage) possédée par une personne (dans les perspectives différentes cette personne peut agir en tant que ressource humaine, i.e, un étudiant, un résolveur de problèmes, etc.) qui est nécessaire pour effectuer une action (atteindre un objectif, contrôler une situation, jouer un rôle, exercer un métier, etc.) avec un certain niveau de performance, dans un contexte (ou un environnement, ou un marché) à l'aide de certaines ressources, afin d'atteindre un objectif. Comme résultat secondaire, la personne gagne en expérience.

La présente définition définit la compétence au sens large. Elle se base sur d'autres définitions de la compétence, citées ci-dessous, mais aussi sur des définitions proposées pour des cas particuliers abordés dans les travaux empiriques. Notre analyse de l'état de l'art étant fondée principalement sur des travaux faits en informatique rend la généralisation que nous avons pu faire limitée en terme d'exhaustivité. Néanmoins une

telle généralisation reste utile pour bien situer notre propre définition de la compétence dans un cadre plus général

Parmi des définitions au sens large, nous pouvons citer celles de Radevski et Trichet [Radevski & Trichet, 2006] qui définissent la compétence comme la combinaison de qualités (connaissances, savoir-faire, etc.) dans un contexte précis, afin d'atteindre un objectif ou accomplir une mission. Les mêmes auteurs, [Radevski & Trichet, 2006a] ont défini la compétence comme capacité de contrôler une certaine situation, d'innover et de trouver de nouveaux chemins de pensée. Cette capacité émerge d'une connaissance, d'un savoir-faire et d'un comportement spécifique.

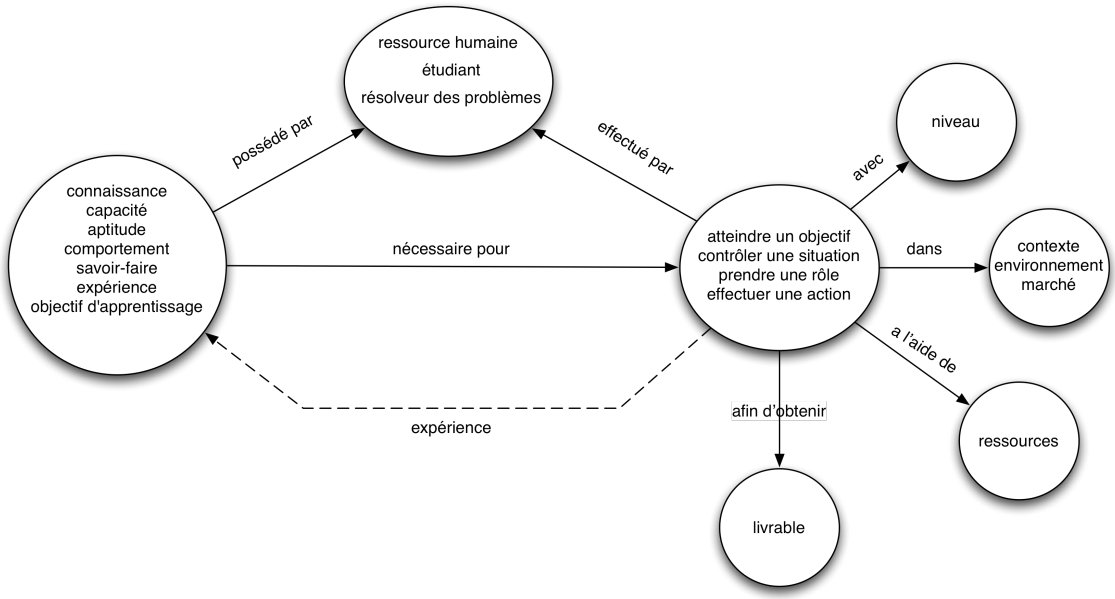


Figure 3 Définition composée de la compétence

Les définitions dans les dictionnaires portent aussi souvent sur le sens large. Le dictionnaire TLFi[10] propose une définition : "Capacité que possède une personne à porter un jugement de valeur dans un domaine dont elle a une connaissance approfondie. " D'après Wikipedia[11], la compétence est : *" est une connaissance (savoir, savoir-faire, savoir-être) mobilisable, tirée généralement de l'expérience et nécessaire à l'exercice d'une activité. Dans ce cas, professionnelle. La notion de savoir renvoie évidemment aux connaissances initialement et ultérieurement acquises indispensables à l'exercice de la profession. La notion de savoir procédural est liée à ce qui est commun au métier, à la profession. L'expérience, le tour de main de l'ouvrier par exemple, est liée à la singularité de l'individu, à sa pratique dans l'exercice de l'activité et en relation avec d'autres, au sein de l'équipe, face aux clients. Pour de nombreux spécialistes du travail, la compétence est un mélange des trois composants que sont les connaissances ou savoirs théoriques et procéduraux, les savoir-faire ou habiletés et les savoirs pratiques ou qualités personnelles."*

D'autres définitions se focalisent sur des aspects particuliers. D'après Gandon et al. [Gandon et al., 2005] elle comprend une action, un livrable (un résultat de l'action), un marché (pour lequel les résultats de cette action sont produits) et des ressources

[10] http://atilf.atilf.fr/tlf.htm

[11] http://fr.wikipedia.org/wiki/Comp%C3%A9tence_(ressources_humaines) , 03/08/2012

(utilisées pour atteindre l'objectif de l'action). Coi et al. [Coi et al. , 2007] définissent la compétence comme la performance effective dans un contexte/domaine avec un niveau de performance.

Des définitions spécifiques de la compétence ont notamment été proposées pour les domaines de la formation en ligne et de la gestion des ressources humaines.

1.3.2. Perspective de la formation en ligne

La perspective de la formation en ligne est notamment prise par le standard IEEE RCD [IEEE, 2000]. Dans ce standard, la compétence est définie comme une forme de connaissance, d'aptitude, de capacité ou d'objectif d'apprentissage qui peut être décrit dans le contexte de l' apprentissage, de l'éducation et de l' entrainement.

Dans ce contexte d'apprentissage, la personne concernée par la compétence est un étudiant. Le contexte d'usage de la compétence est ici moins important que le contexte de son acquisition. La motivation pour l'exprimer n'est donc pas seulement de trouver des personnes avec une certaine compétence, mais aussi de trouver des moyens pour acquérir la compétence nécessaire.

1.3.3. Perspective de la gestion des ressources humaines

Caetano, Pombinho & Tribolet [Caetano et al., 2007] perçoivent les compétences comme une synergie de qualités (connaissances, savoir-faire) et d'actions. C'est un comportement des individus nécessaire dans une situation pour atteindre un niveau de performance.

Draganidis et al. [Draganidis et al., 2006], ainsi que Cristiaens, De Bo & Verlinded [Christiaens et al., 2006] adoptent la définition du consortium chargé du standard HR-XML. Pour eux, la compétence est une connaissance mesurable, une capacité, une aptitude ou autre caractéristique liée à l'embauche, qu'une ressource humaine possède et qui est nécessaire pour effectuer une activité dans un contexte spécifique.

Biesalski & Abecker [Biesalki & Abecker, 2006] définissent la compétence par la capacité de trouver des solutions à certains problèmes. La compétence est nécessaire pour effectuer certaines tâches et pour prendre certains rôles et positions dans une organisation.

Toutes ces définitions considèrent un contexte professionnel. La personne concernée par la compétence est un employé (ou un candidat pour une poste de travail). L'utilité de la compétence est toujours de servir aux objectifs d'une organisation.

1.3.4. Perspective de la résolution de problèmes et de l'innovation

Dans la perspective de la résolution des problèmes, ce qui importe est la capacité de trouver des solutions : de comprendre un problème et son contexte, et de pouvoir se servir des connaissances, du savoir-faire, de l'expérience et de l'intelligence pour proposer de bonnes solutions. Dans cette perspective, l'objectif i.e. la résolution des problèmes, est un élément primordial dans la définition de la compétence. Contrairement à la perspective dans le domaine des ressources humaines, la

compétence n'est pas exigée pour prendre des rôles et positions stables dans une organisation. L'important est ce qu'une personne peut apporter à la résolution d'un problème. L'engagement est de courte durée – la personne investit plutôt son intelligence et ses idées avec un peu de travail occasionnel, alors que dans la perspective des ressources humaines, la personne devient partie prenante d'un projet à long terme ou d'un organisme.

Par l'importance qui est donnée aux objectifs, la définition de Gandon et al. [Gandon et al., 2005] est très proche des besoins de cette perspective, mais elle ne suffit pas, puisque elle ne prend pas en compte la spécificité de la résolution des problèmes. Ce qui importe n'est pas seulement le résultat. Une idée simple, qui n'atteint pas l'objectif en ne donnant pas le résultat demandé, est aussi valable dans les plates-formes de la résolution des problèmes. Une idée peut aider à reformuler le problème, le résoudre en identifiant les problèmes similaires ou connexes, les contraintes critiques dont la résolution dépend, une idée plus précise de l'état de l'art, etc. Ce n'est pas seulement le résultat satisfaisant pour le problème posé qui peut être une contribution valable d'un résolveur de problèmes. Il est donc parfois nécessaire de trouver une définition mieux adaptée aux scénarios pertinents pour les plates-formes de résolution des problèmes.

Nous proposons la définition suivante :

La compétence est l'ensemble des connaissances, des aptitudes, des capacités, du savoir-faire et des expériences, nécessaires ou utiles pour comprendre un problème et son contexte (y compris les contraintes) et pour contribuer à la proposition de solutions correspondant aux attentes, aux livrables demandés. La compétence permet également de mieux poser le problème, de le reformuler, voire de le résoudre par la solution à un problème connexe ou corolaire si celle-ci est transposable.

Il est donc important de souligner que cette définition ne limite pas la compétence aux qualités liées au domaine du problème. Il est donc possible d'envisager la contribution d'un chimiste dans un problème de physique.

Dans ce mémoire nous utilisons les termes suivants : experts, (potentiel) auteur de solutions, porteur d'idées etc. pour désigner une personne compétente pour un problème au sens de cette définition. Autrement dit, un expert est capable de participer, de manière utile, dans la résolution de problème, quelle que ce soit sa formation, son parcours et sa perspective sur le problème lui-même. Une perspective différente de celle tenu par l'auteur du problème est dans nos travaux considérée comme un avantage, plutôt qu'un inconvenant.

1.4. Organisation du mémoire et vision synthétique des résultats

1.4.1. Guide de lecture

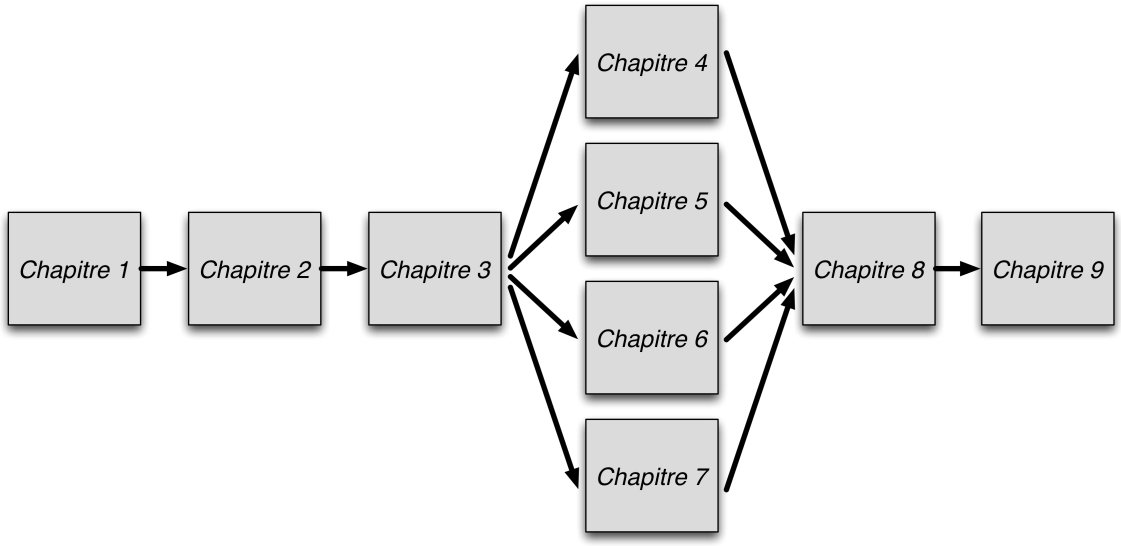


Figure 4 Organisation du mémoire

La thèse est organisée en neuf chapitres dont les dépendances principales sont représentées dans la Figure 4. Le Chapitre 1 a pour objectif d'introduire le lecteur dans les problématiques abordées et de lui présenter le contexte des travaux. Le Chapitre 2 détaille le contexte technique du Web Social Sémantique. Nous découvrons notamment les notions clés à connaître et les technologies existantes qui ont fortement contribué à l'avancement de nos propres travaux. Dans le Chapitre 3, nous définissons le cadre conceptuel qui généralise tous nos travaux. Ce cadre s'appuie largement sur l'idée de raccourci dans le graphe de ressources Web, notion que nous avons évoquée dans la description de la thématique. Les trois premiers chapitres présentent les bases nécessaires pour la compréhension de la thèse.

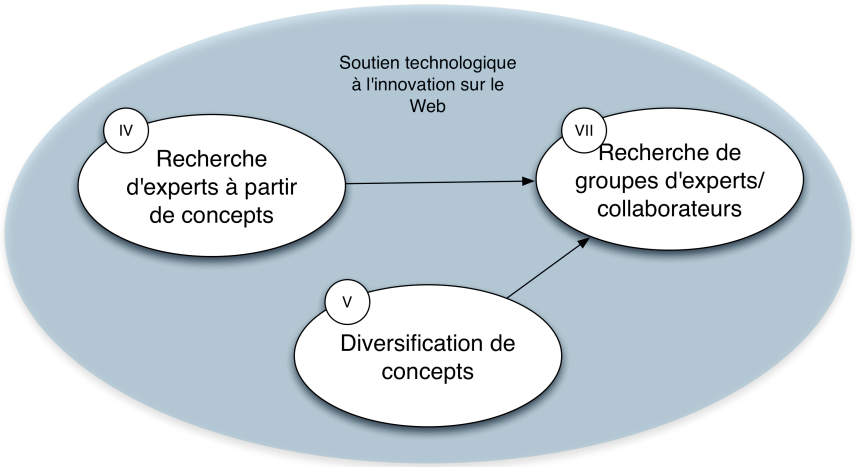


Figure 5 Articulation de chapitres clés

Les chapitres 4 à 7 présentent le cœur de la thèse i.e. les réalisations de différents rapprochements applicables aux graphes de données structurées dans le contexte de l'innovation sur le Web. La Figure 5 illustre l'articulation de ces chapitres clés. Le Chapitre 4 concerne la recherche d'experts à partir d'un ensemble de concepts – sujets de compétence, initialement définis. Dans le Chapitre 5 nous abordons la problématique de la diversification d'un ensemble de concepts donné pour y inclure des concepts pertinents et inattendus, et ce afin d'enrichir les activités de recherche d'experts qui s'appuient sur cet ensemble de départ. Le Chapitre 7 concerne la recherche de collaborateurs susceptibles de répondre aux besoins de la résolution de problèmes pluridisciplinaires. Ces travaux et leur synergie contribuent à l'émergence de l'innovation via le Web. S'ils peuvent à la fois être étudiés indépendamment les uns des autres, ils forment un ensemble cohérent qui les rend véritablement utiles dans notre cas d'utilisation. Le Chapitre 6 présente un travail auxiliaire qui concerne le cas particulier de l'alignement de ressources Web, notamment utiles dans plusieurs scénarios liés à l'innovation. Ce chapitre peut également être perçu comme un prélude pour le Chapitre 7 car il introduit certaines approches reprises dans celui-ci. Le Chapitre 8 présente nos travaux sur la représentation des connaissances liées au domaine d'innovation sur le Web avec la présentation de différentes ontologies. Le Chapitre 9 est le chapitre de conclusion. Il donne des idées sur des pistes à explorer dans des travaux futurs.

Il est important de noter que la revue de l'état de l'art est distribuée dans les chapitres qui concernent les réalisations techniques (donc chapitres 4-7). Au lieu de consacrer un seul et unique chapitre à l'état de l'art, nous avons fais le choix de diviser cette partie en plusieurs sections, situées dans différents chapitres, pour faciliter la lecture et la compréhension de la thèse.

Au début de ce mémoire le lecteur trouvera un glossaire de termes introduits dans cette thèse ou simplement employés avec un sens particulier. Nous espérons que celui-ci facilitera la compréhension du texte.

1.4.2. Résultats et contributions par chapitre

Dans la partie précédente, nous avons décrit l'organisation principale de la thèse et la séparation des chapitres. Nous allons à présent nous intéresser aux résultats et aux contributions majeures apportées dans chaque chapitre.

Chapitre 1 – Introduction

Dans ce premier chapitre, le lecteur se familiarise avec le contexte de la thèse, le cadre institutionnel et avec la problématique scientifique. Nous y présentons les défis qui s'imposent à l'heure actuelle face à des pratiques de plus en plus courantes comme la résolution de problèmes et l'innovation sur le Web. Les axes de recherche, centrés autour du soutien d'une telle pratique grâce à des technologies d'information, ont été identifiés et systématisés.

Chapitre 2 - Web Social Sémantique

Le Chapitre 2 a pour but d'initier le lecteur aux notions et au vocabulaire utilisés dans le domaine du Web. Nous y décrivons notamment les paradigmes du Web, du Web Social

et du Web Sémantique, tout en abordant le contexte historique et technique de leur naissance et de leur évolution. Bien que ces notions ne soient pas nouvelles, les définitions données présentent tout de même un point de vue nouveau, plus moderne et adapté à la situation actuelle. Nous découvrons notamment une convergence du Web Social et du Web Sémantique, baptisé le Web Social Sémantique. Il représente le socle fondateur de nos travaux. Enfin, nous identifions dans ce chapitre les tendances qui émergent dans le développement du Web, ainsi qu'une vision du Web du futur.

Dans un domaine tel que le nôtre, l'analyse de ces tendances est importante car, contrairement à l'étude des phénomènes stables, le Web est en évolution constante, et il l'était au moment de l'écriture de cette thèse. Il est donc particulièrement important de pouvoir observer les tendances afin de parsemer nos choix de pistes de recherche d'un peu de pragmatisme.

Chapitre 3 - Cadre conceptuel: Rapprochement des éléments Web liés à l'innovation

Le Chapitre 3 apporte une introduction aux notions et paradigmes que nous avons croisés lors de nos travaux. Nous y présentons le cadre conceptuel qui généralise nos travaux. Ce cadre découle d'un travail de généralisation important, et va au delà de ce qui a été implémenté dans cette thèse. Afin de mieux comprendre cette généralisation dans nos approches, nous proposons un modèle d'interaction utilisateur-Web. Ce modèle simple permet de comprendre l'origine et la structure d' informations qui constituent le Web. La structure du Web regorge de ressources particulièrement importantes pour les scénarios liés à l'innovation, comme par exemple les problèmes, les auteurs de solutions, les contenus générés par des utilisateurs etc. C'est dans ce chapitre que nous introduisons la notion de **trace utilisateur** – ressources qui émergent de l'activité des utilisateurs sur le Web. Ces traces permettent d'établir un lien entre un utilisateur Web et un concept, sujet de compétence, qui concerne la ressource identifiée comme trace.

Le cadre conceptuel que nous étudions favorise le rapprochement de ressources Web afin de raccourcir les chemins d'accès de manière utile dans les scénarios liés à l'innovation. Cette notion du **rapprochement** est donc une notion clé dans nos travaux. Un rapprochement de ressources dans le graphe Web a pour but principal de déclencher des interactions entre des acteurs et ressources pertinentes pour un processus d'innovation et ainsi faire émerger des solutions innovantes. La notion de rapprochement qui donne un cadre unificateur à nos travaux nous a permis d'exposer nos hypothèses de recherche à un niveau générique. Celles-ci sont concrétisées dans les chapitres suivants, en fonction des aspects abordés dans chacun d'entre eux. Le Chapitre 3 présente également une introduction aux problématiques abordées par des chapitres suivants et permet au lecteur d'obtenir une image globale de la totalité des travaux et des différents aspects de la problématique globale traités par chaque partie.

Si le cadre unificateur présenté dans ce chapitre est sans doute utile pour la compréhension de nos travaux, il est aussi un outil de réflexion permettant de situer nos travaux par rapport aux autres et d'identifier d'autres pistes de recherche que nous pourrions explorer dans l'avenir.

Chapitre 4 - Rapprochement des experts et concepts de leur expertise

Dans le Chapitre 4, nous abordons le scénario de recherche d'experts, notamment à partir des sources de données sémantiquement structurées[12]. Notre travail se focalise sur le rapprochement de ressources Web décrivant des utilisateurs, des experts potentiellement auteurs de solutions, et des concepts associés à leurs centres d'intérêts et à leur expertise. Nous introduisons la notion d'**hypothèse d'expertise** afin de généraliser les nombreuses manières d'identifier les sujets d'expertise d'un expert. Cette notion correspond à une règle d'inférence utilisée dans une approche de recherche d'experts. Elle transforme les données de départ (e.g. traces utilisateurs) en informations sur l'expertise. Nous présentons une systématisation de l'état de l'art selon l'hypothèse d'expertise utilisée. Dans ce chapitre, nous explicitons nos travaux suivants :

- Un cadre d'évaluation de la capacité du Web actuel à s'adapter à chacune de nos hypothèses d'expertise. Nous avons implémenté quatre tests différents dans le but de valider ces éléments.

  Ces tests permettent d'évaluer la faisabilité d'une démarche particulière de recherche d'experts en s'appuyant sur des données structurées disponibles sur le Web. Ils constituent une phase préalable nécessaire à nos autres travaux.

- L'identification des problèmes techniques qui peuvent survenir lors de l'utilisation du graphe Web de données dans le cadre de recherche d'experts ainsi que la proposition de solutions pour faire face à ces problèmes.

  Ces solutions et heuristiques de normalisation de données, souvent hétérogènes, rendent possible l'utilisation de données structurées du Web dans le processus de recherche d'experts.

- La recherche de bonnes pratiques de publication de données structurées, utilisables dans le cadre de la recherche d'experts.

  Les bonnes pratiques de publication de données viennent en complément des heuristiques de normalisation. Ces deux axes contribuent à rendre le Web le plus exploitable pour la tâche de recherche d'expert.

- L'implémentation d'une approche de suggestion d'hypothèse d'expertise à partir d'un sujet donné. Cette approche est fondée sur le fait que des communautés différentes utilisent différents types de canaux de communication et laissent des traces différentes sur le Web. Nous nous sommes appuyés sur la richesse structurelle du Web Social Sémantique pour détecter ces tendances et suggérer des hypothèses d'expertise appropriées.

  La suggestion des hypothèses d'expertise rendue possible par la diversité de types de traces présentes sur le Web Social Sémantique, apporte de la souplesse à la démarche de recherche d'experts et la rend adaptable aux contextes et spécificités de différentes disciplines. Cette approche représente la contribution clé du chapitre.

[12] Par les données sémantiquement structurées nous considérons des ensembles de données mis à disposition sous forme d'un graphe sémantique, et souvent liées aux autres ensembles de données sur le Web (en. Linked Data). Nous définissons cette notion plus en détail dans le chapitre 2. Par simplicité, nous utilisons souvent "données structurées" ou "données liées" pour cette notion dans le reste du mémoire.

- L'implémentation d'un système de recherche d'experts, hy.SemEx, qui utilise les données structurées trouvées sur le Web public et qui applique notre approche de suggestion d'expertise de manière appropriée sur un sujet donné.

  Le système hy.SemEx est l'implémentation technique de notre approche et représente la cristallisation de nos travaux présentés dans ce chapitre. Il nous a permis de nous assurer de la pérennité de l'approche et de la faisabilité d'une démarche de recherche d'experts à l'aide de données structurées issus du Web public.

Les travaux présentés dans ce chapitre jouent un rôle très important dans la réalisation de notre objectif, celui de faciliter l'émergence de l'innovation via le Web. Ils seront complétés pas les travaux présentés dans les chapitres qui suivent.

Chapitre 5 - Rapprochement de concepts

Dans le Chapitre 5, nous traitons le rapprochement de concepts, sujets de compétence. L'objectif de ce rapprochement est de permettre la découverte de concepts pertinents et inattendus à partir d'un ensemble de concepts de départ. Si nous imaginons un problème d'innovation qui concerne certains concepts, il est sans doute utile de découvrir d'autres concepts liées aux concepts de départ qui peuvent permettre une ouverture d'esprit vers d'autres champs de compétence à partir desquelles les solutions innovantes sont susceptibles d'émerger. C'est par cette découverte de concepts que nous souhaitons stimuler la sérendipité dans l'ensemble des situations liées à l'innovation sur le Web.

En plus de l'analyse d'un grand nombre d'approches de recommandation de concepts, issues de l'état de l'art, dans ce chapitre nous présentons les contributions suivantes :

- Une méthode de rapprochement de concepts, adaptée à la nature du graphe de données sur le Web et qui répond aux besoins spécifiques des scénarios liés à l'innovation. Notre approche propose à la fois une notion de proximité dans le graphe de données sémantiques, nommée **hyProximity**, et un algorithme permettant de mettre en œuvre le calcul de la proximité. Plusieurs variations de l'approche sont présentées et comparées.

  HyProximity est la contribution clé de ce chapitre. Elle sera notre outil principal, pour la diversification de champs de compétences étudiés dans nos diverses démarches de recherche et de recommandation

- Plusieurs évaluations et comparaisons de performance entre hyProximity et autres approches de rapprochement de concepts issues de l'état de l'art.

  Les évaluations que nous avons effectuées ont permis de s'assurer des bonnes performances de notre approche, mais aussi d'étudier son utilité dans plusieurs contextes.

- Le prototype que nous avons développé, utilisé de manière continue au sein d'Hypios, nous a également permis d'évaluer son utilité.

  Ce prototype a confirmé la faisabilité de notre approche de recommandation de concepts qui utilise un graphe de données structurées sémantiques. Cette validation porte sur l'utilité de la proposition de concepts ainsi que sur les performances de notre algorithme.

Les travaux présentés dans ce chapitre sont complémentaires de ceux présentés dans le Chapitre 4, et sont réutilisés dans le cadre de travaux présentés dans le Chapitre 7 pour assurer la diversité dans un cas spécifique de la recherche d'experts collaborateurs.

Chapitre 6 - Rapprochement de traces utilisateurs

Le Chapitre 6 approfondit l'idée de rapprochement de traces utilisateurs. Les contributions présentées dans ce chapitre sont principalement :

- La mise en valeur d'une idée de rapprochement des articles scientifiques et des tweets[13], idée ré-exploitée à de nombreuses reprises par la communauté de recherche suite à notre publication en 2010.

  Ce rapprochement permet d'augmenter l'utilité de tweets, dont l'usage est de plus en plus fréquente au sein des communautés scientifiques, pour l'identification de compétences des utilisateurs et donc pour la recherche d'experts. Il vient en complément de nos travaux présentés dans le Chapitre 4.

- Le cadre conceptuel de sémantisation de tweets et de rapprochement entre tweets et articles est appelé RATED. En ce qui concerne la mise en relation entre les tweets et les articles, nous décrivons les solutions possibles que nous avons étudiées en collaboration avec le KMI (Knowledge Media Institute), au Royaume Uni.

  Ce cadre conceptuel met en œuvre le rapprochement entre tweets et articles scientifiques. Dans le contexte de la présentation de ce cadre, nous présentons également des méthodes de traitement de tweets qui sont utilisées dans le Chapitre 7.

- Au travers de l'analyse des centres d'intérêt des utilisateurs de Twitter et de la popularité des tweets générés lors des conférences académiques, nous avons mesuré l'utilité des correspondances tweet-article et leur valeur ajoutée par rapport au simple usage des tweets ou des articles.

  L'étude de l'utilité des alignements tweet-article nous a permis de nous assurer de la valeur de notre démarche dans des situations réelles.

Les travaux présentés dans ce chapitre ont un caractère complémentaire et viennent en complément des piliers de la thèse présentés dans les chapitres 4,5 et 7.

Chapitre 7 - Rapprochement de collaborateurs dans les scénarios d'innovation sur le Web

Le Chapitre 7 explore la notion de rapprochement complexe. De manière générale, il consiste à rapprocher de potentiels collaborateurs dans le contexte d'un problème à résoudre conjointement. Dans ce contexte nous présentons les contributions suivantes :

- Une approche générale pour la recommandation de collaborateurs potentiels pour un utilisateur qui s'est déjà engagé dans un processus de résolution de problème en ligne. Notre approche cherche à mettre en avant des collaborateurs

[13] Les tweets (autrement dits billets de microblog) sont des messages courts, destinés au grand public ou bien à des groupes d'utilisateurs ciblés, que les utilisateurs partagent dans les sites Web sociaux. Ce concept sera défini plus en détail dans le Chapitre 2.

qui sont à la fois complémentaires avec l'utilisateur en termes de compétence, et capables de bien s'entendre et de bien travailler ensemble.

Ce rapprochement complète notre approche de recherche d'experts présentée dans le Chapitre 4 dans la mesure où il met en relation des experts avec de bons collaborateurs, leur permettant ainsi d'affronter les problèmes pluridisciplinaires. Il répond à un besoin réel, celui de l'innovation en collaboration, souvent vu comme indispensable dans les démarches industrielles.

- Plusieurs méthodes d'extension de profils utilisateurs pour mieux cibler les recommandations. Nous présentons aussi plusieurs évaluations de notre approche et de ses différentes versions alternatives.

  L'extension de profils permet de mettre en valeur notre approche de découverte de concepts, hyProximity, dans le cadre de l'élargissement des profils d'experts, et de comparer ses performances à celles des approches issues de l'état de l'art.

Les travaux présentés dans ce chapitre combinent les expériences acquises lors de la réalisation des travaux précédents, et notamment ceux présentés dans les chapitres 5 et 6 pour mettre un œuvre un rapprochement complexe et véritablement utile dans des scénarios d'innovation réels.

Chapitre 8 - Ontologies pour faciliter les rapprochements

Dans le Chapitre 8, nous présentons des approches de modélisation des éléments introduits en premier lieu dans le Chapitre 3, puis détaillés dans les chapitres suivants. Ces éléments sont essentiels dans notre étude du cadre conceptuel. La modélisation concerne à la fois les ressources Web qui sont sujets à des rapprochements, et les rapprochements eux-mêmes. La description complète de ces derniers est importante, notamment pour le partage et la réutilisation des résultats de rapprochements passés dans un autre processus de rapprochement. Nous décrivons un tel scénario à la fin du Chapitre 8.

Chapitre 9 - Conclusion

Le Chapitre 9 apporte les conclusions générales qui ressortent de cette thèse, et notamment les évaluations et la validation des hypothèses de départ. Nous abordons aussi la question de la généralisation de nos travaux à des situations plus variées, et réfléchissons à d'autres domaines d'application. Pour finir, nous suggérons des idées de pistes pour de futurs travaux.

Bibliographie

[Andel & Bourcier, 2009] Andel, P. V., & Bourcier, D. (2009). *De la sérendipité dans la science, la technique, l'art et le droit : Leçons de l'inattendu*. L'Act Mem. Retrieved from http://www.amazon.fr/sérendipité-dans-science-technique-droit/dp/2355130183

[Berners-Lee, 2000] Berners-Lee, T. (2000). *Weaving the Web*. San Francisco, CA, USA. Harper San Francisco.

[Biesalki & Abecker, 2006] Biesalski, E., & Abecker, A. (2006). Similarity measures for skill-profil matching in enterprise knowledge management. *8th International Conference on Enterprise Information Systems (ICEIS)*, 23–27. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Similarity+measures+for+skill-profil+matching+in+enterprise+knowledge+management#0.

[Bush, 1945] Bush, V. (1945). As We May Think. *The Atlantic Monthly*, *18*(3). Retrieved from http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/3881/

[Caetano et al., 2007] Caetano, A., Pombinho, J., & Tribolet, J. (2007). Representing organizational competencies. *Proceedings of the 2007 ACM symposium on Applied*. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Representing+Organizational+Competencies#3.

[Chesbrough, 2003] Chesbrough, H. W. (2003). *Open Innovation: The New Imperative for Creating and Profiting from Technology*. Harvard Business Press. Retrieved from http://www.amazon.com/Open-Innovation-Imperative-Profiting-Technology/dp/1578518377

[Christiaens et al., 2006] Christiaens, S., De Bo, J., & Verlinden, R. (2006). Competency model in a semantic context: Meaningful competencies (position paper). *LECTURE NOTES IN COMPUTER SCIENCE*, *4278*, 1100. Springer. Retrieved from http://www.springerlink.com/index/05657381p7317873.pdf.

[Coi et al. , 2007] Coi, J. D., Herder, E., Koesling, A., Lofi, C., Olmedilla, D., Papapetrou, O., et al. (2007). A model for competence gap analysis. *Proc. of WEBIST*. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:A+model+for+competence+gap+analysis+*#1.

[Draganidis et al., 2006] Draganidis, F., Chamopoulou, P., & Mentzas, G. (2006). An ontology based tool for competency management and learning paths. *Proc. of I-KNOW'06*. Citeseer. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:An+Ontology+Based+Tool+for+Competency+Management+and+Learning+Paths#0.

[Englebart, 1962] Engelbart, D. (1962). Augmenting human intellect: a conceptual framework, *SRI Summary Report AFOSR-3223*. Washington, USA. Retrieved from http://www.dougengelbart.org/pubs/augment-3906.html

[Gandon et al., 2005] Gandon, F., Corby, O., Giboin, A., Gronnier, N., & Guigard, C. (2005). Graph-based inferences in a semantic web server for the cartography of competencies in a telecom valley. *Lecture Notes in Computer Science*, *3729*(3), 247. Springer. Retrieved from http://www.springerlink.com/index/7r648t660w1u023n.pdf.

[IEEE, 2000] IEEE. (2000). Specification of Reusable Competency Definitions, Version 0.4, 0-3. Retrieved from http://ltsc.ieee.org/wg20/files/RCD_0_4.pdf.

[Jeppesen & Lakhani, 2009] Jeppesen, L. B., & Lakhani, K. R. (2009). Marginality and Problem Solving Effectiveness in Broadcast Research. *Organization Science*, *20*. Retrieved from http://dash.harvard.edu/handle/1/3351241

[Nelson, 1965] Nelson, T. (1965). The Hypertext. *Proc World Documentation Federation Conference*.

[Radevski & Trichet, 2006] Radevski, V., & Trichet, F. (2006). Ontology-Based Systems dedicated to Human Resources Management: an application in e-recruitment. *Lecture Notes in Computer Science*, *4278*, 1068–1077. Berlin: Springer-Verlag, Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Ontology-Based+Systems+dedicated+to+Human+Resources+Management:+an+application+in+e-recruitment#0.

[Radevski & Trichet, 2006a] Trichet, F., & Radevski, V. (2006). Competency-Based Systems dedicated to e-Recruitment. *4th International Conference on Computer Science and*

Information Technology (CSIT'2006), Jordanie Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Competency-Based+Systems+dedicated+to+e-Recruitment#1.

2. Web Social Sémantique

> *Ce chapitre définit les notions clés autour desquelles sont centrés nos travaux de recherche. Nous introduisons la notion de Web (§2.1) et ses différentes dimensions, tels que le Web Social (§2.2) et le Web Sémantique (§2.3). Nous expliquons ensuite la distinction importante entre l'état technique du Web, que ces différentes dimensions font progresser, et le Web en tant que paradigme unique du rêve au service de l'Homme connecté. Nous verrons par la suite que ce rêve se concrétise peu à peu par la convergence du Web Social et du Web Sémantique, aussi connu sous le nom de Web Social Sémantique (§2.4). Enfin nous présentons quelques tendances de l'évolution du Web et donnons une vision du Web du futur (§2.5), vision qui a par ailleurs influencé nos choix de recherche.*

2.1. Web

2.1.1. Contexte historique

Si vers la fin des années 80, quand le Web n'existait pas encore, nous avions voulu engager une personne capable de le construire, on n'aurait probablement pas soupçonné que Tim Berners-Lee, physicien de formation, aurait pu le faire. Néanmoins, l'histoire a pris un cours inattendu comme c'est souvent le cas avec les grandes inventions. A cette époque le Conseil Européen de la Recherche Nucléaire (CERN) hébergeait un grand nombre de chercheurs, provenant de différents pays, utilisant différents ordinateurs qui fonctionnaient avec des formats de fichiers différents. Ces chercheurs voulaient travailler ensemble, mais l'échange des fichiers était très complexe suite à la nécessité d'effectuer des conversions entre différents formats. L'idée d'un espace informationnel commun s'imposait naturellement. Tim Berners-Lee était à l'époque chercheur au CERN et il avait compris, qu'un tel moyen d'échanges

d'informations, clairement nécessaire pour mieux connecter ses collègues au CERN, pourrait potentiellement connecter toute l'humanité.

A l'époque de la création du Web, l'Internet existait déjà. Sous la direction de Leonard Kleinrock, la première connexion de deux ordinateurs sur ARPANET[14] a été effectuée entre un ordinateur à UCLA[15] et un autre à l'Institut de Recherche de Stanford (SRI) le 29 octobre 1969 [Gromov, 1995]. ARPANET se fusionnera ensuite avec d'autres réseaux pour former l'Internet. Le réseau global s'est développé d'abord en incitant des universités à se joindre, pour se propager ensuite au grand public dans les années 80. C'est à cette époque-là que les premières entreprises qui fournissaient l'accès commercial à l'Internet ont vu le jour et que l'Internet est devenu accessible à tout le monde. L'infrastructure de ce réseau repose encore aujourd'hui sur des protocoles de transfert de paquets de données, plus précisément sur TCP/IP (*Transmission Control Protocol* et *Internet Protocol*) proposés par Cerf et Kahn en 1974 [Cerf & Kahn,1974]. Si cette infrastructure permettait l'échange de fichiers elle n'était pas suffisante pour permettre à un chercheur de se servir d'informations produites par un autre chercheur à cause des formats de fichiers hétérogènes. Au lieu de faire une multitude de transformateurs de formats de fichiers, Tim Berners-Lee envisagea un outil beaucoup plus puissant qui pourrait intégrer des informations produites par différents utilisateurs dans un espace informationnel commun, dans lequel celles-ci pourraient être consultées, réutilisées et référencées.

Le concept d'hypertexte était également déjà connu à l'époque de la création du Web. Les travaux de Ted Nelson en 1965 [Nelson, 1965], dans le cadre de projet Xanadu[16], ont introduit le mot « hypertexte » pour signifier un système d'organisation de connaissances où les éléments textuels sont liés entre eux, ressemblant à un réseau. Ces travaux proposaient une concrétisation des idées de Vanevar Bush [Bush, 1945] qui envisageait un système conceptuel, appelé MEMEX, dont l'objectif était d'étendre le mémoire humaine. Vers la fin des années 70, les travaux sur hypertexte ont finalement porté leurs fruits sous forme des premières applications, telles que *Guide* développé en 1992 à l'Université de Kent [Gracca Campos Pimentel, 1992], mais ces applications ont majoritairement été conçues pour fonctionner sur un seul ordinateur ou dans un réseau local et n'avait pas la capacité de se connecter à l'Internet.

Tim Berners-Lee voyait une opportunité dans la combinaison de l'hypertexte et de l'Internet. Il croyait que l'hypertexte était un moyen idéal pour interconnecter des informations de type différent, en approvisionnant une interface unique par différentes formes d'informations, telles que rapports, notes, bases de données, etc. [Berners-Lee & Cailliau, 1990]. C'est en envisageant un système hypertexte où les utilisateurs pourraient créer des liens entre des contenus résidant dans différents ordinateurs, et naviguer dans cet espace d'informations connectées à l'aide de l'Internet, qu'il a créé le Web [Berners-Lee, 2000], vision conceptuelle, tout comme les premiers programmes qui ont permis de mettre en place cette vision.

---

[14] en. *Advanced Research Projects Agency Network*, destiné à l'échange de messages entre des universités participant dans des projets financés par l'agence de financement des projets de recherche militaire – DARPA (Defense Advanced Research Projects Agency).

[15] en. University of California, Los Angeles

[16] http://xanadu.com/

Il est très important de souligner la distinction entre le paradigme du Web en tant que moyen de connexion entre un humain et le monde extérieur comprenant des mémoires, des expériences et d'autres humains, et la réalisation technique de ce paradigme. Nous décrivons ces deux notions avec plus de détails dans la suite de cette section. Si le paradigme est resté inchangé, l'implémentation technique du Web a évolué au fil du temps, notamment dans le sens de l'amélioration des standards, des protocoles et des technologies qui font vivre le Web. Cette évolution a eu une influence sur la pratique de consommation du contenu Web, aussi bien que sur la pratique de mise à disposition du contenu via le Web. Cette influence a également rendu l'usage du Web en général plus facile. Plus le Web devenait facile à utiliser, plus le nombre de ses utilisateurs grandissait. Plus le nombre de ces utilisateurs devenait important, plus il était économiquement intéressant de le faire avancer et le rendre encore plus utile. C'est ce cercle vertueux qui a conduit le progrès rapide du Web et sa transformation en une plate-forme qui augmente les capacités cognitives de millions de personnes, dans une variété de situations quotidiennes.

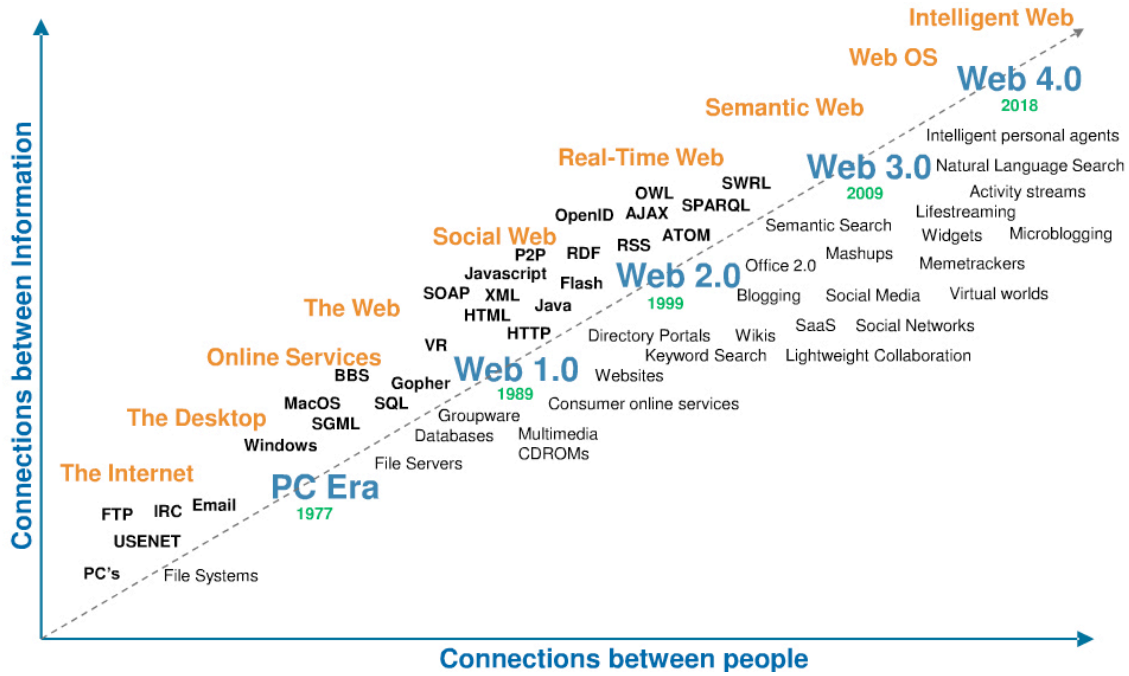


Figure 6 Evolution du Web. Repris de Nova Spivack [Spivack, 2009]

Certaines améliorations de l'infrastructure du Web ont un impact tellement important sur la perception du Web par ses utilisateurs qu'elles créent l'impression d'avoir produit un nouveau Web. Cet effet a pu être remarqué pour la première fois avec l'introduction des sites Web qui permettent la création facile du contenu Web et la collaboration avec d'autres utilisateurs. Ce fut la naissance de Web 2.0, connu aussi sous le nom Web Social[17]. Ensuite, l'initiative de rendre le contenu du Web plus structuré et plus manipulable par des machines fut appelée Web Sémantique ou Web 3.0. Par Web des objets nous entendons souvent un Web où les appareils et d'autres objets physiques sont connectés et référençables sur le Web. Même si la multiplicité d'appellations des

[17] Dans ce mémoire nous allons éviter d'appeler les différentes dimensions de l'évolution technique de Web par des numéros de version tels que 2.0 et 3.0, afin d'insister sur le fait qu'il ne s'agit pas d'une nouvelle version du concept de Web, mais plutôt d'une meilleure implémentation d'une de ces facettes, telles que l'aspect social (dans le cas de 2.0) et sémantique (dans le cas de 3.0).

différentes phases et différents aspects du développement technique du Web ne semble pas converger vers une nomenclature commune, une catégorisation proposée par Nova Spivack [Spivack, 2009] s'est imposée par sa clarté et reste parmi les plus citées sur le Web. Cette catégorisation, présentée sur Figure 6, explique les phases du développement du Web en fonction de leur apport en terme de la connectivité des personnes (sur l'axe X) et de la connectivité d'informations (sur l'axe Y) qu'elles apportent à la réalisation technique du Web. Cette perspective propose une vision de l'évolution du Web vue comme une oscillation permanente entre les directions de développement mettant un accent sur la connectivité des personnes, et celles mettant un accent sur la connectivité des informations. Si cette vision a permis d'expliquer le développement du Web à ce jour, il serait difficile d'expliquer son futur développement en fonction de ces deux dimensions. Nous verrons par la suite, dans ce chapitre, que l'avenir du Web annonce une connectivité des nouveaux éléments, au delà des personnes et des informations, et que la compréhension de futurs développements devrait reposer sur plusieurs nouvelles dimensions, telles que la connectivité d'objets physiques, des animaux, des applications automatiques, etc.

L'implémentation concrète du Web continue d'évoluer mais toujours dans le sens d'un même rêve de l'Homme connecté, de manière intense et synergétique, à son entourage informationnel, social, physique, etc.

2.1.2. Le paradigme

Le fait d'avoir insisté sur le caractère unique et inchangé du paradigme du Web à travers son histoire, nous oblige de donner une définition de ce paradigme. Nous présentons ici, une définition qui, nous espérons, passera l'épreuve du temps :

Le Web est un système d'extension de la mémoire de l'homme et de la société, où les éléments de mémoire sont organisés par des liens entre eux et forment un graphe. Grâce à ses caractéristiques sociales, ce système est en même temps un moyen de communication entre différentes entités qui peuvent s'y connecter.

Il est important de souligner la dimension sociale du Web. Si les premières initiatives qui essayaient d'étendre l'intelligence humaine, telles que NLS Augment, avaient pour objectif principal d'augmenter les capacités personnelles d'un utilisateur, le Web, suite à son caractère collaboratif et mondial, vise également à étendre la mémoire collective de la société. L'exemple de Wikipedia[18], qui cherche à préserver la connaissance encyclopédique collectée par l'humanité pendant des siècles de son existence, en est une bonne illustration.

Si ce moyen de communication est mondial et collectif, il sert par conséquent à connecter des utilisateurs humains entre eux mais également à d'autres éléments qui peuvent se voir connectés au Web, tels que les machines, les objets physiques ou même d'autres organismes. Le progrès récent dans la production des interfaces organiques, réalisé par des chercheurs de l'Université de Californie, Berkeley, permet par exemple de contrôler un insecte à l'aide d'une puce fixée sur son corps [Singer, 2009]. Il est donc nécessaire d'admettre qu'il n'y a pas désormais d'obstacles conceptuels pour différents types d'organismes ou d'objets à se connecter au Web, à condition qu'il y ait les interfaces appropriées.

[18] http://www.wikipedia.org

Le contenu du Web consiste en différents éléments de mémoire, sous différents formats qui sont reliés entre eux par des liens et forment un graphe (**graphe Web** par la suite). La nature des liens et des éléments de mémoire peuvent varier. Si par éléments de mémoire nous entendons principalement des informations, incréments de la connaissance, rien n'empêche l'inclusion d'émotions, expériences et autres éléments, à condition qu'ils puissent être captés et enregistrés. Les liens ont déjà subi une évolution, en passant de simples hyperliens non typés à des liens typés popularisés par la vision du Web Sémantique.

2.1.3. La réalisation technique

La réalisation technique initiale du Web fut modeste par rapport au potentiel du paradigme, et même par rapport à son état au moment de la rédaction de ce mémoire. La première raison en est certainement le niveau de développement des systèmes informatiques de l'époque, certains étant même dépourvus d'une interface graphique. Une autre raison est le peu de soutien apporté par son entourage à Tim Berners-Lee. Il a effectué lui-même une partie importante des réalisations de logiciels et ce, avec peu de ressources. Les trois éléments de la réalisation initiale, développés par Berners-Lee en 1990 sont :

- **HyperText Transfer Protocol** (HTTP), le protocole permettant d'effectuer le transfert d'éléments d'hypertexte à travers Internet. La version initiale, ensuite enrichie, est décrite dans [Berners-Lee, 1991].
- **HyperText Markup Language** (HTML), le langage de descriptions des éléments d'hypertexte et de leurs liens. C'est dans ce langage que tous les éléments du Web initial ont été exprimés.
- **WorldWideWeb**, renommé ensuite à Nexus pour éviter la confusion avec le concept entier du *World Wide Web*. C'était le premier navigateur du contenu du Web, qui permettait à la fois d'accéder au contenu et d'en créer. Les navigateurs qui ont été conçus plus tard ont majoritairement été destinés à la lecture et ont conduit à la naissance du Web non participatif qui dominait dans la première décennie d'existence du Web.

La coordination de ces trois éléments, reposait sur un élément clé permettant d'identifier des éléments du Web comme cibles d'hyperliens – l'URL (en. Uniform Resource Locator). Ce concept a depuis évolué en URI (en. Uniform Resource Identifier) [Berners-Lee, 1994] qui a une syntaxe plus riche. Si l'URL était suffisante dans le Web de départ où tous les éléments du graphe Web étaient des contenus que l'utilisateur consultait de manière directe, dans le graphe actuel la notion plus générale d'URI est nécessaire pour permettre de faire référence aux éléments qui ne sont pas destinés à la visualisation par l'utilisateur final.

Le perfectionnement permanent des standards du Web est principalement piloté par le W3C[19] (en. World Wide Web Consortium), un organisme de standardisation, dont les membres sont des universités, des centres de recherche et des entreprises impliquées dans l'industrie du Web. Ces acteurs, par un processus de consensus modélisent les nouveaux standards en fonction des besoins réels provenant des nouvelles pratiques d'utilisation du Web. Néanmoins, les groupes indépendants continuent à proposer des

[19] http://w3.org

standards en dehors de W3C, preuve de la nature démocratique du Web qui échappe à tout contrôle centralisé.

2.2. Web Social

« Le Web est une création sociale plutôt que technologique. Je l'ai créé en envisageant un effet social - pour faciliter la collaboration entre des personnes – ce n'est pas un jouet technique. » affirme Tim Berners-Lee dans son livre Weaving the Web [Berners-Lee, 2000]. Malgré cette intention, dans les premières années du Web il était difficile de se servir du Web pour communiquer. Le contenu du Web était créé principalement par une minorité d'utilisateurs formés au HTML, ce qui rendait le Web un canal de diffusion non participatif pour non initiés.

C 'est avec l'émergence des blogs, wikis, forums et d'autre sites destinés à la collaboration que le Web a commencé à prendre la forme d'un outil d'intégration dans l'aspect social du Web. Ce phénomène et son impact sur l'industrie du Web et sur l'apparition de nouveaux services innovants ont notamment été repérés par O'Reilly en 2005 [O'Reilly, 2005]. Deux éléments clés ont permis au Web de devenir plus interactif et plus participatif :

- **La facilité de création du contenu**. De nombreux sites ont proposé des pages Web qui permettaient à l'utilisateur de fournir ses propres textes, documents, images, vidéos etc. de les incorporer au Web très facilement, sans devoir rédiger de code HTML. Ce sont des composants de ces nouveaux sites, appelés « applications Web », qui se chargent de transformer les données de l'utilisateur aux formats du Web. Le contenu généré par les utilisateurs (en. *user-generated content*) était à l'origine d'une nouvelle excroissance du Web, qui persiste encore au moment de l'écriture de cette thèse. Parmi les nombreux types de sites sociaux nous pouvons distinguer notamment :
    - **Blogs –** sites qui permettent à chacun d'avoir son propre journal sur le Web. Très populaire à l'époque de la naissance du Web Social, cette pratique évolue de plus en plus vers la publication de billets très courts, reconnue sous le nom « microblogging » ; Nous appelons ces billets courts (d'une taille souvent limitée à 140 caractères) les tweets[20] (ou parfois « billets de microblog »).
    - **Wikis** - sites rendant possible la rédaction collaborative de documents. Un site de type wiki, est à l'origine de Wikipedia ;
    - **Sites de partage de photos et vidéos** tels que Flickr[21] et Youtube[22], permettent aux utilisateurs d'afficher leurs collections photographiques et leurs vidéos et de les exposer aux commentaires de la communauté en ligne ;

[20] Nous pouvons tracer l'origine de l'appelation « tweet» dans le nom de site Twitter (http://twitter.com) grace auquel la pratique d'echange de tweets est devenue fréquente. Si lors d'une conversation avec Jérome Euzenat nous avons pu trouver une traduction appropriée de ce mot par le mot francais *gazouille*, nous avons choisi d'utiliser le mot original dont l'usage est déjà très courant.

[21] http://flickr.com

[22] http://youtube.com

- **Forums -** sites qui facilitent la discussion au sein des communautés en ligne ;
- **Réseaux Sociaux** – sites destinés à la socialisation et la mise en relations des individus permettant les échanges de différents éléments (e.g. messages, photos, liens, commentaires)

- **Rechargement des pages en temps réel**. Les techniques pour l'affichage partiel des parties de pages Web, telles qu'Ajax (Asynchronous JavaScript and XML) ont commencé à émerger vers 2005 [Garrett, 2005]. Ces techniques permettent de mettre à jour certains éléments de pages Web, sans même l'intervention de l'utilisateur, créant ainsi un effet de dynamique et de continuité. Un bon exemple de cette pratique est le site Twitter[23] qui l'utilise pour donner l'impression d'un flux de messages qui passent en temps réel devant les yeux de l'utilisateur.

Si cette nouvelle facilité d'interaction avec le Web a rendu plus fréquents les échanges entre les utilisateurs, elle a également fait émerger un nouveau phénomène – la socialisation centrée sur l'objet – repéré d'abord par Engestrom [Engestrom, 2005] et puis étudié par Breslin and Decker [Breslin & Decker, 2007]. En interagissant avec des objets Web, les utilisateurs laissent des traces, par exemple sous forme de commentaires visibles par d'autres utilisateurs interagissant avec le même objet. Cette pratique peut conduire à des conversations et à l'établissement de nouveaux contacts. La socialisation autour du contenu Web est donc à la fois la cause et la conséquence de la création de contenus par des utilisateurs.

Il est important de souligner deux tendances qui émergent incontestablement sur le Web Social et qui transforment de manière significative la nature des éléments du Web :

- **La publication des éléments Web de petite taille**, appelés microposts. Les sites Web proposent de plus en plus des interactions nécessitant peu d'effort de la part de l'utilisateur. Par exemple sur le réseau social Facebook[24], l'utilisateur peut apprécier un objet Web en cliquant sur un bouton « j'aime » qui y est associé. Ces actions résultent de la création des nombreux micro-objets et intensifient l'interaction entre le Web et l'utilisateur. Un autre réseau social, Foursquare[25] permet aux utilisateurs de se déclarer présents dans un lieu localisé par GPS [NRC, 1995] en effectuant un « check-in » qui sert à signaler leur présence dans ce lieu en temps réel à leurs contacts. L'émergence de tweets courts, faciles à créer, publier et republier fait également preuve de la réalité de cette tendance.
- **L'émergence du contenu en tant que conséquence indirecte d'une action de l'utilisateur**. Certaines applications Web, de plus en plus nombreuses, identifient les activités de l'utilisateur et génèrent du contenu Web, sans l'intervention de l'utilisateur. C'est le cas de Spotify[26], le service de streaming musical, qui publie sur Facebook, de manière automatique l'information sur chaque chanson que l'utilisateur écoute.

Dans cette optique de croissance de la quantité et de la granularité d'informations publiées sur le Web, il est particulièrement important de pouvoir traiter et utiliser de

[23] Twitter est un site Web destiné à la publication des messages courts, dites tweets http://twitter.com

[24] http://facebook.com

[25] http://foursquare.com

[26] http://spotify.com

manière pertinente toutes ces informations. La conscience de ce défi ainsi que de ces deux tendances à imposé certains choix dans notre démarche scientifique. Elle peut, en partie, expliquer la nécessité d'une convergence du Web Social avec le Web Sémantique, qui lui, envisage de capter le sens du contenu du Web et de le rendre ainsi traitable par des machines de manière automatique.

2.3. Web Sémantique et Données Liées

2.3.1. Vision du Web Sémantique

« Le Web Sémantique est une extension du Web actuel dans laquelle l'information est enrichie de sens bien défini, permettant aux humains et aux ordinateurs de mieux travailler ensemble. » Tim Berners-Lee, Jim Hendler et Ora Lassila ont décrit le Web Sémantique dans leur article publié en 2001 dans le journal *Scientific American* [Berners-Lee et al., 2001]. Cet article, présente une vision d'un Web techniquement plus sophistiqué que le Web initial qui pourrait conduire à la réalisation de certains rêves qui ont obsédé les spécialistes du Web pendant des années. Les agents intelligents, envisagés dans les travaux liés à l'intelligence artificielle [Russel & Norvig, 1995], pourront, à l'aide de l'information structurée sur le Web Sémantique, automatiser certains processus qui exigeaient auparavant une intervention humaine, tels que le raisonnement, l'intégration d'informations de plusieurs sources hétérogènes, etc. Un tel agent pourrait par exemple, consulter la liste d'achats d'un utilisateur qui contient « gâteau d'anniversaire » et faire sonner son téléphone portable pour lui indiquer un achat possible lorsqu'il passe devant un supermarché, tout en déduisant que ce supermarché particulière héberge une pâtisserie qui pourrait bien vendre des gâteaux d'anniversaire. Il pourrait également tenir compte des avis des consommateurs sur la qualité du pain de cette boulangerie et en avertir l'utilisateur. Pour se rapprocher de la réalisation de cette vision des choses, il fallait faire un progrès technique important par rapport à l'état du Web, qui à l'époque en était à peine au stade du Web Social. Inspirés par certains travaux existants dans le domaine de la logique informatique et de la représentation des connaissances, les auteurs de cette conception ont indiqué comme particulièrement important le fait d'améliorer la capacité de capter le sens du contenu Web et le partager entre différentes applications. Cette conception a donné un souffle à des centaines de personnes qui essayent de proposer des perspectives différentes sur la manière d'aborder le problème et de contribuer à la réalisation d'une ou de plusieurs parties de la vision qui y est associée. Les travaux issus de ces efforts peuvent par exemple être trouvés dans les actes des conférences ISWC[27], ESWC[28], WWW[29] mais aussi dans des journaux tels que *Semantic Web Journal*[30], *Journal of Web Semantics*[31], etc.

Afin de suggérer une direction de développement technique pour le Web Sémantique, Tim Berners-Lee a proposé un schéma appelé « Semantic Web Layer Cake » (Figure 7). Ce schéma présente les technologies et les standards nécessaires pour compléter la

[27] International Semantic web Conference, http://swsa.semanticweb.org/content/international-semantic-web-conference-iswc

[28] Extended Semantic Web Conference, http://eswc-conferences.org/

[29] World Wide Web Conference, http://www.iw3c2.org/

[30] http://www.semantic-web-journal.net/

[31] http://www.journals.elsevier.com/journal-of-web-semantics/

vision du Web Sémantique, ainsi que leurs interdépendances (chaque couche de la pile dépend des couches qui sont au-dessous). Si les couches basses correspondent à des standards et des technologies existantes, les couches supérieures évoquent les capacités susceptibles d'être développées dans l'avenir, telles que les vérifications automatiques de faits (en. proof) et la gestion confiance par des applications Web d'une manière qui ressemble à la notion de confiance dans les relations humaines.

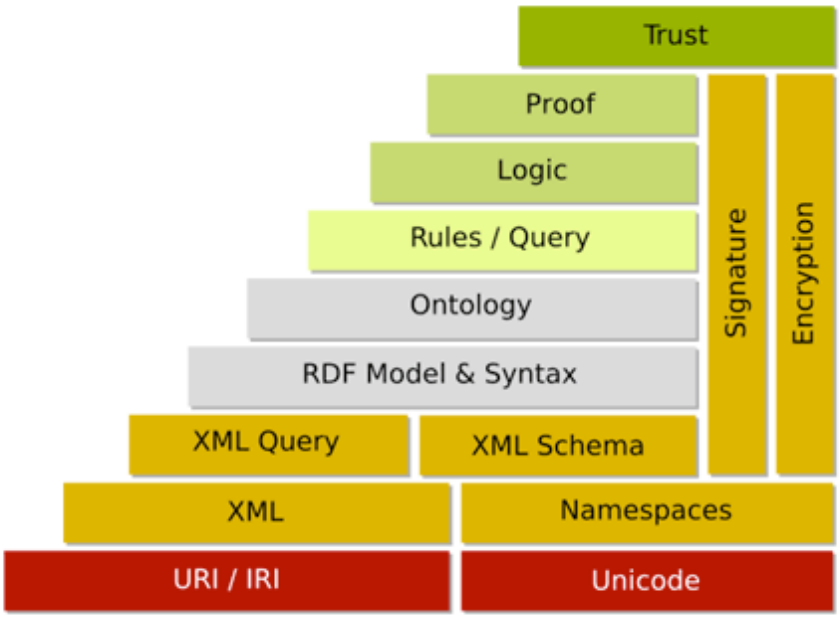


Figure 7 « Semantic Web Layer Cake », repris de Tim Berners-Lee[32]

D'une manière générale, dans la structure de ce nouveau Web à l'information structurée nous pouvons distinguer trois couches (de bas en haut, Figure 8):

La **couche de documents** correspond au Web initial dans lequel on trouve les éléments textuels, visuels, auditifs etc. qui sont destinés à la consommation directe d'un utilisateur. Pour simplifier on appelle tous ces éléments « documents ».

La **couche de données** qui contient des données structurées, qui sont au cœur de ce nouveau Web. Chaque nœud dans cette couche est identifié par son URI, et peut être lié aux autres nœuds par des liens typés. Ils peuvent également contenir des références vers des éléments de la couche de documents. Un tel lien est présenté dans la Figure 8 où un document est associé à une ressource de la couche de données qui le représente.

La **couche ontologique** est d'une importance vitale pour l'inférence des faits à l'aide de la couche de données. C'est la couche ontologique qui explique la signification des éléments de la couche de données. Cette signification est précisée dans la couche ontologique en associant les éléments de la couche de données à des structures de connaissance plus génériques qui permette de tirer des conclusions sur ceux-ci. Dans l'exemple présenté en Figure 8, le type d'objet identifié par son URI http://example.org/res/doc est désigné dans la couche ontologique comme http://ontology.org/Document. Dans la couche de données cet objet est lié par la propriété « hasTopic » à un autre objet http://example.org/WebSémantique. Grâce à la

[32] http://www.w3.org/2004/Talks/0412-RDF-functions/slide4-0.html

définition de cette propriété dans la couche ontologique, qui précise la nature des éléments avec lesquels il est légitime d'établir une relation de type « hasTopic » comme http://ontology.org/Topic, il est donc possible d'inférer le type d'objet http://example.org/WebSémantique, dans notre cas http://ontology.org/Topic. Cet exemple de base, démontre le principe d'inférence dans sa forme la plus simple. En fonction des différentes caractéristiques du formalisme utilisé pour exprimer les éléments des différentes couches, les possibilités de raisonnement peuvent varier.

Par la suite nous présentons ces différents formalismes de représentation des connaissances pour la couche de données et la couche ontologique, ainsi que des formalismes permettant à un agent intelligent d'effectuer des requêtes sur ces couches et d'obtenir des informations précises.

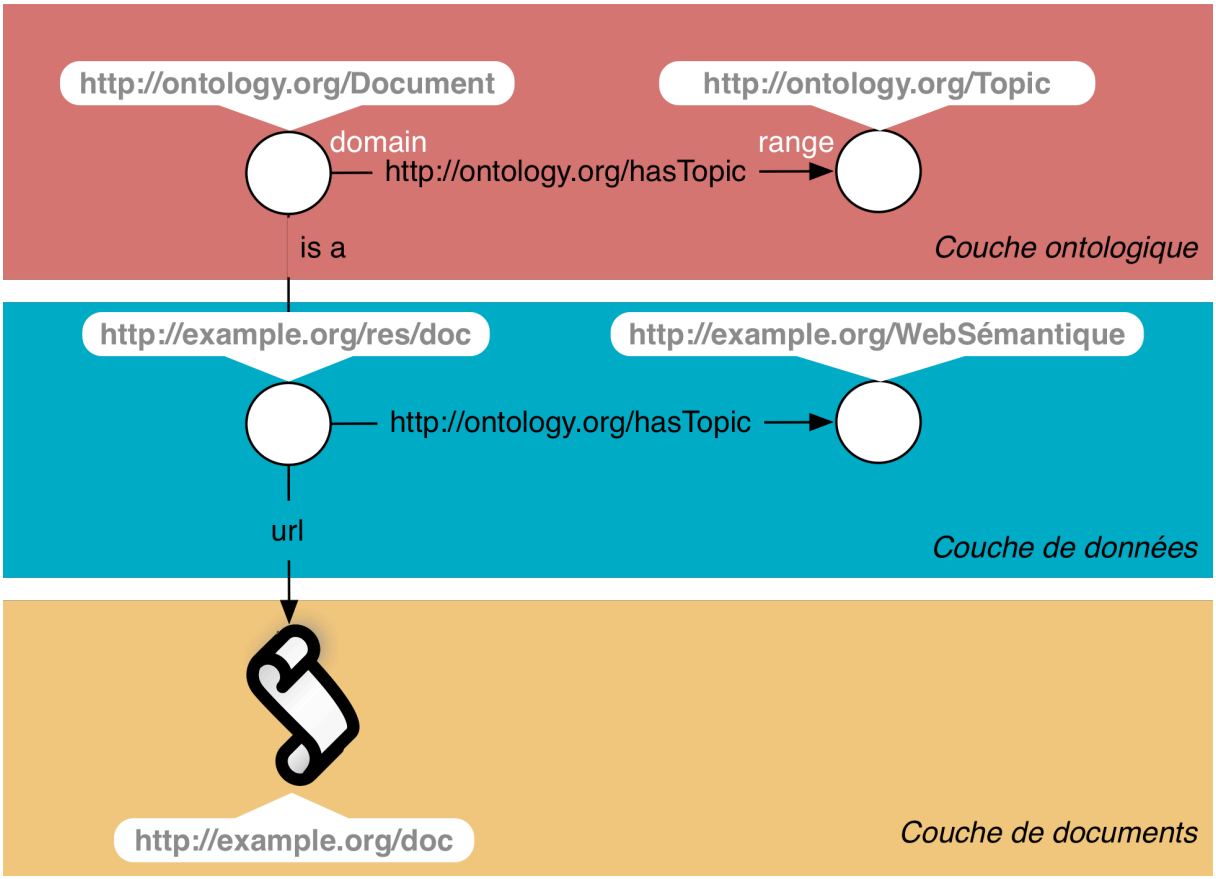


Figure 8 Trois couches du Web Sémantique

2.3.1.1. Représentation des données

Au fil des 10 dernières années le W3C a consacré beaucoup d'efforts à la création des standards de la représentation des données pour alimenter le Web Sémantique. Le standard RDF (en. Resource Description Framework) [W3C, 2004], dont la première version fut publiée en 1999 [W3C,1999], est au cœur de ces activités. RDF permet de décrire des graphes comme des ensembles de triplets <sujet, prédicat, objet>. À la place du sujet nous trouvons toujours une ressource Web, identifiée par son URI. Le prédicat est une propriété, éventuellement définie dans une ontologie et identifiée par son URI. Il est possible de distinguer deux types de propriétés : les propriétés d'objets et les propriétés de données. Les propriétés d'objets ont toujours une ressource en tant qu'objet dans le triplet, alors que les propriétés de données ont une valeur primitive

(e.g. numérique, textuelle). Le graphe présenté dans la couche de données sur Figure 8 pourrait donc être représenté comme triplet <http://example.org/res/doc, http://ontology.org/hasTopic, http://example.org/WebSémantique>.

Si l'organisation de l'information par triplets donne un cadre général de modélisation dans le modèle RDF, plusieurs formalismes permettent de traduire ce modèle conceptuel en un format lisible par des ordinateurs et traitable de manière automatique. Une syntaxe, appelée souvent RDF/XML [W3C, 2004a], se base sur le standard XML (en. *eXtensible Markup Language*) [W3C, 2008] et traduit le modèle de triplet en structure d'arbre imposé par ce standard. La structure d'arbre est parfois encombrante et l 'aspect visuel de la syntaxe orientée machines peut faire peur aux utilisateurs humains, comme les développeurs d'applications. Une autre syntaxe, Turtle (en. *Terse RDF Triple Language*) [W3C, 2011] se base sur la notation N3 (ou Notation 3) [W3C, 2011a] qui a été développée par Tim Berners-Lee et Dan Conolly afin de rendre la notation d'RDF plus facile à lire par des utilisateurs humains. Cette syntaxe rend ainsi la structure des triplets plus apparente.

@PREFIX ont : <http://ontology.org/>.

@PREFIX rdf :< http://www.w3.org/1999/02/22-rdf-syntax-ns#>.

<http://ontology.org/res/doc> ont :hasTopic <http://example.org/WebSémantique>.

<http://ontology.org/res/doc> rdf :type <http://ontology.org/Document>.

Figure 9 Exemple de graphe (décrivant la couche de données représentée sur Figure 8) en format Turtle.

Apres avoir introduit les deux syntaxes, la question qui se pose naturellement est comment les graphes, une fois exprimés en RDF, sont mis à la disposition des consommateurs, notamment les applications. Trois modalités sont particulièrement utilisées dans la pratique :

- Mettre à disposition des fichiers RDF, ou bien fournir du contenu RDF en réponse à une requête HTTP qui contient une demande de ressource décrite par le contenu RDF.
- Héberger les données RDF dans des entrepôts de données permettant aux consommateurs de lancer des requêtes et obtenir l'information qui leur est utile. Une interface d'interrogation peut également permettre d'interroger des données présentes dans des fichiers RDF. Nous allons détailler la mise à disposition et l'intégration de ces données dans des entrepôts dans la partie 2.3.1.2
- Incorporer des données RDF dans les pages HTML.

Cette dernière technique, permet essentiellement de réduire la redondance de données en publiant une même information qui est identifiée à la fois par des marqueurs HTML standards indiquant la manière dont elle doit être visualisée et par des marqueurs sémantiques indiquant le sens de l'information. Ces marqueurs sémantiques sont principalement définis par RDFa [W3C, 2012] – une syntaxe appuyée par le W3C. Malgré l'apparition de quelques initiatives qui proposaient des alternatives standards permettant d'incorporer les données structurées dans des pages HTML, telles que les microformats[33], RDFa a passé le test du temps. Ce standard est aujourd'hui utilisé par

[33] http://microformats.org/

des grands producteurs de données du Web, tels que Facebook[34], et il est préconisé par des moteurs de recherche Web majeurs, dans le cadre du projet schema.org qui assemble différents vocabulaires ou schémas de données pouvant être compris dans le processus d'indexation des pages HTML sémantiquement effectué par des moteurs de recherche. Les fichiers HTML contenant du contenu RDFa peuvent facilement être convertis en RDF grâce à des logiciels spécifiques, dits logiciels transformateurs, tels que RDFa distillateur[35].

GRDDL (en. Gleaning Resource Descriptions from Dialects of Languages) [W3C, 2007], le standard de transformation de formats semi-structurés tels que XML, XHTML[W3C, 2002] en RDF, , s'ajoute à la liste des initiatives qui envisagent la réutilisation du contenu du Web standard sous une forme sémantiquement riche. Des transformations GRDDL ont été créées pour de nombreux microformats, et permettent de transformer en RDF les données concernant des personnes, des événements, etc. Même avec de plus en plus de contenu publié en RDF, ces transformations restent encore utiles surtout dans la transition des parties du Web initial vers un « Web de données ». Ce terme est souvent utilisé pour distinguer le Web émergent, caractérisé par une couche de données de plus en plus riche et volumineuse.

2.3.1.2. Ontologies, sens de données et raisonnement

Les ontologies, servent à la fois de vocabulaire commun pour échanger des données structurées entre plusieurs acteurs de manière précise et de support au raisonnement automatique et à l'inférence de faits. Définie de manière précise, *une ontologie est une spécification formelle et explicite d'une conceptualisation partagée*, d'après une définition proposée par Gruber [Gruber, 1993] et puis enrichie par Studeret al. [Studer et al., 1998].

Le W3C est à l'origine de la création de plusieurs standards de modélisation de données permettant de créer des ontologies. RDFS (RDF schéma) [W3C, 2004b] est un de ces standards, très simple, à expressivité réduite. OWL (en. Ontology Web langage) [W3C, 2004b] est un autre langage d'une expressivité plus élevée et graduée. Plusieurs versions d'OWL ont différentes capacités expressives. Plus une version d'OWL est expressive, moins il est possible d'automatiser le raisonnement avec des données exprimées.

Aujourd'hui, OWL , dans sa version 2.0 [W3C, 2009], se base sur des expériences récoltées pendant quelques années d'utilisation de la première version. Le langage comprend 3 versions, appelées *profils,* OWL EL, OWL QL et OWL RL. L'usage des éléments de la langage restreint à ceux appartenant à OWL EL garantit les meilleures performances de raisonnement en échange d'une expressivité réduite. OWL QL est consacré aux usages dans les ontologies qui représentent une structure de données similaire à celle des bases de données relationnelles, et OWL RL garantit des bonnes performances de raisonnement avec OWL dans le cadre de systèmes de raisonnement à base de règles.

[34] https://developers.facebook.com/docs/opengraph/

[35] http://www.w3.org/2007/08/pyRdfa/

2.3.1.3. Interrogation du Web Sémantique

Certains fournisseurs de données structurées, de plus en plus nombreux, mettent à disposition des quantités de triplets RDF importantes. Pour des raisons pratiques et afin de faciliter l'accès à l'information par des parties externes, ils proposent souvent un point d'accès qui permet d'interroger leur données RDF et d'obtenir des informations précises concernant certaines ressources. Par exemple, dans un entrepôt contenant des informations démographiques, un tel point d'accès permettrait de sélectionner les villes qui ont plus d'un million d'habitants et qui ont été fondées avant 500 av. J.-C. et d'afficher leur nom et leurs coordonnées géographiques.

SPARQL [W3C, 2008a] est un langage développé par le W3C qui permet d'effectuer ce type de requête. Sa syntaxe s'approche de la syntaxe Turtle de RDF et s'inspire fortement de celle d'SQL[Codd, 1970] ce qui la rend facile à appréhender et surtout facile à utiliser en connaissant la structure du graphe qui est à interroger.

2.3.2. Données liées – Le Web Sémantique d'aujourd'hui

La fin des années 2000 a vu l'émergence de l'initiative *Linking Open Data*[36] qui représente un effort coordonné pour rendre conscient le public de la nécessité de publier des données dans des formats structurés, et cela de manière accessible au grand public surtout concernant des données qui sont d'intérêt public. Évidemment, la quantité de données disponibles est essentielle pour la croissance du Web Sémantique et la réalisation de la vision qui y est associée. Dans son célèbre discours « Raw Data Now » [Berners-Lee, 2009] Tim Berners-Lee va même au delà de cet objectif et insiste sur le caractère essentiel de la publication des données pour le bien de la société, quel que soit leur format. En effet avoir des données dans le domaine public est un point de départ important, et ce message a été bien reçu par des gouvernements qui ont commencé à publier des données produites par des agences publiques sur le Web, certains même en RDF. Ceci a donné lieu à la création de portails de données publiques aux Etats Unis[37], en Grande Bretagne[38], récemment en France[39] et dans une dizaine d'autres pays.

Résultant de la même initiative, il est question d'augmenter la quantité de données structurées présentes sur le Web et d'étendre le plus possible la couche de données en RDF. Le moyen d'atteindre cet objectif réside certainement dans la transformation des sources non structurées ou semi structurées existantes. Ce type d'effort est à l'origine de la création de DBpedia[40], une version RDF de Wikipedia, ou bien encore de la *sémantification* de Geonames[41], une grande base de données géographiques. DBpedia est très souvent utilisé dans nos travaux, et en cette raison nous la décrivons plus en détails dans la partie §2.3.3.

Si les données structurées ont une utilité incontestable, ce sont les liens qu'on peut établir entre ces données qui les rendent encore plus utiles. Pour chaque nouveau lien

[36] http://linkeddata.org/

[37] http://www.data.gov/

[38] http://www.data.gov.uk/

[39] http://www.data.gouv.fr/

[40] http://dbpedia.org

[41] http://www.geonames.org/

établi entre une ressource et une autre, la probabilité de retrouver ces deux ressources dans un parcours de graphe augmente, et par conséquent sa valeur informative devient plus élevée. L'idée d'organiser la couche de données en graphe de données liées qui s'apparente à celui du Web de documents a inspiré la formulation d'un guide des bonnes pratiques de mise à disposition de données, proposé par Tim Berners-Lee en 2006 [Berners-Lee, 2006]. Le guide consiste en 4 principes garantissant la formation d'un espace de données hyperliées :

1. Utiliser les URIs en tant que noms des objets Web ;
2. Utiliser les URIs HTTP pour que les utilisateurs puissent déréférencer[42] ces noms ;
3. Fournir des informations utiles, en utilisant des standards tels que RDF et SPARQL quand quelqu'un déréférence une URI ;
4. Inclure des liens vers d'autres URIs, pour que des utilisateurs puissent découvrir d'autres choses.

En suivant ces principes des différentes acteurs publient des ensembles de données structurés (nous les appellerons EDS par la suite) reliés éventuellement avec des EDS déjà publiés par des autres. Par exemple nous pouvons imaginer un EDS (EDS 1) contaminant les données publiques qui concernent les entreprises en France publié par une agence gouvernementale qui contiens des liens vers des ressources définies dans un autre EDS publique qui fournis des ressources décrivant des lieux géographiques en France. Au lieu de définir ces propres ressources pour les lieux géographiques le EDS 1 fais référence à une source crédible pour éviter la redondance.

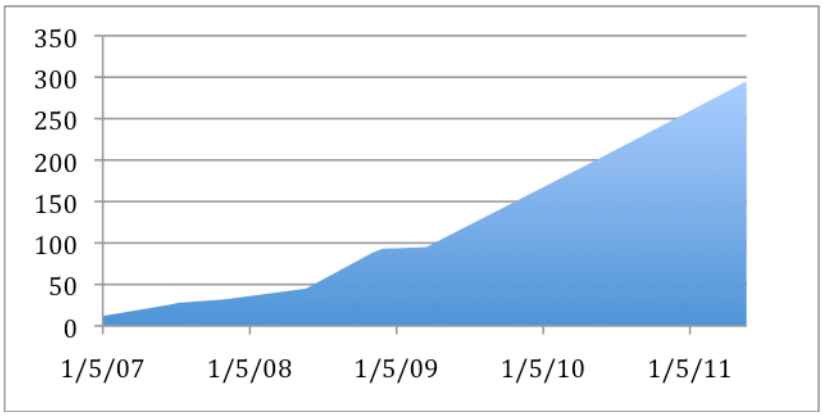


Figure 10 Nombre d'ensembles de données liées ouvertes, D'après les statistiques de Richard Cyganiak[43]

Nous appelons le graphe de données résultant de la pratique de la publication qui suit ces principes, graphe de données liées, ou simplement « données liées » (en. *Linked Data*). Les données liées ouvertes (en. *Linked Open Data, LOD par la suite*) représentent le sous-graphe de données liées qui est accessible de manière publique. La quantité de données liées ouvertes a connu une croissance exceptionnelle ces dernières années. La

[42] dans le sense du protocol HTTP

[43] http://richard.cyganiak.de/2007/10/lod/

Figure 10 démontre une croissance du nombre de jeux de données liées ouvertes depuis l'année 2007.[44] Un extrait de ce graphe est présenté sur la figure Figure 12.

2.3.3. La désambigüisation des sujets d'intérêt sur le Web de données

La composition actuelle du graphe de données liées ouvertes est présentée en Figure 12 qui présente une densité variable de liens entre des jeux dans les différentes parties de ce graphe. Des parties du graphe qui représentent des domaines avec plusieurs jeux de contenu similaire ont, naturellement plus de connexions. Par exemple, les entrepôts contenant des données relatives aux publications scientifiques, e.g. ACM, DBLP, IEEE, situées dans les parties droites du diagramme comprennent des données sur des mêmes entités (auteurs, articles, etc.) et par conséquent sont fortement référencées. Il est particulièrement important de souligner le rôle central de DBpedia dans ce graphe, et sa connectivité avec un grand nombre d'autres sources. Ce rôle émerge comme conséquence de l'utilisation des ressources de DBPedia pour désigner des sujets d'intérêt, dans d'autres entrepôts contenant des données sur les contenus Web, tels que les blogs, les articles scientifiques, etc. Le fait d'être composé de plus de 3.5 millions de ressources représentant des sujets de connaissance humaine, une quantité qui dépasse largement la taille des autres thesaurus et vocabulaires à contenu similaire (e.g., Wordnet[45], OpenCyc[46], etc.) rend DBpedia le mieux placé pour servir de « langage » commun.

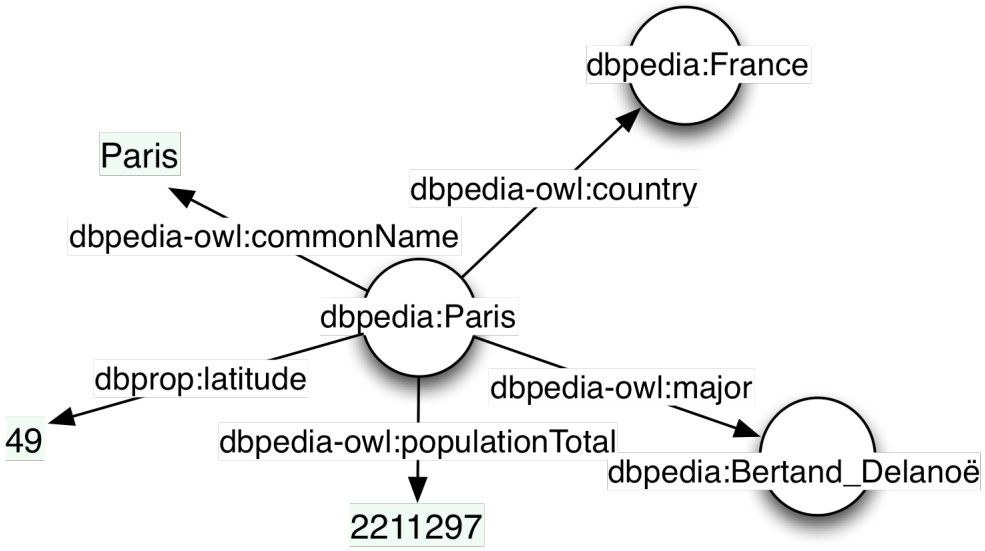


Figure 11 Extrait de l'information concernant Paris sur DBPedia

Utiliser des URIs DBPedia au lieu de définir ses propres ressources désignant des sujets d'intérêt est un choix fréquemment fait par des producteurs de données liées qui visent à augmenter la réutilisabilité de leurs données. Annoter des ressources Web par des concepts DBpedia est aussi très avantageux par rapport à l'ancienne pratique de

[44] Seulement des entrepots contenant plus de 1000 triplets sont pris en compte.

[45] http://wordnet.princeton.edu/

[46] http://www.cyc.com/

création des tags textuels puisqu' elle permet une identification précise des sujets d'intérêt. Par une URI. Il est possible de distinguer la notion de Paris, capitale de France de la notion de Paris, ville au Texas, ou bien de la notion de Paris Hilton ; ce qui n'aurait pas été tellement facile avec un tag « Paris ».

Un autre avantage de cette pratique d'utilisation des URIs DBPedia est sans doute le fait que ceux-ci donnent l'accès aux connaissances supplémentaires concernant les ressources qu'ils représentent. Dans la Figure 11 nous présentons un extrait de graphe qui concerne la ressource « Paris » définie au sein de DBPedia. Il est clair que la richesse informationnelle de DBPedia ouvre la porte aux nombreuses exploitations. L'utilité de DBPedia en tant que référentiel de concepts va donc au delà de simple désambigüisation de termes.

Si DBPedia est aujourd'hui le cœur du graphe de données liées, c'est grâce à sa nature de référentiel de ressources représentant des sujets d'intérêt qui est d'une importance essentielle pour la de publication de données liées à long terme. Sans intention de négliger la valeur de DBPedia et de l'excellent travail fait pour le rendre généralement utile [Bizer et al., 2009], il est possible d'imaginer une autre source de données à sa place, à condition que cette autre source soit riche en ressources représentant des sujets d'intérêt humain. Freebase, une source de taille plus petite que celle de DBpedia, est un exemple d'un tel entrepôt de données liées. Il fournit des informations, souvent plus fiables que celles de DBpedia, ce qui le rend utile malgré sa taille.

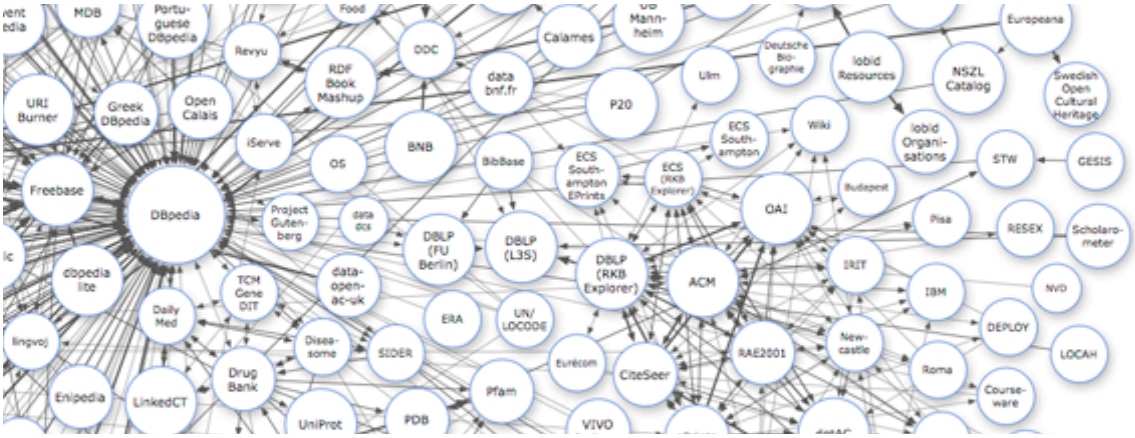


Figure 12 Un extrait du graphe de données liées ouvertes, en Septembre 2011. D'après Richard Cyganiak[47]

2.4. Web Social Sémantique

Le Web Social et le Web Sémantique ont été considérés, au départ de leur développement, comme deux visions opposées, largement en raison de leur origine dans différentes communautés, celle de l'industrie et celle de la recherche. La communauté scientifique a commencé à se rendre compte de la complémentarité des bénéfices que ces deux visions apportent à l'idée du Web autour de l'année 2005 quant les premiers travaux envisageant une convergence des deux visions ont vu le jour. Comme bons exemples de ces initiatives nous pouvons citer des travaux fondateurs de Breslin et Decker [Breslin & Decker, 2007], ainsi que la thèse de doctorat d'Alexandre Passant

[47] http://richard.cyganiak.de/2007/10/lod/

[Passant, 2009] dédiée principalement à ce sujet. Un nombre important des travaux liés à ce sujet est publié dans les ateliers *Social Data on the Web* de 2008 à ce jour[48].

Si cette convergence des deux initiatives (présentée dans Figure 13) implique un enrichissement des contenus sociaux, générés par des utilisateurs, par des formats sémantiques, elle implique aussi un changement de paradigme dans la création de plus en plus collaborative des données sémantiques sur le Web.

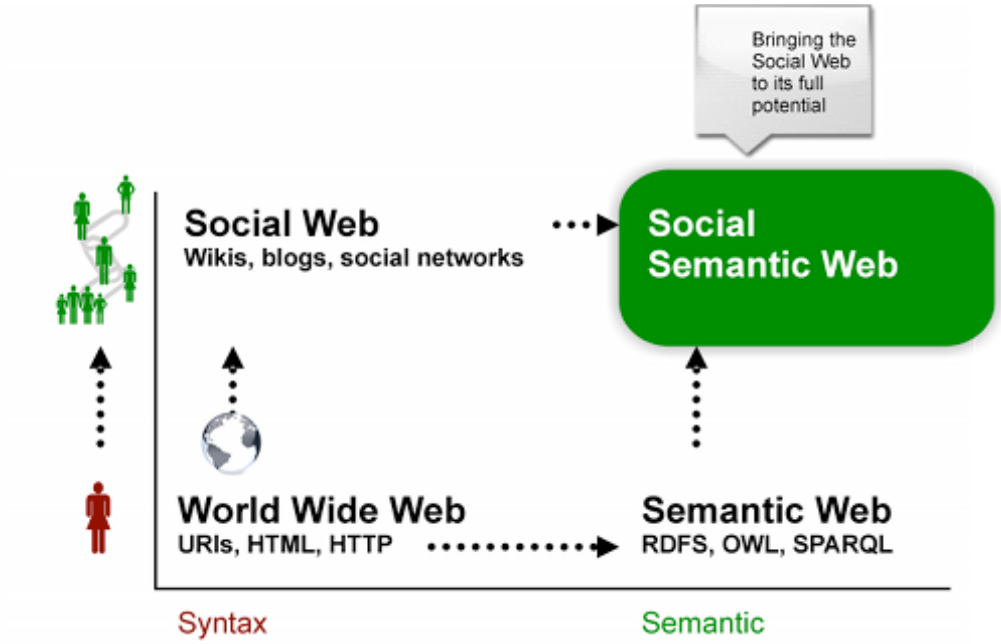


Figure 13 Convergence du Web Sémantique et Web Social. Repris de [Breslin et al., 2009]

Dans nos travaux, nous partons de l'hypothèse que le Web Social Sémantique est aujourd'hui une vision unique qui explique la nature du Web dans son état d'avancement actuel. Si nos contributions principales s'appuient particulièrement sur ces deux dimensions du Web, leur utilité à long terme est assurée par la prise de conscience de l'importance d'autres dimensions, présentées dans la partie 2.5, qui enrichiront le Web dans l'avenir.

2.4.1. Vers des données structurées à partir du contenu créé par les utilisateurs

Il est possible d'envisager des voies différentes dans la création des données issues des activités sociales qui seraient mises à disposition dans des formats sémantiquement riches. Une voie comprend certainement la transformation des contenus produits par les sites de Web Social en RDF. Dans ce but, un nombre de transformateurs a été mis en place. Par exemple nous pouvons citer l'extracteur du réseau de partage de photos - Flickr[49], celui de moteur de blogs Wordpress[50]. Dans le cadre du projet Semion [Nuzzolese et al., 2010] il est possible de trouver une approche de « sémantifisation » automatique du contenu des bases de données qui s'enrichit d'une phase de perfectionnement de résultats par des utilisateurs humains. Une autre branche

[48] http://sdow.semanticweb.org/

[49] http://www4.wiwiss.fu-berlin.de/flickrwrappr/

[50] http://wordpress.org/extend/plugins/wp-rdfa/

d'approche propose l'utilisation des méthodes d'analyse des réseaux sociaux (en. Social Network Analysis) pour caractériser des relations entre des utilisateurs des sites du Web 2.0 et ensuite représenter ces relations en RDF afin d'enrichir le graphe Web. Erétéo et al. présentent une telle approche [Erétéo et al., 2008].

Dans le cadre de la construction de nouveaux services et sites du Web Social, dont les créateurs sont partisans de cette vision, il est possible d'envisager la production de données sémantiques directement au fur et à mesure que l'utilisateur crée du contenu. Ceci est le cas des nombreuses applications issues du monde de la recherche, comme par exemple l'outil de microblogging sémantique SMOB [Passant et al., 2010] qui, lors de la création d'un billet de microblog, publie en RDF les éléments de contexte tels que la localisation de l'utilisateur et les éventuels sujets d'intérêt mentionnés dans le billet. Revyu [Heath & Motta, 2007] – site qui capte les avis des internautes sur différents produits ou objets et les publie sous un format sémantiquement riche est un autre exemple de ce type d'outil. Une approche également attractive est la création des données structurées dans les wikis, comme par exemple dans le cadre de projets Semantic Media Wiki[51] et SweetWiki [Buffa et al., 2008]. La pratique de création des données structurées comme conséquence des activités sociales est reprise de plus en plus dans les applications industrielles. Facebook[52], par exemple, met à disposition une partie des informations sur une entité (e.g. utilisateur, page) en fonction des droits d'accès du consommateur dans le format RDF[53] en utilisant son vocabulaire Open Graph[54].

Quelle que soit la voie par laquelle nous arrivons à produire des données sociales dans des formats sémantiques, il est souvent question, dans une mesure plus ou moins importante, d'extraire des concepts (sujets d'intérêt) définis dans le LOD à partir d'éléments textuels. Les utilisateurs sont habitués à produire du texte et des éléments multimédia. Dans certains systèmes ils fournissent également des annotations concernant des sujets d'intérêt liés à leur contenu, souvent sous forme de tags textuels, mais ils font rarement le choix des ressources du graphe de données liées qui représentent ces sujets. S'il est relativement facile d'automatiser la production des métadonnées relatives à la date et au temps de la création du contenu, de son créateur, etc. l'extraction et la désambigüisation automatique des sujets d'intérêt liés aux éléments textuels restent un vrai défi.

2.4.1.1. Extraction de ressources représentant des sujets d'intérêt, définies dans le LOD

Certains travaux de recherche et même de systèmes commerciaux ont contribué à relever ce défi de l'extraction des ressources définis dans le LOD à partir du texte. Par l'extraction nous comprenons **l'action d'identification de concepts et/ou ressources, préalablement définis dans le LOD, mentionnés dans un texte**. Ces systèmes permettent d'annoter des contenus textuels par des URIs de sujets déjà décrits dans des EDS ouvertes, et pour cela intègrent ces contenus de Web des documents dans la couche de données du Web Sémantique (appelé souvent Web de données).

[51] http://semantic-mediawiki.org/

[52] http://www.facebook.com

[53] https://developers.facebook.com/docs/reference/api/

[54] http://developers.facebook.com/docs/opengraph/#types , http://ogp.me/ns/ogp.me.ttl

Parmi les outils commerciaux, nous pouvons citer Zemanta[55], OpenCalais[56], Alchemy[57], comme étant les plus connus à l'heure actuelle. Zemanta est un outil de suggestion des ressources (les images, blogs, aussi bien que les ressources de DBPedia et Freebase) pour les auteurs de blogs qui s'en servent pour enrichir et annoter leur contenu et le lier avec d'autres éléments du Web. OpenCalais provient d'une initiative de Reuters pour annoter leurs contenus. Il est devenu disponible au grand public à travers un API Web. Alchemy est un ensemble d'outils de traitement de langue naturelle, qui, entre autres, peut servir pour la détection de mentions des ressources de données liées dans les textes.

La communauté de recherche a également produit un certain nombre d'outils d'extraction. Dans le projet Stanbol[58], il est possible de trouver un outil d'annotation de textes, principalement utilisé dans des sites Web permettant la création du contenu généré par des utilisateurs. Les résultats de ces projets sont majoritairement disponibles en source ouverte et de manière gratuite. Un autre extracteur, DBPedia Spotlight [Mendes et al., 2011], à source ouverte, créé par l'Université de Berlin, travaille exclusivement avec les ressources de DBPedia.

Une étude comparative de plusieurs services d'extraction de ressources de LOD à partir du texte fut présentée par Rizzo et Troncy [Rizzo & Troncy, 2011]. Dans cette étude, différents aspects de fonctionnement de ces systèmes sont étudiés. Zemanta est identifié comme le système le plus performant en ce qui concerne la détection des mentions des ressources de graphe de données liées dans les textes.

Conscients de la complexité de la tâche de reconnaissance des ressources représentant des sujets d'intérêt, très important pour la cohérence du Web, mais externes à nos problématiques de recherche et celles d'Hypios, nous allons nous appuyer sur des services existants, notamment Zemanta, dans nos travaux.

2.5. Une vision partielle du Web d'avenir

De plus en plus, nous pouvons remarquer qu'en plus des utilisateurs humains, des appareils peuvent se connecter sur le Web, ils utilisent son contenu et produisent du nouveau contenu. Dans l'idée même de l'Internet des objets (en. *Internet of things*) [Ashton, 2009] différents appareils domestiques (fours, réfrigérateurs, etc.), différents capteurs (capteurs de la pollution dans des grandes villes, cameras, capteurs de bruit, etc.), et d'autres objets pourront interagir avec le Web, même sans avoir besoin d'être contrôlés par un utilisateur humain. Certains projets dont l'ambition est très visionnaire proposent même des appareils qui permettent de surveiller l'état physiologique des animaux et reporter cette information sur le Web, particulièrement sur Twitter [Axon, 2010]. D'autres proposent des capteurs de l'état de santé d'une personne et leur mise à la disposition des médecins via le Web [Martin, 2010].

[55] http://www.zemanta.com/api/

[56] http://www.opencalais.com/

[57] http://www.alchemyapi.com/

[58] http://incubator.apache.org/stanbol/

S'il une telle vision du Web d'avenir peut paraître trop technophile et trop enthousiaste, elle reflète une certaine diversité, peut contestable, d'entités et d'interfaces qui vont se connecter au Web dans l'avenir.

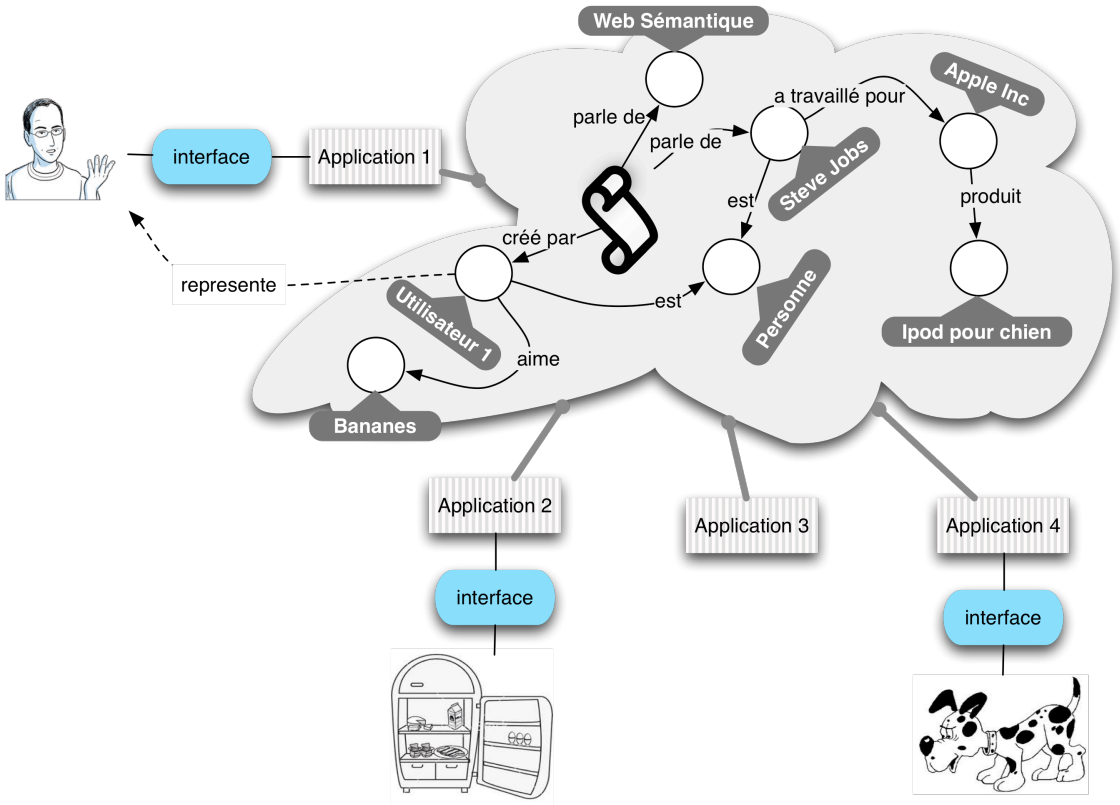


Figure 14 La vision d'un Web d'avenir

La diversité des entités, créatrices de l'information sur le Web, nous amène naturellement à considérer que le Web est un espace informationnel qui peut être utilisé et enrichi par une entité utilisatrice dont la nature peut varier. Cette entité accède au Web à travers une interface physique qui lui est adaptée (un ordinateur, ou téléphone portable pour les humains, un capteur spécial pour les animaux, etc.). A cette interface physique s'ajoute une application qui sert à effectuer différentes opérations nécessaires à l'information du Web. Pour les utilisateurs humains, cette application est le plus souvent un navigateur Web ; pour les animaux de notre exemple cité ci-dessus, c'est une application qui récupère des informations d'un capteur et les met à disposition sur le Web. Certaines applications, comme différents robots d'indexations Web (*en. Web crawlers*), peuvent agir de manière automatique, sans utilisateur et donc sans le besoin d'une interface physique. Cette compréhension du Web de l'avenir, représenté sur la Figure 14, rentre fondamentalement dans le cadre de la définition originellement donnée du Web dans la partie 2.2.1, mais elle précise la dimension des liens (externes) qui se créent entre le Web et le monde externe.

Dans cette nouvelle optique, il est possible de distinguer deux types de connectivité :

- Connectivité intra-Web, entre les éléments du graphe Web. Ce type de connectivité crée le Web au sens pur, tel qu'il est défini dans le §2.2.1.
- Connectivité extra-Web, entre les différentes entités utilisatrices et le Web. Cette connectivité crée un Web au sens large, où les utilisateurs humains et les différentes machines et objets sont connectés pour fonctionner ensemble, et

peuvent être référencés. Nous entendons même d'applications (hébergées sur des serveurs où sur les ordinateurs des utilisateurs) qui permettent d'effectuer différentes opérations sur le graphe Web comme faisant partie de cet espace conceptuel. Les interfaces physiques qui permettent l'activité de ces applications Web font également partie de ce Web au sens large.

Si l'évolution du Web à ce jour à pu être expliquée par le progrès dans la connectivité entre des utilisateurs d'une part et dans la connectivité entre des informations d'autre part, présentées dans la partie 2.1, pour comprendre le Web de demain, il serait nécessaire d'étudier également la connectivité entre les utilisateurs et les informations, en prenant en compte la diversité des types d'utilisateurs et de leurs interfaces vers l'information.

Nous avons définis les différents apports de cette thèse en faisant le constat de la diversité croissante, et parfois imprévisible, des acteurs et des entités utilisateurs du Web.

2.6. Conclusions

Dans ce chapitre, nous avons défini les notions clés nécessaires à la compréhension de la thèse et donné une vision globale du Web. De l'histoire de l'évolution du Web, les standards et pratiques utilisés, en passant par l'amélioration des outils techniques, nous avons vu une nouvelle définition de ce paradigme qui correspond à la situation actuelle. Par cette définition nous considérons le Web comme un outil d'extension de la mémoire collective humaine. Le Web est également un outil de communication et de collaboration. Une conscience de tous ces aspects du Web permet d'identifier des usages potentiels du Web dans le cadre de différents scénarios liés à l'innovation.

Nous avons également présenté les tendances de développement du Web à la fois sur le plan social et technique, ce qui nous a permis de mieux orienter nos choix d'axes de recherche. Ces tendances, ainsi que la vision du Web d'avenir décrite dans le chapitre §2.9, sont certainement utiles pour d'autres travaux qui ne traitent pas forcement le cas d'innovation sur le Web. A titre d'exemple, la tendance des internautes à publier ce que nous appelons des « microposts », c'est-à-dire des contenus de plus en plus petit, a inspiré un certain nombre de travaux utilisant par exemple le traitement de la langue naturelle, ou bien des travaux liés à la construction de systèmes de recommandation, etc. L'atelier « Making Sense of Microposts » que nous avons organisé en 2011[59] et 2012[60] autour de cette thématique a notamment permis de recueillir des exemples pour illustrer de telles approches.

Bibliographie

[Ashton, 2009] Kevin Ashton (2009), That 'Internet of Things' Thing. In: *RFID Journal*, 22 July 2009. Retrieved 8 April 2011, http://www.rfidjournal.com/article/view/4986

[59] http://research.hypios.com/msm2011/

[60] http://socsem.open.ac.uk/msm2012/

[Axon, 2010] Axon, S. (2010), Tweeting Dog Collar Posts Your Dog's Movements to Twitter, Mashable Blog, http://mashable.com/2010/02/13/tweeting-dog-collar/

[Berners-Lee & Cailliau, 1990] Berners-Lee, T. & Cailliau, R. (1990) WorldWideWeb: Proposal for a HyperText Project , http://www.w3.org/Proposal.html

[Berners-Lee et al., 2001] Berners Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*.

[Berners-Lee, 1991] Berners-Lee, T (1991), Hypertext Transfer Protocol, http://www.w3.org/Protocols/HTTP/AsImplemented.html

[Berners-Lee, 1994] Berners-Lee, T. (1994), Universal Resource Identifiers in WWW. Request For Comments 1630. Internet Engineering Task Force. http://tools.ietf.org/html/rfc1630

[Berners-Lee, 2000] Berners-Lee, T. (2000). *Weaving the Web*. San Francisco, CA: HarperSanFrancisco.

[Berners-Lee, 2006] Tim Berners-Lee (2006), LinkedData, http://www.w3.org/DesignIssues/LinkedData.html

[Berners-Lee, 2009] Tim Berners-Lee on the next Web, a TED talk, http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html

[Bizer, 2009] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). DBpedia - A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web,* 7(3), 154-165. doi:10.1016/j.websem.2009.07.002

[Bresil & Decker, 2007] Breslin, J., & Decker, S. (2007). The Future of Social - The Need for Semantics. *Ieee Internet Computing*, 5, 86-90.

[Breslin et al., 2009] Breslin, J., Passant, A., & Decker, S. (2009). *The Social Semantic Web*. Heidelberg: Springer-Verlag.

[Buffa et al., 2008] Buffa, M., Gandon, F., Ereteo, G., Sander, P., & Faron, C. (2008). SweetWiki: A semantic wiki. *Web Semantics: Science, Services and Agents on the World Wide Web*, *6*(1), 84-97. doi:10.1016/j.websem.2007.11.003

[Bush, 1945] Bush, V. (1945). As We May Think. *The Atlantic Monthly*, 18(3). Retrieved from http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/3881/

[Cerf & Kahn,1974] Cerf, V.G. Kahn, R. E. "A Protocol for Packet Network Intercommunication", IEEE Transactions on Communications, Vol. 22, No. 5, May 1974 pp. 637-648

[Codd, 1970] Codd, Edgar F (1970). "A Relational Model of Data for Large Shared Data Banks". *Communications of the ACM* (Association for Computing Machinery) 13 (6): 377–87.DOI:10.1145/362384.362685. Retrieved 2007-06-09.

[Engestrom, 2005] Engestrom, J (2005) Why some Social Networks Work and Others dont or the Case for Object-centered Sociality. http://www.zengestrom.com/blog/2005/04/why-some-social-network-services-work-and-others-dont-or-the-case-for-object-centered-sociality.html

[Erétéo et al., 2008] Erétéo, G., Buffa, M., Gandon, F., Grohan, P., & Sander, P. (2008). A State of the Art on Social Network Analysis and its Applications on a Semantic Web, *Proceedings of SDoW2008*, workshop at ISWC, Karslruhe, Germany, 1-6.

[Garrett, 2005] Garrett, J. J., (2005) Ajax: A New Approach to Web Applications". Adaptive Path. http://www.adaptivepath.com/ideas/ajax-new-approach-web-applications

[Gracca Campos Pimentel, 1992] Gracca Campos Pimentel, M. (1992) Showing the destination of hypertext links: a new approach for guide. Technical Report 8-92, University of Kent

[Gromov, 1995] Gromov, G. (1995) *Roads and Crossroads of Internet History*

[Gruber, 1993] Gruber, T.R. (1993) A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*. 5: p. 199–220.

[Heath & Motta, 2007] Heath, T. and Motta, E.: Revyu.com: a Reviewing and Rating Site for the Web of Data . *Proceedings of the Semantic Web Challenge 2007*, ISWC Conference, http://ceur-ws.org/Vol-295/paper12.pdf

[Martin, 2010] Martin, R. (2010), Wearable wireless health sensor for remote bio-monitoring. Gizmag blog, http://www.gizmag.com/health-monitoring-system-japan/14047/

[Mendes et al., 2011] Mendes P.N., Jakob M., García-Silva A., Bizer C. (2011) DBpedia Spotlight: Shedding Light on the Web of Documents. In the *Proceedings of the 7th International Conference on Semantic Systems* (I-Semantics 2011). Graz, Austria, September 2011. [DOI 10.1145/2063518.2063519]

[Nelson, 1965] Nelson, T. (1965). The Hypertext. *Proc World Documentation Federation Conference*.

[NRC, 1995] National Research Council (U.S.). Committee on the Future of the Global Positioning System; National Academy of Public Administration (1995). *The global positioning system: a shared national asset: recommendations for technical improvements and enhancements*. National Academies Press. p. 16. ISBN 0-309-05283-1., Chapter 1, p. 16

[Nuzzolese, 2010] Nuzzolese, A. G., Gangemi, A., Presutti, V., & Ciancarini, P. (2010). Fine-tuning triplification with Semion. *EKAW workshop on Knowledge Injection into and Extraction from Linked Data* (KIELD2010) (pp. 2-10).

[O'Reilly, 2005] Tim O'Reilly (2005). O'Reilly Network : What Is Web 2.0 : Design Patterns and Business Models for the Next Generation of Software. http://www.oreillynet. com/lpt/a/6228

[Passant et al., 2010] Passant, A., Bojars, U., Breslin, J. G., Hastrup, T., Stankovic, M., & Laublet, P. (2010). An Overview of SMOB 2: Open, Semantic and Distributed Microblogging. *4th International Conference on Weblogs and Social Media*, ICWSM (pp. 303-306). Retrieved from http://swl.slis.indiana.edu/files/smob-icwsm-poster.pdf

[Passant, 2009] Passant, A. (2009) *Technologies du Web Sémantique pour l'Entreprise 2.0,* Thèse de doctorat. Université Paris-Sorbonne IV

[Rizzo & Troncy, 2011] Rizzo, G., & Troncy, R. (2011). NERD : Evaluating Named Entity Recognition Tools in the Web of Data. *Proceedings of the 11th Interational Semantic Web Conference ISWC2011*. Bonn, Germany.

[Russel & Norvig, 1995] Russel, S., & Norvig, P. (1995). Artificial Intelligence: A Modern Approach . Prentice Hall; 1st edition. Retrieved from http://www.amazon.com/Artificial-Intelligence-Approach-Stuart-Russell/dp/0131038052

[Singer, 2009] Singer, E. (2009). Bilogical Machines. *Technology Review*, MIT, http://www.technologyreview.com/biomedicine/22111/page1/

[Spivack, 2009] Spivack, N. (2009). Web Evolution. Presentation sur Slideshare.net. http://www.slideshare.net/novaspivack/web-evolution-nova-spivack-twine

[Studer at al., 1998] Studer, R., Benjamins, R. and Fensel, D. (1998) : Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, 25(1-2):161-198

[W3C, 1999] W3C (1999), Resource Description Framework: Model and Syntax Specification, http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/

[W3C, 2001a] W3C (2011), Notation3 (N3): A readable RDF syntax, http://www.w3.org/TeamSubmission/n3/

[W3C, 2002] W3C (2002), XHTML 1.0 The Extensible HyperText Markup Language (Second Edition), http://www.w3.org/TR/xhtml1/

[W3C, 2004] W3C (2004), Resource Description Framework: Concepts and Abstract Syntax, http://www.w3.org/TR/rdf-concepts/

[W3C, 2004a] W3C (2004), RDF/XML Syntax Specification, http://www.w3.org/TR/rdf-syntax-grammar/

[W3C, 2004b] W3C (2004), RDF Vocabulary Description Language 1.0: RDF Schema, http://www.w3.org/TR/rdf-schema/

[W3C, 2004c] W3C (2004), OWL Web Ontology Language, http://www.w3.org/TR/owl-features/

[W3C, 2007] W3C (2007), Gleaning Resource Descriptions from Dialects of Languages (GRDDL), http://www.w3.org/TR/grddl/

[W3C, 2008] W3C (2008), Extensible Markup Language XML 1.0 (Fifth version), http://www.w3.org/TR/2008/REC-xml-20081126/

[W3C, 2008a] SPARQL Quelry Language for RDF, http://www.w3.org/TR/rdf-sparql-query/

[W3C, 2009] W3C (2009), OWL2 Web Ontology Language, http://www.w3.org/TR/owl2-overview/

[W3C, 2011] W3C (2011), Terse RDF Triple Language, http://www.w3.org/TR/2011/WD-turtle-20110809/

[W3C, 2012] W3C (2012), RDFa Core 1.1. Syntax and processing rules for embedding RDF through attributes, http://www.w3.org/TR/rdfa-syntax/

3. Cadre conceptuel : Rapprochement des éléments Web liés à l'innovation

Ce chapitre expose le cadre conceptuel qui oriente nos recherches et guide le choix de nos hypothèses de recherche. Ce cadre repose sur l'idée d'une interdépendance entre le graphe de concepts et le graphe social qui relie les utilisateurs Web entre eux. Cette interdépendance dépend notamment des activités quotidiennes sur le Web qui sollicitent en permanence les deux graphes tout en contribuant à leur évolution. Un modèle d'interaction entre les utilisateurs et le Web est présenté dans le §3.1. Ce modèle facilite la compréhension du rôle des activités dans l'évolution des graphes. Nous proposons ensuite (§3.2) une idée générale de rapprochement des éléments Web. L'objectif de ce rapprochement est de faciliter l'accès, pour une situation donnée, à certains éléments pertinents, réduisant ainsi la qualité de l'interaction (e.g. navigation) avec le graphe Web. Nous nous focalisons notamment sur les rapprochements des éléments liés à l'innovation, et dont la mise en place pourrait faciliter l'émergence de l'innovation sur le Web. Tous ces éléments sont décrits dans la partie §3.3. Dans la partie §3.4 nous présentons tous les rapprochements possibles de ces éléments en nous focalisant sur ceux qui seront décrits en détail dans les chapitres qui suivent.

3.1. Accès à l'information sur le Web

Comme nous avons pu le voir dans le Chapitre 2, le Web est composé de ressources informationnelles potentiellement utiles pour un utilisateur dans une situation particulière. Nous appelons ces composants informationnels « les éléments Web ». Ces éléments peuvent être des contenus textuels ou visuels créés par des utilisateurs, ou bien des représentations d'objets tels que des utilisateurs, des produits, des lieux géographiques, etc. Ils peuvent aussi bien avoir une représentation dans la couche de documents que dans la couche de données. Pour mieux comprendre la capacité informationnelle du Web, il est intéressant d'étudier l'accessibilité, la création et l'interaction de ces éléments. Pour ce faire, nous présentons dans cette partie un scénario générique d'usage du Web qui nous permettra de définir les notions relatives à l'accessibilité informationnelle.

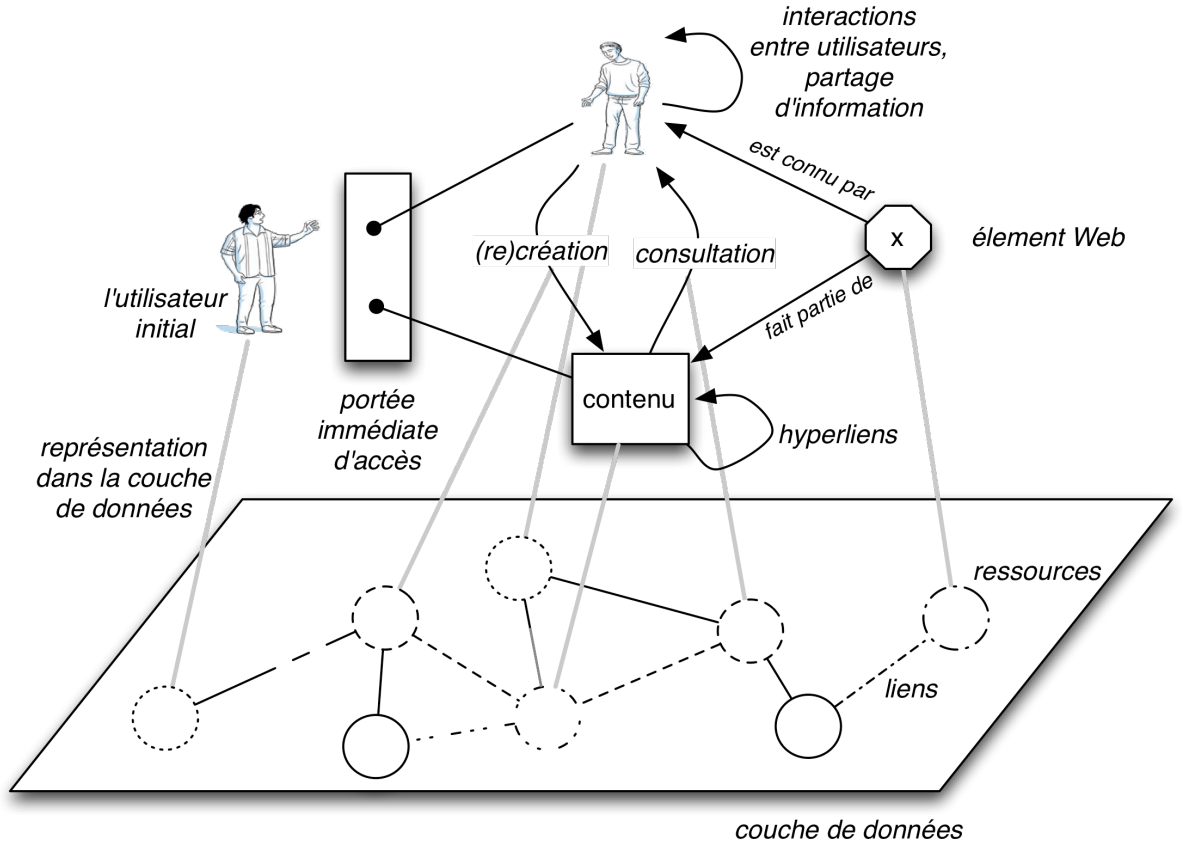


Figure 15 Propagation de l'information à travers le Web et sa conséquence dans le graphe Web

Imaginons un élément Web utile à un utilisateur (utilisateur initial par la suite) dans une situation donnée. Cet élément peut concerner une personne, un document ou un autre élément informationnel. La vie de cet élément et les informations le concernant peuvent prendre des voies différentes sur le Web. La Figure 15 illustre un tel élément (marqué par le signe X) et ces différentes voies. L'élément en question peut par exemple faire l'objet d'un document/contenu Web, interlié avec d'autres contenus par des hyperliens. Il peut également être introduit sur le Web par un utilisateur dans le cadre de ses activités créatrices de contenu. Au cours des interactions entre utilisateurs, le contenu peut se voir partagé avec des autres utilisateurs, republié, commenté, etc. Toutes ces

voies créent des chemins entre l'élément X et l'utilisateur initial qui le consulterait à un moment donné dans le cadre de notre scénario.

Considérons maintenant le scénario où notre utilisateur initial *U*[61] interagit avec le Web pour trouver un élément *X* qui peut lui être utile dans sa situation particulière *S*. Il accède au Web par l'intermédiaire des navigateurs Web et d'autres applications. *L'ensemble des éléments Web auxquels il peut accéder à un moment donné représente ce que nous appelons « la portée immédiate d'accès »* (PIA par la suite). Par exemple, si le navigateur affiche une page typique de réseau social. Tous les messages, amis, et autres ressources visibles sur cette page, seront considérés comme éléments de son PIA. Prenons un autre exemple, celui d'une carte interactive que l'utilisateur consulte par le biais de son appareil mobile. Toutes les ressources représentes sur cette carte (restaurants, rues, institutions, musés, etc.) font alors partie de son PIA.

En interagissant avec des applications Web (e.g., en cliquant sur un lien dans un navigateur) l'utilisateur change l'ensemble des éléments qui sont à sa portée immédiate d'accès. En partant d'un ensemble d'éléments de départ, l'utilisateur explore différentes voies en enchaînant des interactions de nature diverse jusqu'au moment où l'élément utile X se retrouve dans son PIA. L'utilité de X peut soit être connue à l'avance si l'utilisateur cherche un élément déjà connu, soit évaluée au fur et à mesure s'il fait des découvertes inattendues lors de ses recherches. Le parcours de l'utilisateur dans ce scénario est représenté par la Figure 16.

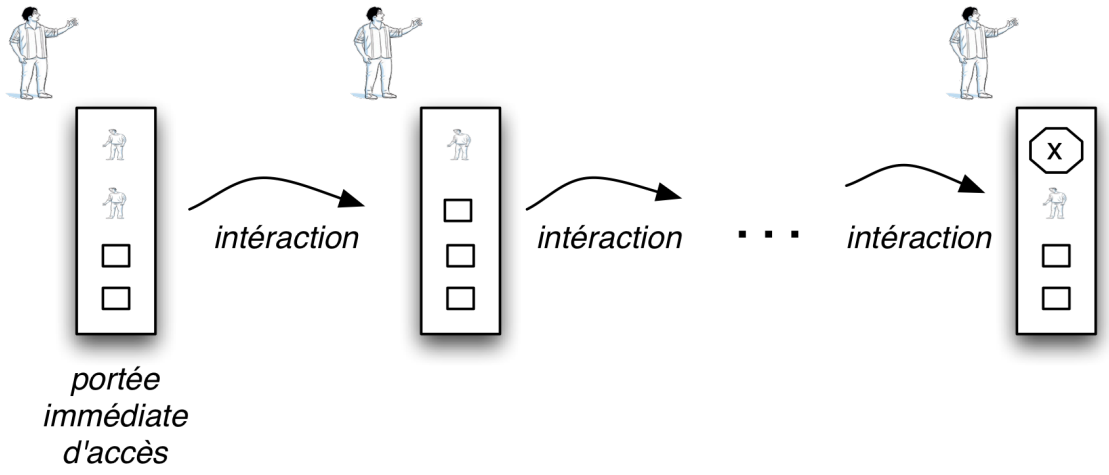


Figure 16 Changement de la portée immédiate d'accès à l'information du graphe Web lors de l'interaction itérative

Le nombre d'interactions nécessaires pour arriver à un élément utile peut donc varier. Nous pouvons dire qu'il existe une probabilité $P(i<n, X \in PIA_i \mid U,S)$ qu'un élément X, utile à l'utilisateur U par rapport à une situation S, se trouve dans le PIA dans l'itération *i* de l'interaction, où *i* est inférieure à un nombre prédéfini *n*.

Pour toute interaction avec le graphe Web, quelle que soit son mode et sa nature, il est intéressant d'augmenter cette probabilité pour des n aussi petits que possible. En analysant la nature des distances entre les éléments Web formés par des chemins construits lors des interactions sociales des utilisateurs et lors de leurs interactions avec

[61] Il est important de noter que, dans le rôle de l'utilisateur initial, il est possible de trouver aussi bien des utilisateurs humains que des programmes - agents automatiques.

les contenus, il apparaît clairement qu'un nombre de facteurs sociaux influencent la probabilité P. Ces facteurs sont entre autres :

- La densité des liens sociaux et la position de l'utilisateur dans le graphe social ;
- La pratique de la consommation, du partage des informations et de l'interaction avec des contenus au sein des réseaux sociaux.

La conception des interactions qui favorise un partage pertinent et efficace serait certainement utile pour améliorer les chances de l'utilisateur de trouver rapidement les éléments recherchés. Une innovation d'un tel niveau pourrait faire l'objet d'études ergonomiques et d'un travail sur l'interaction homme-machine, mais est hors du cadre de cette thèse. Dans ce mémoire nous nous intéressons principalement aux facteurs liés à la structure et l'accès au graphe Web. La manipulation de ces facteurs permettrait d'augmenter la probabilité que l'utilisateur retrouve dans son PIA des éléments utiles en fonction de sa situation dans les interactions à faible degré de complexité (moins d'itérations). Ces facteurs sont principalement:

- La connectivité des éléments utiles pour la situation S dans le graphe Web, i.e. le nombre et la longueur des chemins séparant l'utilisateur U des éléments utiles ;
- La capacité des applications, utilisées pour interagir avec le graphe Web, de le traiter de manière pertinente par rapport aux besoins du scénario i.e. de faire des hypothèses pertinentes sur des directions susceptibles d'aboutir en guidant l'interaction dans le sens attendu.

Une voie d'amélioration de la connectivité des éléments utiles est sans doute une représentation riche et adéquate des ressources dans la couche de données. Nous abordons plus en détails cette question de la représentation des connaissances pertinentes pour le scénario de l'innovation sur le Web dans le Chapitre 8.

En se qui concerne le deuxième facteur, il appartient aux applications Web de « rapprocher » des éléments Web i.e. de faire en sorte que les éléments Web utiles dans une situation puissent se trouver plus facilement et plus rapidement dans le PIA de l'utilisateur. La partie suivante est consacrée à la définition plus formelle d'un tel processus de rapprochement.

3.2. Notion du rapprochement

3.2.1. Définition du rapprochement des éléments du graphe Web

Introduisons d'abord la notion de rapprochement qui nous permettra de se référer à la capacité de systèmes de bien guider l'interaction avec le graphe d'une manière pertinente en s'appuyant sur sa structure. *Un* ***rapprochement des éléments Web*** *est une opération, ayant des effets temporaires ou permanentes, effectuée au niveau du graphe Web, et qui permet de considérer deux (ou plusieurs) éléments utiles l'un pour l'autre dans une situation particulière comme étant plus proches dans le sens de l'accessibilité mutuelle, ainsi augmentée.* De manière générale, un rapprochement permet d'associer un élément ou un ensemble d'éléments de départ, à un autre ensemble d'éléments potentiellement utiles dans un contexte donné et dans une situation particulière dans laquelle se trouve l'utilisateur. La formule 1 représente le

rapprochement (*R*) en tant qu'une opération qui établit une relation entre un ensemble d'éléments du graphe Web (éléments de départ, *E*) et l'ensemble d'éléments (*E'*) considérés proches dans un contexte donné.

$$R : E \xrightarrow{contexte} E' \quad (1)$$

Le résultat pratique d'un rapprochement est la capacité des applications Web de faire figurer dans le PIA d'un utilisateur un élément Web utile dans sa situation particulière plus rapidement i.e. avec une probabilité *P* augmentée par rapport à un parcours du Web qui consiste à suivre des liens explicites.

Un rapprochement sous-entend souvent une ***hypothèse de pertinence****, règle qui définit les conditions nécessaires pour considérer un élément comme pertinent pour un autre, dans le cadre d'une situation particulière.* Un exemple d'hypothèse de pertinence peut être donné par la règle suivante : un élément Web, qui contient des mots-clés utiles pour la situation de l'utilisateur et qui est (re)partagé par plus d'utilisateurs peut être considéré comme plus pertinent qu'un autre élément partagé par moins d'utilisateurs.

Les diverses manières d'effectuer ces rapprochements et des diverses hypothèses de pertinence vont se montrer utiles dans différents scénarios d'usage concrets. Dans la partie §3.2.2, nous présentons quelques exemples de rapprochements mis en place par des applications Web existantes qui profitent principalement de l'expressivité de la couche de documents. Plus tard, dans la partie §3.4.1, nous présentons des avantages que la couche de données peut apporter au processus de rapprochement et nous introduisons de manière préliminaire et générale nos rapprochements qui profitent de l'expressivité de cette couche.

Finalement, pour mieux délimiter le cadre de cette thèse il est important de noter que la capacité réelle de l'utilisateur de trouver un élément utile ne dépend pas seulement de la probabilité de trouver cet élément dans son PIA, mais aussi de l'attention que cet utilisateur porte à la PIA. La conception des interfaces utilisateurs favorisant la navigation facile et pertinente reste donc un défi important et complémentaire à celui que nous avons abordé dans cette thèse.

3.2.2. Exemples de la pratique de rapprochement dans le contexte du Web de documents

Les rapprochements faits par les moteurs de recherche Web sont sans doute un exemple auquel les internautes sont le plus fréquemment confrontés. Ces applications produisent dans une page Web un nombre de liens classés par pertinence en fonction de mots-clés fournis par les utilisateurs. Cela permet d'éviter un parcours du Web susceptible d'être très long, et pendant lequel l'utilisateur devra suivre des liens hypertextes successifs jusqu'à ce qu'il trouve les contenus utiles à sa situation. Les moteurs de recherche créent des index de contenus Web et se servent de différentes hypothèses de pertinence pour délivrer des résultats susceptibles de satisfaire le besoin de leurs utilisateurs. Google, par exemple, utilise leur célèbre formule PageRank™ [Brin & Page, 1998] pour classer des pages Web selon leur importance. Plus les pages Web ont de liens, plus elles sont importantes. Dans le calcul de classement, cette formule récursive tient également compte du classement des pages d'origine des liens.

Un autre exemple très connu peut être souligné dans le cas des sites d'e-commerce tels qu'Amazon.com. Amazon propose des articles susceptibles d'intéresser l'utilisateur en

s'appuyant sur les articles que l'utilisateur a déjà acheté. La technique se base sur l'historique d'achats d'autres utilisateurs pour trouver les articles fréquemment achetés avec les articles du panier. Si les deux produits, rendus plus proches grâce à la recommandation, sont complémentaires, Amazon le propose à l'utilisateur, lui évitant ainsi de parcourir le catalogue d'Amazon.

Le réseau social Facebook a une approche similaire à Amazon. Il recommande « des personnes que vous êtes susceptible de connaître ». Cette recommandation pourrait se faire en fonction de la similarité des connections sociales des internautes selon l'hypothèse que les personnes membres des mêmes cercles sociaux sont susceptibles de se connaitre. De nombreux autres rapprochements sont effectués par différents systèmes de recommandation sur le Web. Une récapitulation très complète des approches techniques utilisées dans ces systèmes peut être trouvée dans [Adomavicius & Tuzhilin, 2005] et [Rao et al., 2010].

Il est important de préciser que ces quelques exemples ont pour objectif d'illustrer l'idée générale des rapprochements, et ne représentent pas une récapitulation exhaustive des travaux effectués sur ce sujet.

3.3. Les éléments du graphe Web dont le rapprochement est utile dans les scénarios liés à l'innovation

Les problèmes d'innovation et leurs différentes représentations sur le Web (e.g. pages Web, etc.) ont sans doute un rôle important dans les scénarios d'innovation sur le Web. Néanmoins, l'innovation elle-même repose sur des acteurs humains, souvent utilisateurs du Web. Nous nous intéressons notamment aux parties du graphe Web qui caractérisent ces acteurs humains et leurs interactions avec le Web. Nous nous intéressons notamment aux parties du graphe Web qui caractérisent ces acteurs humains et leurs interactions avec le Web. L'ensemble d'éléments du graph Web que nous souhaitons rapprocher dans des scénarios de l'innovation consiste donc de ressources représentant :

- les problèmes d'innovation ;
- les utilisateurs, auteurs potentiels de solutions;
- les traces d'activité des utilisateurs, qui seront définis plus en détails dans la partie suivante ;
- le sujet d'intérêt/compétence, i.e. des concepts.

Nous détaillons ci-dessous la notion de trace utilisateur qui est au cœur de nos travaux. La focalisation sur cette notion nous donnera l'occasion de définir des autres éléments pertinents dans le contexte de l'innovation sur Web, à l'exception de la notion des problèmes, déjà introduite dans la partie §1.1.2.

3.3.1. Traces utilisateurs

Dans leur pratique quotidienne sur le Web, les utilisateurs interagissent avec le graphe et par conséquent, l'enrichissent en y insérant de nouveaux éléments. Cette pratique est amplifiée par l'émergence des microposts et des activités à faible effort dont la conséquence est la production des contenus Web. Si celle-ci est principalement motivée

par la socialisation sur le Web, elle donne également lieu à une analyse des centres d'intérêt de l'utilisateur et même de ses compétences. Introduisant une transparence sur les compétences d'un utilisateur, ces éléments sont d'une utilité énorme pour l'émergence de l'innovation sur le Web. Nous les appellerons les « traces utilisateur ».

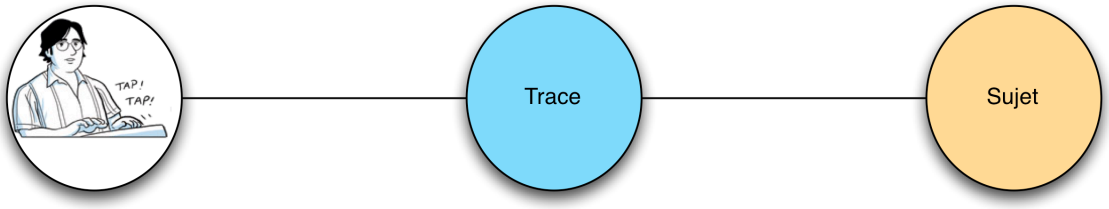


Figure 17 Exemple de trace (triplet utilisateur-trace-sujet) sous forme graphique

Une trace utilisateur[62] *est un élément Web, ou bien un ensemble d'éléments Web, qui relient un utilisateur et un concept - sujet de connaissance humaine.* La Figure 17 illustre ce triplet conceptuel utilisateur-trace-sujet. A titre d'exemple, un article de blog sur la gestion d'entreprise peut être considéré comme trace. Une page Web annonçant la participation d'un utilisateur à un congrès et dont les interventions portent sur le Web Sémantique est un autre exemple, un peu plus complexe mais néanmoins réaliste. Les sujets des traces jouent, bien évidemment, un rôle important dans l'appréciation de l'opportunité d'engagement d'un utilisateur dans un processus d'innovation particulier.

L'innovation est une activité générale, relative aussi bien aux domaines techniques qu'aux sciences humaines. Il est donc nécessaire de considérer tous les domaines de connaissance humaine en tant que sujets possibles des traces utilisateurs. Dans cette thèse et dans le contexte du graphe Web, nous identifions ces domaines par des *concepts*. *Un concept est donc un élément du graphe Web, identifié par son URI, désignant un sujet de connaissance humaine*.

Si l'utilité des traces laissées dans le Web public est très évidente, il ne faut pas négliger l'importance de traces laissées de manière privée, partagées seulement avec certains cercles sociaux et accessibles seulement par certaines applications. La protection de la vie privée de l'utilisateur est d'une importance primordiale pour la survie du Web et pour la confiance que les utilisateurs lui accordent. Naturellement, accorder aux utilisateurs la possibilité de contrôler l'accès aux données qu'ils créent devient une pratique de plus en plus présente dans les applications Web[63]. Cela est même devenu une obligation légale dans certains pays, dont la France[64]. Selon l'hypothèse que l'utilisateur a intérêt à participer à un processus d'innovation sur le Web, quelle que soit sa nature, il est raisonnable de s'attendre à ce qu'il laisse une application Web accéder à ses traces. Celle-ci pourrait les utiliser pour impliquer l'utilisateur dans un processus où

[62] Il est important de noter que notre notion de trace utilisateur est différente de celle utilisée dans les modèles d'usage d'ordinateur dans les travaux issus de l'interaction homme-machine, comme apr exemple chez [Prié, 2011]

[63] Facebook, par exemple, continue à faire evoluer sa ploitique de vie privée afin de satisfaire des demandes des utilisateurs qui sont de plus en plus conscients de dagners de leur vie exposée en ligne. http://www.facebook.com/help/privacy.

[64] Conformément à la loi du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés, les personnes concernées par des données disposent d'un droit d'accès, de modification, de rectification et de suppression de données qui les concernent.

il pourrait valoriser ses compétences et dont il pourrait lui-même bénéficier. Quel que soit son niveau d'accessibilité, une trace est donc utile pour le processus d'innovation sur le Web. Elle permet également le rapprochement d'un utilisateur à d'autres éléments du graphe Web qui peuvent lui être utiles.

Il est important de souligner que le contenu Web n'apporte pas toujours des traces bien structurées, ce qui rend la distinction du concept et de l'utilisateur concerné difficile. Il est parfois nécessaire, surtout avec des éléments de Web de documents, d'effectuer des transformations sur le contenu brut pour pouvoir l'utiliser sous forme de traces structurées. Par la suite nous aborderons les transformations que nous avons effectuées et leur nécessité en cas de traitement de certains types de contenu Web.

Enfin, il est important de noter que la notion de trace utilisateur désigne principalement la partie du graphe Web qui concerne les scénarios de l'innovation. Cette notion n'implique ni l'existence d'une expérience, ni l'existence d'une expertise, ni l'intérêt de l'utilisateur pour le concept en question. Les traces peuvent néanmoins souvent jouer un rôle dans l'appréciation de telles qualités. Nous rappelons également que dans le cadre de cette thèse, les notions de compétence, d'expertise et d'intérêt n'ont pas un sens classique, mais plutôt un sens influencé par le contexte de l'innovation ouverte et de l'innovation sur le Web, précisé par la définition dans la partie §1.3.4, page 18.

3.3.1.1. Abstractions : profil social et profil conceptuel

Afin de rendre plus compréhensible nos analyses des différents phénomènes rencontrés sur le Web Social et Sémantique nous introduisons une abstraction - la notion de profil utilisateur, à laquelle nous ferons référence à plusieurs reprises dans ce mémoire. De manière générale, dans le sens de nos travaux, nous entendons par profil utilisateur *un ensemble d'éléments du graphe Web qui est en quelque sorte associé à l'utilisateur*. La nature de cette association et des éléments peut varier selon le type de profil. Dans ce mémoire nous allons principalement nous intéresser à deux types de profils, social et conceptuel. Par profil social, *PS,* nous entendons un ensemble d'utilisateurs $\{u_1, u_2 \ldots, u_n\}$ avec lesquels l'utilisateur *u* concerné par le profil, est connecté par une relation sociale (d'amitié, familiale, de travail, etc.).

$$PS(u) = \{u_1, u_2, \ldots, u_n\} \quad (2)$$

L'ensemble des concepts $\{c_1, c_2, \ldots, c_n\}$[65] liés à l'utilisateur *u* (représentant par exemple les sujets de son intérêt ou bien de son expertise) constitue le profil conceptuel, *PC*. En fonction de besoins particuliers dans un cas d'utilisation, ce profil peut être constitué de manière différente en fonction des traces utilisateurs. Différentes techniques seront présentées dans des chapitres suivants, dans le contexte des besoins qui y sont considérés.

$$PC(u) = \{c_1, c_2, \ldots, c_n\} \quad (3)$$

De manière plus générale, il est possible d'associer avec des concepts une quantification de leur importance par rapport à l'utilisateur eu question. Un profil conceptuel représenté en tant qu'un ensemble de couples <concept,importance> est donné par la formule 3. Il est également possible d'associer d'autres informations avec les concepts dans les profils, mais cela dépasserait les besoins de nos cas d'utilisation.

[65] Les concepts sont représentés par leurs URIs

$$PC(u) = \left\{(c_1,i_1),(c_2,i_2),...,(c_n,i_1)\right\} \quad (4)$$

Par analogie il est également possible de construire des profils sociaux comprenant des valeurs d'importance, mais ceux-ci ne seront pas utilisés dans ce mémoire. Nos définitions de profils sont assez abstraites, notamment sur la manière dont ils sont créés et sur la définition d'éligibilité des éléments à paraître dans les profils. Cette généralité vise principalement à établir une uniformité de notion qui serait utilisée sous différentes formes dans différentes parties de nos travaux selon les besoins.

3.3.1.2. La genèse de traces - les activités

L'enrichissement des profils sociaux et conceptuels est permanent avec l'usage du Web au quotidien. Les activités de l'utilisateur sont à l'origine de cet enrichissement d'une double nature et qui se manifeste par:

- l'évolution unilatérale des profils ;
- la coévolution des profils sociaux et conceptuels.

En interagissant avec des éléments du Web liés à ces centres d'intérêt, l'utilisateur découvre de nouveaux intérêts et enrichit son profil conceptuel. Similairement, par l'intermédiaire des utilisateurs faisant partie de son profil social, il découvre de nouveaux amis et crée de nouveaux contacts.

Si les activités sont d'une importance critique pour l'enrichissement des profils, cet enrichissement est encore amplifié par une interdépendance du profil social et du profil conceptuel qui est à l'origine de leur coévolution, et cela dans les deux sens :

- **Le profil conceptuel influence le profil social** : C'est en interagissant avec des éléments du graphe Web liés à des concepts de leur intérêt que les utilisateurs établissent de nouveaux liens sociaux. Par exemple, en lisant un billet de blog relatif à la macroéconomie, un utilisateur peut découvrir des personnes qui ont laissé des commentaires sur ce même blog et qui partagent donc son intérêt pour le sujet. Ce phénomène est bien connu sous le nom « socialisation centrée sur des objets », phénomène décrit plus en détails dans la partie §2.2.
- **Le profil social influence le profil conceptuel** : C'est en interagissant avec les autres utilisateurs qu'un utilisateur peut découvrir de nouveaux intérêts et même développer de nouvelles compétences en apprenant de ces contacts.

Si les profils utilisateur sont composés des éléments du graphe Web, une interdépendance des deux types de profils implique certainement une interdépendance des sous-graphes du graphe Web desquels ils sont dérivés. Le graphe social et le graphe conceptuel du Web coévoluent donc grâce aux activités des utilisateurs (Figure 18). Cette coévolution est notamment évidente dans des cas où les activités des utilisateurs sont représentées par la création de traces utilisateur, tels que nous les avons définies dans ce chapitre.

Imaginons par exemple qu'un utilisateur lise un billet de microblog sur un sujet qui l'intéresse et écrit par un utilisateur inconnu auparavant. Il aime le billet et décide de le republier auprès de son propre cercle d'amis. En même temps, il se met à suivre l'auteur du billet de microblog initial. Cette situation, très commune dans le Web d'aujourd'hui, décrit un enrichissement du profil social de l'utilisateur qui est suivi par une trace concrète, ici le billet de microblog republié, et qui établit un nouveau lien indirect entre

l'utilisateur, auteur du billet de microblog, et celui qui a effectué la republication. Le billet republié crée également un lien entre l'utilisateur et les concepts éventuellement liés à ce billet.

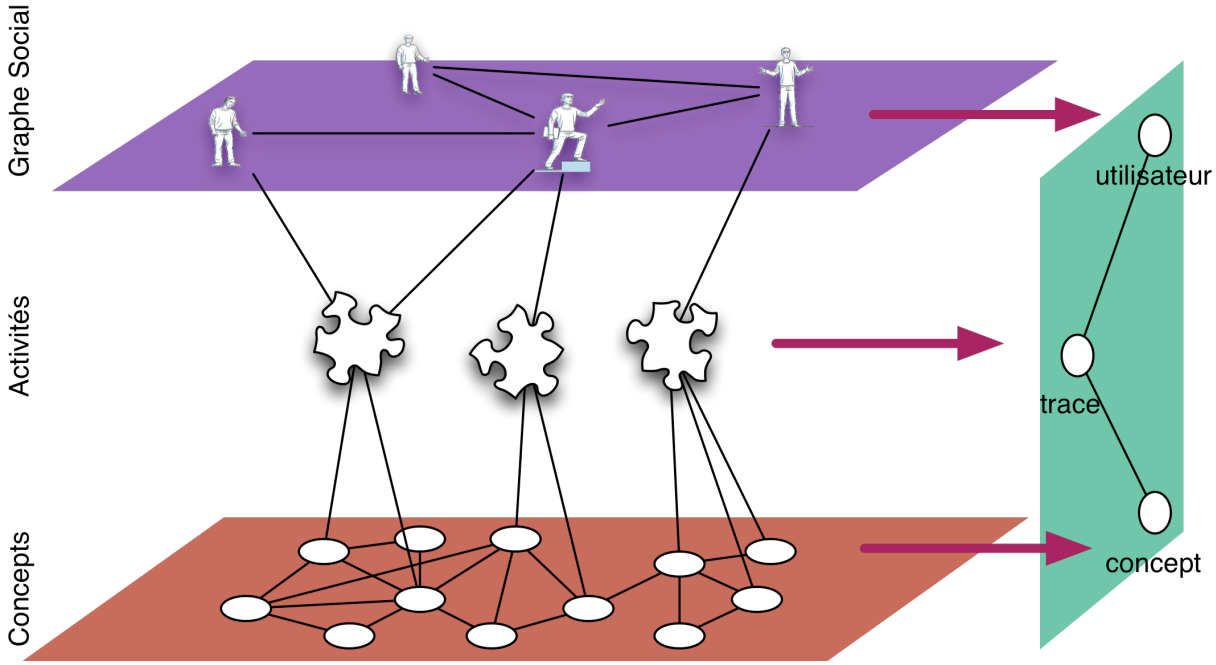


Figure 18 Interdépendance entre le graphe social et le graphe conceptuel à travers les activités des utilisateurs

Les traces émergent en conséquence des activités, et peuvent faire l'objet des nouvelles activités qu'ils inspirent chez les utilisateurs qui les consomment. Elles peuvent donc être perçues comme un lien entre le plan social et le plan conceptuel du graphe Web, conduisant le Web à une cohésion plus élevée. Il ressort d'une telle influence un Web dans lequel il est possible de trouver des éléments qui représentent des utilisateurs, des concepts, des traces utilisateurs et d'autres types d'objets associés par des liens de types différents. Ces traces, et leurs éléments joueront un rôle principal dans les démarches visant à faciliter l'émergence de l'innovation sur le Web.

Toutes les différentes interdépendances évoquées rendent évident un mécanisme plus général de traçabilité des processus cognitifs chez des utilisateurs qui émergent grâce au Web Social Sémantique. La sémantique très explicite des traces structurées expose certaines connaissances de l'utilisateur, en termes de sujets qui lui sont connus ainsi qu'en termes des personnes qu'il connaît dans le cas des traces sociales. Une trace reflète un processus de réflexion sur des éléments concernés par la trace. D'un autre coté, la consommation des traces peut générer de nouvelles connaissances. Les traces sont donc utiles pour la création de la connaissance et vice versa.

La nature sociale des activités sur le Web contribue à la prolifération des activités traçables, et la nature sémantique et structurée des traces leur donne une utilité supérieure. Ces deux aspects, sémantique et social, garantissent un certain dynamisme du Web en termes de capacité d'évolution et de partage de connaissances. C'est pourquoi le Web Sémantique Social, et notamment les traces utilisateurs qui en font partie, sont d'un intérêt vital pour la construction des systèmes visant à faire émerger l'innovation et la recherche.

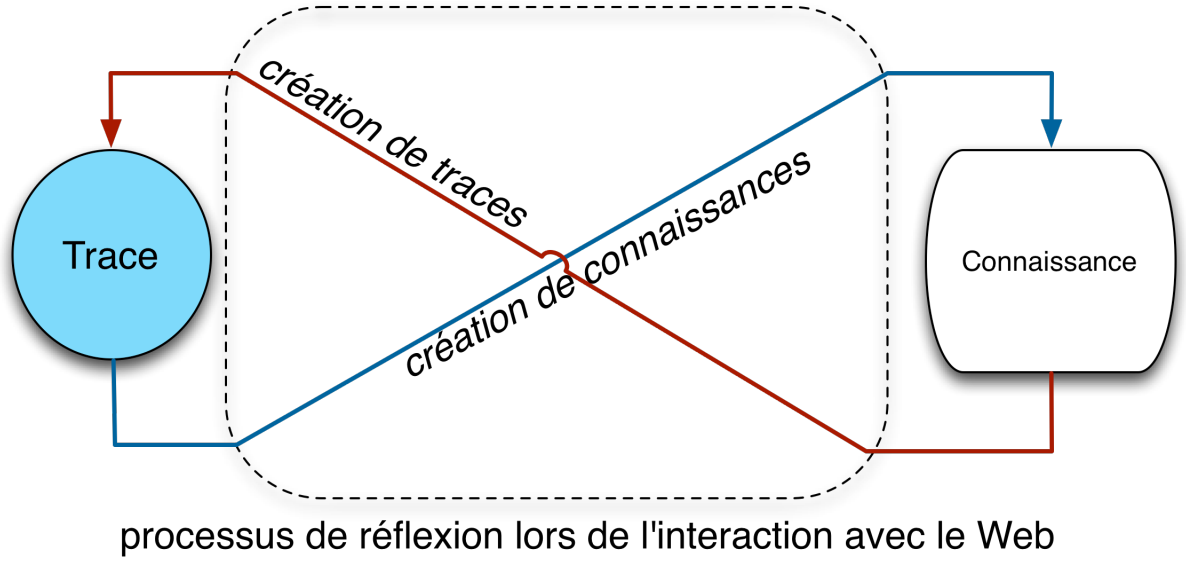


Figure 19 Interdépendance des traces et des connaissances

Dans les travaux présentés dans ce mémoire, nous exploitons fréquemment la nature dualistique, sociale et conceptuelle, de l'évolution du graphe Web, à la fois en se focalisant sur des traces que émergent de cette dualité, et en utilisant des profils conceptuelles et sociales.

3.4. Rapprochement dans le contexte du Web de données

3.4.1. Les avantages de Web de données pour la pratique des rapprochements

Si le graphe de liens entre des pages Web est utilisable pour mettre en place des rapprochements entre des éléments de la couche de documents, la question des rapprochements dans la couche de données reste peu abordée. Cependant, la richesse de types de liens, et notamment la présence des traces et leur nature sociale dans la couche de données, rendent cette couche particulièrement apte à la conception des rapprochements diverses et raffinés.

La Figure 15, illustre la projection des utilisateurs, de leurs activités et des contenus sous forme de traces ou d'autres ressources (représentées par des cercles sur la figure). Ces ressources se situent sur la couche de données, i.e. sur la partie « Web Sémantique Social » de cette couche. En étudiant cette projection, il devient évident que cette couche bénéficie d'une richesse de liens plus élevée et plus descriptive que les ressources elles-mêmes. Une ressource représentant une trace utilisateur par exemple, peut être liée à une ressource représentant l'utilisateur créateur de la trace, elle-même liée à une autre ressource représentant un concept qui est un sujet d'intérêt pertinent pour la trace. Ces ressources sont de types différents et sont liées par des liens de types différents. Certaines ressources, telles que les concepts, sont exclusivement présentes dans la couche de données, ce qui la rend non seulement plus structurée mais aussi plus riche en information. Cette richesse à la fois structurelle et informationnelle, devrait également rendre envisageable un certain nombre d'approches de traitement du graphe

Web qui seraient plus intelligentes par rapport à ceux qui s'appuient uniquement sur la couche de documents.

Notre hypothèse principale est donc la suivante : *en s'appuyant sur la diversité des types de ressources et des types de liens entre eux, il est possible de concevoir des systèmes capables de rapprocher deux éléments Web de manière utile par rapport à une situation particulière. Ce rapprochement est fondé sur des hypothèses sur la pertinence des éléments et sur l'importance de leurs liens.*

Autrement dit, un rapprochement peut exploiter la richesse de la couche de données pour réduire les distances entre des éléments Web qui sont pertinents l'un pour l'autre en favorisant leur accessibilité.

Des hypothèses de pertinence, dans ce contexte de graphes typés, identifient certains types de liens, de chemins, et d'éléments comme plus pertinents que d'autres, permettant ainsi de déduire une notion de proximité entre les éléments du graphe. Ce raisonnement dépasse la simplicité des mesures basées uniquement sur la longueur des chemins et de propriétés statistiques. Un exemple d'hypothèse de pertinence dans ce contexte peut être donné par la règle suivante : deux concepts sont pertinents l'un pour l'autre s'ils sont liés directement à la même ressource de type blog. Dans le cas de cette hypothèse c'est l'usage commun des mêmes contenus qui définit la pertinence. Nous verrons par la suite les hypothèses basées sur les caractéristiques des liens, sur la connectivité des ressources dans le graphe, et sur d'autres critères encore.

Dans la suite de ce chapitre, nous considérons le cas spécifique des rapprochements entre les éléments de types différents et pertinents pour des scénarios liés à l'innovation ouverte.

3.4.2. Rapprochements abordés dans cette thèse

Dans la partie précédente nous avons expliqué comment les activités génèrent des traces utilisateurs et font évoluer le Web Social Sémantique. Ces traces sont à l'origine des chemins qui relient les informations à travers le Web. Différents cas de rapprochement des éléments Web dans les scénarios liés à l'innovation

Pour faire émerger l'innovation sur le Web, il est notamment intéressant d'effectuer des rapprochements entre les éléments du graphe Web liés à l'innovation. Nous rappelons que ces éléments, sont notamment :

- utilisateurs ;
- problèmes d'innovation ;
- traces utilisateurs ;
- concepts.

Dix rapprochements (r1…r10 sur Figure 20) sont théoriquement possibles entre ces quatre éléments. Nous avons isolé, parmi ces dix rapprochements, un nombre de scénarios de rapprochements particulièrement intéressants pour les plateformes d'innovation sur le Web, que nous avons favorisées dans nos travaux et que nous présenterons en plus de détails dans les sous-parties §3.4.2.1-§3.4.2.4.

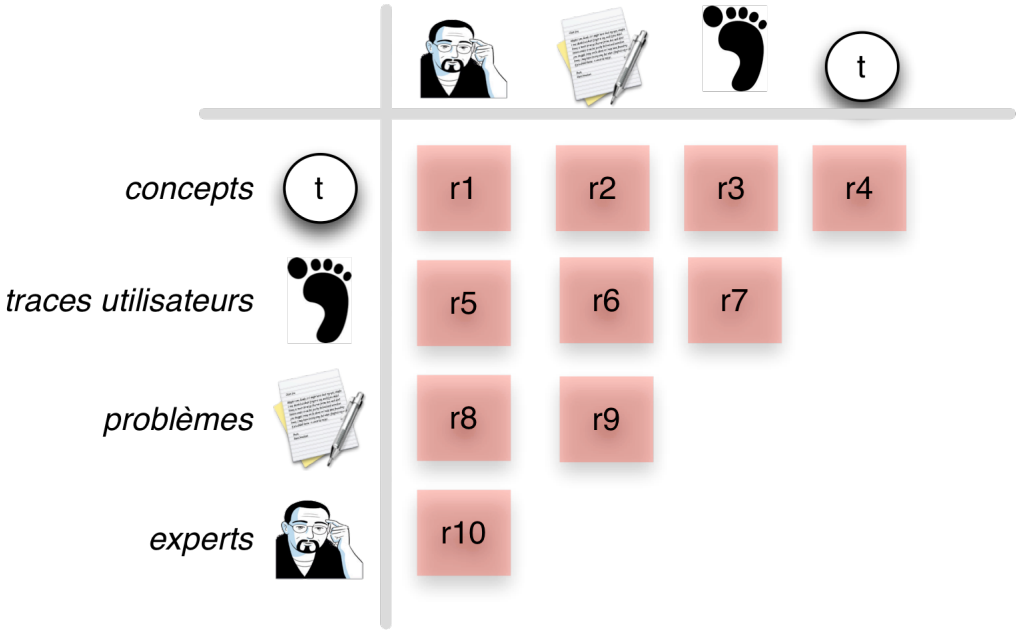


Figure 20 Rapprochements théoriquement possibles entre des éléments du graphe Web liés à l'innovation

Certains rapprochements, théoriquement possibles, restent peu nécessaires dans la pratique, grâce à la nature du Web qui lie déjà entre eux ces éléments de manière très directe. Par exemple le rapprochement r5 entre des traces et des utilisateurs est peu utile étant donné que les traces sont par définition créées par un utilisateur. Il est plutôt question de mettre en relation du contenu Web avec les utilisateurs concernés dans des cas où ce lien n'est pas déjà établi, afin d'en construire des liens éventuellement manquants entre des traces et les utilisateurs qu'elles concernent. Le rapprochement entre des utilisateurs et des traces des autres utilisateurs, malgré son incontestable utilité n'est pas abordé dans cette thèse. Celui-ci pourrait être effectué en s'inspirant de nos travaux qui concernent le rapprochement r7, discutés par la suite, qui font face aux mêmes défis de recherche. Les rapprochements r2 et r3, entre des problèmes/traces et des sujets sont rendus inutiles dans le cas d'éléments déjà annotés par des concepts. Dans d'autres cas, si ces problèmes/traces contiennent des éléments textuels, il est possible de dériver ce lien de manière automatique ou semi-automatique à l'aide des outils d'extraction décrits dans la section §2.4.1.1.

D'autres rapprochements posent des défis plus importants sur lesquels se focalise ce travail doctoral. Plus précisément nous aborderons les défis suivants :

- **Rapprocher de potentiels auteurs de solutions et experts concernés d'un concept donné** (rapprochement r1). Ce rapprochement vise d'abord à identifier des experts et les rapprocher de concepts décrivant leur compétence.
- **Rapprocher des concepts représentant des sujets connexes d'un ensemble de concepts initiaux donné** (rapprochement r4). Ce rapprochement vise à satisfaire le besoin de latéralité[66] et de diversité dans des scénarios liés à l'innovation sur le Web. Dans une multitude de tâches liées à l'innovation sur le Web il est nécessaire d'identifier des concepts pertinents de manière latérale,

[66] Par lateralité nous comprenons la mise en connection des idées, solutions, personnes provenant des spheres intellectuelles et concepteulles differentes mais neanmoins pertinentes

pour ensuite par exemple trouver des experts dans des domaines connexes ou diffuser un problème.

- **Rapprocher des éléments du graphe Web généré par des utilisateurs des autres éléments** qui pourraient enrichir son contenu informationnel et le transformer en une trace utilisateur utile (rapprochement r7).
- **Rapprocher des potentiels collaborateurs d'un utilisateur engagé dans la résolution d'un problème d'innovation** (rapprochement 10). Ce rapprochement vise à mettre en valeur la nature sociale du processus de l'innovation et faciliter la création d'équipes. Ce rapprochement a un caractère complexe dû à son rôle multidimensionnel : rapprocher un auteur de solutions des personnes dont les compétences seront complémentaires, relativement à un problème donné (rapprochements r8, r2, r3) et qui soient compatibles avec l'utilisateur pour effectuer un travail collaboratif (r10).

Nous présentons ces quarte rapprochements en plus de détails dans les sous-parties suivantes, car ceux-ci sont élaborés très amplement dans les quarts chapitres suivants de la thèse. La répartition des rapprochements par chapitres est donnée par le schéma présenté sur la Figure 21.

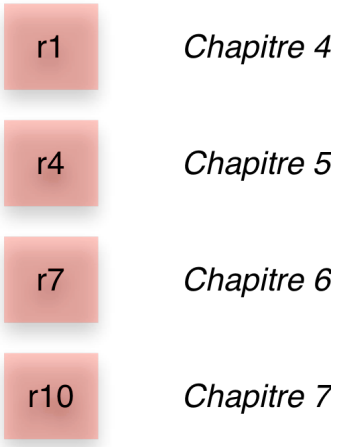


Figure 21 Chapitres dans lesquelles les rapprochements sont présentés

Le rapprochement r8 émerge comme une conséquence naturelle de la suite d'autres rapprochements : rapprochement de personnes et concepts (r1), celui des problèmes et concepts (r2) et celui des concepts entre eux (r4). Un rapprochement prenant en compte autant de paramètres viserait principalement à augmenter la visibilité d'un problème auprès des personnes qui sont susceptibles d'y contribuer. Si un mode d'interaction qui comprend la recommandation de problèmes aux experts est le plus évident, il ne faut pas exclure les modes dans lesquels un internaute cherche explicitement des opportunités pour mettre en valeur ses compétences. Ce rapprochement est au cœur du scénario de l'innovation sur le Web, et par conséquent celui-ci est pratiqué à plusieurs reprises dans nos travaux seule ou en combinaison avec des autres.

Les rapprochements r6 et r9 ne sont pas abordés dans cette thèse. Si le rapprochement entre des traces et des problèmes (r6) est tout à fait envisageable et même potentiellement utile pour des auteurs de solutions qui peuvent s'inspirer des traces rattachées à leur problème, la complexité de cette rapprochement ne représente pas un grand défi. Nous considérons que les démarches présentées dans cette thèse, visant à

effectuer d'autres rapprochements, pourront avec quelques modifications, inspirer la réalisation de ce rapprochement.

A l'inverse, le rapprochement r9, entre des problèmes, représente un défi beaucoup plus important, nécessitant une démarche qui dépasse le cadre de cette thèse. Pour aller au-delà d'un simple rapprochement en fonction des concepts que deux problèmes partagent, il faudrait analyser la structure intrinsèque du problème. Nombreux sont les travaux qui ont été menés à ce sujet dans le domaine de la gestion de l'innovation mais aussi dans le domaine de l'intelligence artificielle. La représentation des compétences fais souvent partie des travaux effectués dans le cadre d'analyse de la structure des problèmes, et nous connaissons ainsi la complexité d'une telle tâche et les débouchées peu certaines de cette démarche. Parmi les approches de la gestion de l'innovation nous trouvons TRIZ [Mann, 2001], une méthodologie d'innovation visant à systématiser la créativité et qui définit le problème comme un conflit de forces, une contrainte. Une vue similaire est prise dans la classification des problèmes architecturaux en Design Patterns – solutions typiques des problèmes typiques [Alexander, 1979], qui définit les problèmes comme des conflits des contraintes spatiales et besoins d'usage. De nombreux travaux sont issus de l'intelligence artificielle et notamment de l'effort de produire un système capable de résoudre des problèmes de manière automatique [Russel & Norvig, 1995]. Si la notion des problèmes dans ces travaux ne corresponde pas entièrement à la nature industrielle des problèmes avec lesquelles nos sommes confrontés dans nos travaux, il est néanmoins possible que ces travaux pourront servir d'inspiration à de futurs efforts de décrire l'essence des problèmes et de les comparer entre eux. Une telle description de problèmes devrait permettre de découvrir des problèmes analogues et ainsi faciliter leur résolution. Malgré l'existence de quelques pistes, une telle possibilité reste un rêve à l'heure actuelle, principalement en raison de la jeunesse de la pratique de résolution de problèmes en ligne qui n'a pas à ce jour produit suffisamment de problèmes analogues pour pouvoir analyser et identifier des caractéristiques rendant deux problèmes similaires, analogues ou tout simplement comparables.

Par la suite nous présenterons des rapprochements auxquels nous avons accordé le plus d'attention en restant toujours au niveau abstrait, pour ensuite présenter dans les chapitres qui suivent des détails de nos implémentations et évaluations.

3.4.2.1. R1 : Rapprochement des concepts et des experts

Ce rapprochement a pour objectif de rapprocher des concepts d'expertise avec un utilisateur, afin de construire son profil conceptuel, d'une manière pertinente par rapport à une situation donnée. Ce rapprochement repose notamment sur différents types de traces et différents sujets. Notre travail autour de ce rapprochement se base principalement sur l'hypothèse que différents types de traces sont utilisés différemment par différentes communautés d'experts et que, par conséquent, cette invariance pourrait donner lieu à une identification des « bons » types de traces à suivre pour trouver des experts pertinents sur un sujet spécifique. Le chapitre 4 est dédié à ce rapprochement.

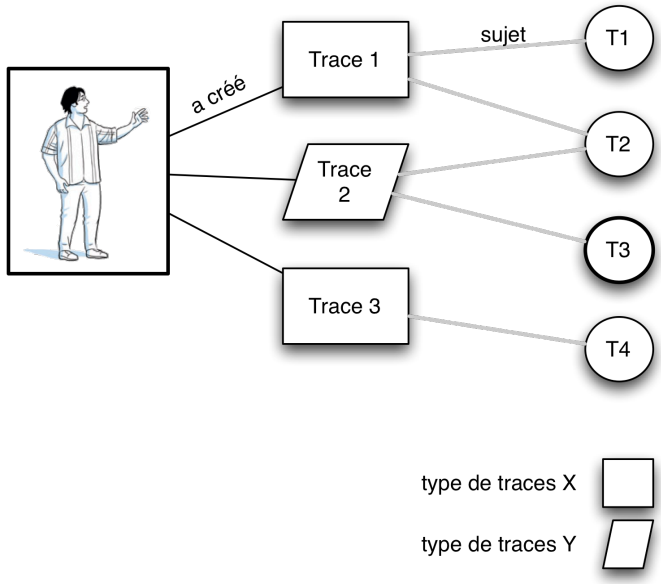


Figure 22 Rapprochement basé sur des types de traces

3.4.2.2. R4 : Rapprochement des concepts entre eux

L'objectif de ce rapprochement est de rapprocher un concept à un autre dans un graphe de concepts connectés et cela d'une manière pertinente par rapport à une situation. Dans notre approche, nous exploitons particulièrement la pluralité de types de liens dans le graphe de concepts, et nous considérons des chemins entre des concepts établis à travers différents liens comme ayant une importance différente. C'est ici que la richesse des données liées, incomparable avec celle des autres sources de concepts, permet de montrer l'utilité d'une représentation de connaissance à la fois riche et bien structurée.

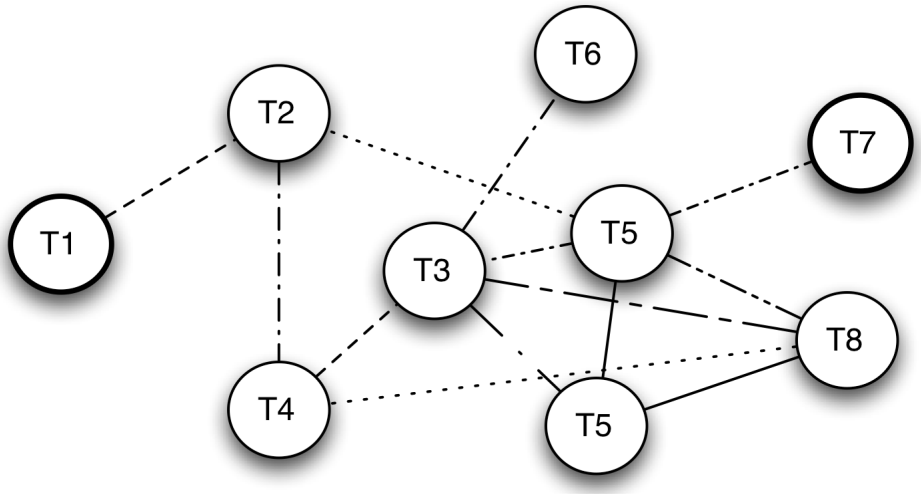


Figure 23 Rapprochement des concepts utilisant différents types de liens

Confronté à un graphe de concepts assez conséquent qui est disponible aujourd'hui au sein du LOD, il est question de trouver une notion de proximité de concepts adaptée à des scénarios liés à l'innovation. La spécificité de ces scénarios réside notamment dans

le besoin de considérer comme proches des concepts qui ne sont pas perçus comme liés à un même « domaine de compétence » ou parcours de formation mais qui peuvent être pertinents pour la résolution d'un même problème. Il est important de souligner que la notion de domaine de compétence est, à notre avis, la conséquence d'une perception souvent subjective de la classification des domaines, et qu'elle doit être évitée dans toute démarche sérieuse visant une objectivité de vue. La répartition de la science en domaines, et même l'ampleur de la science elle-même, sont perçues très différemment selon les communautés de pratique ou les personnes de parcours et formation différents. Cet effet peut être observé sur l'exemple des différentes taxinomies de domaines de compétences, y compris la taxinomie des domaines informatiques par ACM[67] qui ne contient même pas une catégorie « Social Web », qui est à notre avis importante. L'effort de classifier des domaines ne date pas d'hier, et pourtant il est difficile de dire qu'un accord universel sur une classification a émergé. En cette raison nous avons abandonné tout espoir de modéliser cette notion de domaine, et nous travaillons avec les concepts et leurs ensembles. Heureusement, les ensembles de concepts suffisent amplement pour situer dans l'espace conceptuel des objets de notre intérêt tels que les problèmes, les experts etc. Le chapitre 5 est dédié à ce rapprochement.

3.4.2.3. R7 : Rapprochement visant à enrichir des traces utilisateur

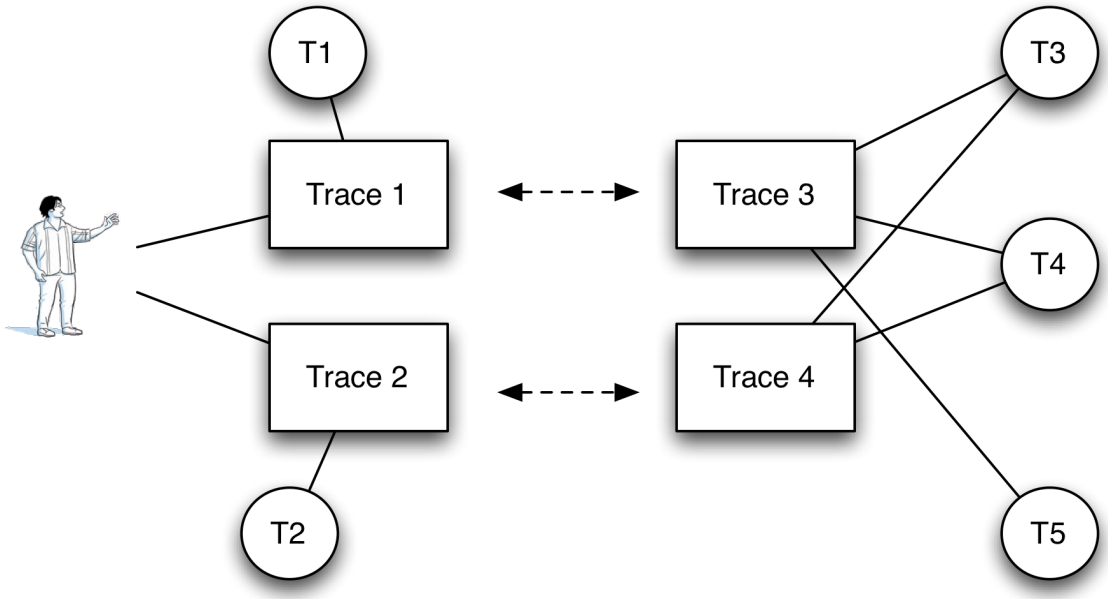


Figure 24 Rapprochement basé sur des mises en correspondance des traces

Ce rapprochement visant à enrichir des traces se base principalement sur des techniques de mise en correspondance des traces entre elles. Cette mise en correspondance pourrait s'avérer notamment utile dans le cas des traces à faible contenu informationnel, comme des tweets. Une fois les tweets mis en correspondance avec un autre élément Web auquel ils font référence, un article scientifique par exemple, il est possible d'accéder à une pluralité de concepts, associés aux deux traces (l'initial et la trace cible). Cet enrichissement est susceptible d'être utile dans la création des profils

[67] http://www.computer.org/portal/web/publications/acmtaxonomy

utilisateurs et dans plusieurs autres scénarios. Le chapitre 6 est dédié à ce rapprochement.

3.4.2.4. R10 : Rapprochement des experts entre eux

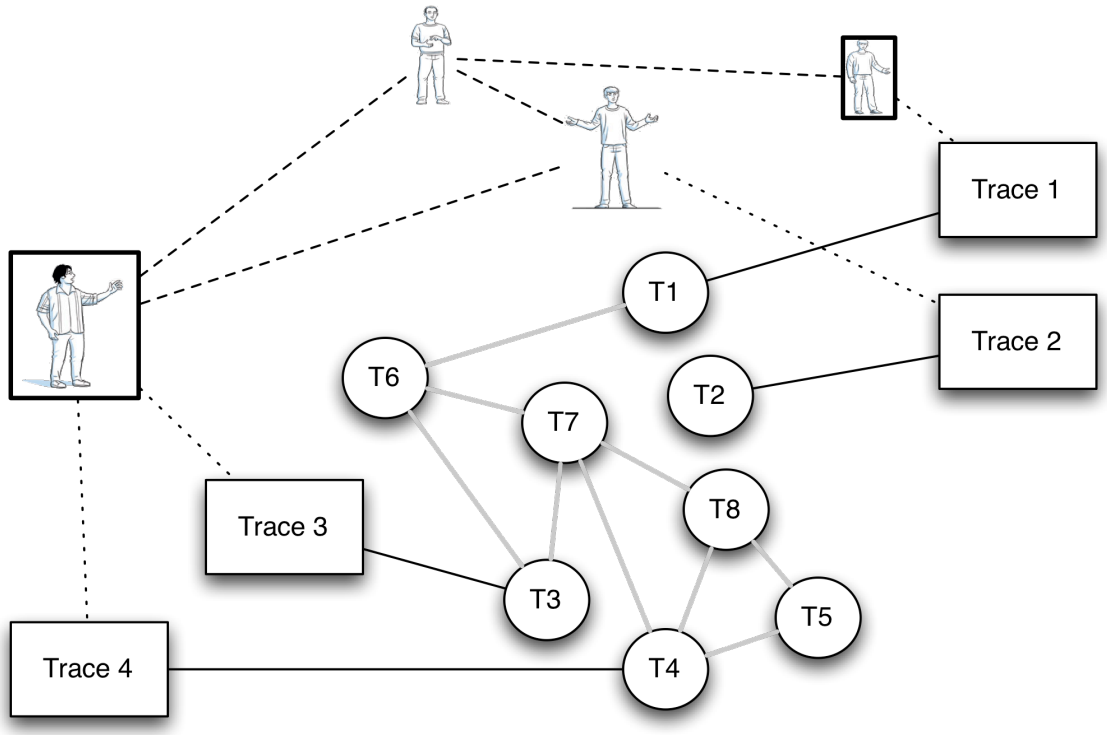


Figure 25 Rapprochement en plusieurs dimensions

Le rapprochement des experts dans le cadre d'une collaboration est particulièrement complexe parce qu'elle implique une notion de proximité qui dépend de plusieurs facteurs. Deux collaborateurs proches dans le contexte d'un problème ne le sont pas dans le contexte d'un autre. Il est donc nécessaire de développer une notion de pertinence qui reflète à la fois la compatibilité des personnes et leur complémentarité par rapport à un problème d'innovation.

La complexité de ce rapprochement émerge également de la nature des chemins qui séparent deux collaborateurs, qui comprennent des liens sociaux entre personnes, des liens entre des personnes et des contenus, contenus et concepts ainsi que des liens entre concepts entre eux. Notre approche consiste donc à combiner plusieurs rapprochements entre ces différents éléments. Nous nous appuyions sur des rapprochements déjà développés pour des situations précitées, et nous en développons des nouveaux, spécifiques à la situation en question.

Ce rapprochement composé exemplifie la combinaison des rapprochements élémentaires, applicables dans une gamme de situations réelles dont la complexité est plus élevée que celle de la plupart de nos exemples précédents. Le chapitre 7 est dédié à ce rapprochement.

3.4.3. Quelques caractéristiques du rapprochement des éléments Web

3.4.3.1. Effets des rapprochements dans les scénarios liés à l'innovation

Dans la partie §3.2.1 nous avons brièvement évoqué l'utilité d'un rapprochement des éléments Web comme étant principalement liée à l'accessibilité de l'information dans le graphe Web. Si cette raison motive le rapprochement dans une large gamme de scénarios, les scénarios liés à l'innovation impliquent un certain nombre d'autres effets désirables et envisagés par le rapprochement. Dans cette partie nous donnons une vision plus complète de tous ces effets.

- Bien évidement, **augmenter l'efficacité d'accès à l'information** est la motivation principale des rapprochements des éléments du graphe Web. Quelle que soit la situation dans laquelle un utilisateur interagit avec le graphe Web, les informations qui lui sont utiles sont souvent éloignées de son point de repère dans le graphe et sont réparties dans des endroits différents. Le rapprochement réduit ces distances. Ce qui est utile à un internaute varie forcément en fonction du scénario d'usage. Les deux effets désirables présentés ci-après concernent notamment le scénario de l'innovation sur le Web.
- L'innovation sur le Web repose sur la participation d'utilisateurs. Le rôle d'un rapprochement est donc également de **faciliter la mise en valeur des capacités intellectuelles des internautes**. En rapprochant les utilisateurs avec des concepts, des collaborateurs potentiels, et des problèmes il est possible de stimuler leur participation en tant qu'auteurs de solutions.
- La vision de l'innovation sur le Web promet, entre autres, de faire émerger des transferts de connaissances à travers des « domaines » et d'inciter la récolte d'idées et de pistes de réflexions inattendues. Pour ce faire, il est nécessaire de pouvoir **stimuler la sérendipité** par des rapprochements des éléments Web.

Par sérendipité nous entendons l'effet d'effectuer des découvertes inattendues mais pertinentes. Elle émerge en conséquence soit d'un résultat non recherché mais néanmoins utile, soit d'une voie inattendue qui a mené à un résultat recherché. La notion de sérendipité est généralement peu abordée dans la littérature de recherche. Quelques travaux néanmoins tâchent de définir et étudier ce phénomène. Dans le domaine des sciences humaines, les travaux de Danièle Bourcier et Andel Pek Van [Andel & Bourcier, 2009] proposent une analyse des différentes occurrences des découvertes fortuites dans la science, en droit, dans l'art etc. Ils insistent notamment sur le fait que la sérendipité n'est pas un hasard mais qu'elle émerge en conséquence d'une synergie de plusieurs facteurs tels que la personnalité de la personne expérimentant la sérendipité, la préparation de l'esprit, etc. Les chercheurs de MIT [André et al., 2009], [André & schraefel, 2009] soutiennent également l'hypothèse que la sérendipité peut être stimulée et suggèrent des pistes de développement des interfaces utilisateurs qui augmenteraient la probabilité d'émergence de la sérendipité.

Il est également possible de trouver des descriptions de quelques systèmes de recommandation qui ont été conçus pour inciter la sérendipité. Iaquinta et al. [Iaquinta et al., 2008] proposent un système de recommandation de différents objets ayant une description textuelle qui est basé sur des heuristiques afin de stimuler la sérendipité.

Dans un système de recommandation de programme télévisé Kamahara et al. [Kamahara et al., 2005] proposent l'introduction de recommandations de programmes inattendus afin de permettre à l'utilisateur d'élargir le champ de ses intérêts. D'autre travaux comme [Ge et al., 2010], [Murakami et al., 2008], [McNee & Riedl, 2006] proposent d'évaluer la sérendipité dans les systèmes de recommandation. Malgré certaines différences au niveau de la méthode d'évaluation, dans ces travaux la sérendipité est généralement perçue comme une synergie de la pertinence et du caractère inattendu des recommandations. La nature des recommandations en fonction de ces deux dimensions peut être résumée par le diagramme, présenté sur la Figure 26. Les recommandations qui sont à la fois inattendues et pertinentes sont de caractère fortuit. Pour l'utilisateur elles représentent une découverte et sont les plus utiles puisqu'elles enrichissent l'utilisateur. La pertinence seule, sans cet aspect inattendu, résulte en recommandations attendues (en haut à gauche) qui sont néanmoins d'une certaine utilité. Les recommandations qui manquent de pertinence et qui sont très évidentes (en bas à gauche) ne devront en aucun cas être proposées par un système de recommandation. Celles qui sont inattendues, mais manquent de pertinence (en bas à droite) sont également d'une utilité très faible.

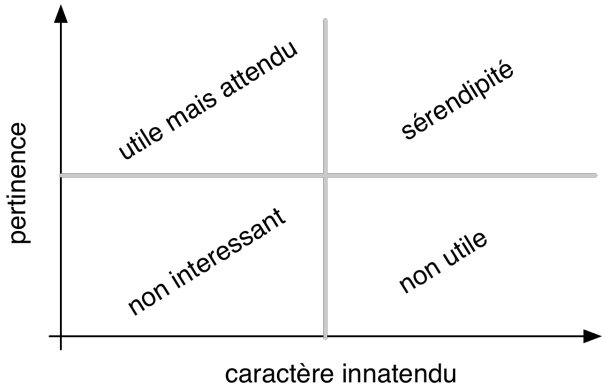


Figure 26 Rapport entre la pertinence et l'aspect inattendu des recommandations, diagramme inspiré par [Akiyama & Obara, 2010]

Si la pertinence d'une recommandation dans un contexte utilisateur reste assez stable dans le temps, le caractère inattendu à tendance à varier. Une fois l'utilisateur confronté à une recommandation dans un contexte spécifique, il peut s'attendre à avoir cette même recommandation la prochaine fois qu'il se trouve dans ce même contexte. Comme la perception d'une recommandation en tant qu'inattendue dépend de l'état d'esprit, de la connaissance et des expériences préalables de l'utilisateur il est difficile d'envisager la conception de systèmes capables de produire un tel effet, mais il est néanmoins possible, comme le montrent nos travaux, de concevoir des systèmes de recommandation favorables à la découverte inattendue qui la rendent plus probable que dans les systèmes focalisés sur la pertinence seule.

3.4.3.2. Rapprochements permanents et temporaires

Il est important de considérer les aspects temporels de rapprochements. Dans le cas général, le rapprochement peut être compris comme une vue temporelle sur le graphe Web et ses distances, n'affectant pas celui-ci de manière permanente. Dans un tel cas, celui-ci sert dans un moment précis à une application qui l'a effectuée, mais reste éphémère.

Dans des cas où l'utilité d'un rapprochement dépasse les besoins d'un moment particulier, il est possible de créer de nouveaux liens dans le graphe Web pour représenter le raccourcissement qu'un rapprochement introduit de manière permanente et réutilisable. Une description riche de ces nouveaux liens, permettant de les distinguer de liens qui existent indépendamment du rapprochement semble être une nécessité pour le bon fonctionnement des applications Web. Nous présenterons des manières de décrire ces nouveaux liens et les réintroduire dans le graphe Web dans le chapitre 8.

3.5. Travaux Liés

Étant donné la nature récente du Web de données et sa taille, devenue importante seulement récemment, il n'est pas étonnant de voir que le nombre d'approches trouvées pour effectuer les rapprochements des éléments de cette couche du Web soient assez faibles. Nous présentons dans cette partie les travaux qui ont émergé du monde de la recherche au cours des quelques dernières années. Ces travaux sont organisés en deux catégories :

- Les systèmes de recommandation qui utilisent le Web de données (partie §3.5.1);
- Les approches pour la découverte de liens dans le Web de données (partie §3.5.2).

Ces approches s'inscrivent dans le cadre général de rapprochements dans le Web de données. Même si elles ne s 'appliquent pas forcément au contexte de l'innovation, nous les comptons parmi les travaux liés.

Deux autres idées au cœur de nos travaux sont l'interdépendance du plan social et conceptuel du graphe Web et l'usage de ces deux aspects dans le cadre des rapprochements. Si de telles approches sont encore rares, un certain nombre d'approches basées sur les idées connexes qui nous ont, par ailleurs, partiellement influencés sont présentés dans la partie §3.5.3.

Si le nombre de travaux effectuant des rapprochements dans le Web de données reste faible à ce jour, la communauté de recherche a produit un vaste corpus de travaux pertinents pour chacun de nos rapprochements, parmi lesquels les systèmes de recommandation et les systèmes de recherche d'experts. Pour faciliter la lecture de ce mémoire, nous présentons les travaux liés dans les parties du mémoire concernées. Dans la présentation de chacun de nos rapprochements, nos nous referons aux travaux qui nous ont inspiré et à ceux qui ont tenté d'obtenir des résultats similaires aux nôtres.

3.5.1. Systèmes de recommandation et de recherche qui utilisent le Web de données

L'usage de données liées et notamment du LOD dans les systèmes de recommandation a émergé récemment. Par exemple, des systèmes de recommandations de musique ont été proposés dans [Passant et al., 2009] et [Passant, 2010] en 2009 et 2010. La recommandation se base sur des mesures de distance sémantique comme LDSD (en. Linked Data Semantic Distance) qui permettent de rapprocher deux ressources. Ces mesures quantifient l'importance des liens qui relient deux ressources dans le graphe de LOD. L'hypothèse d'une telle approche est qu'un utilisateur intéressé par une ressource

(e.g., un artiste) est susceptible d'être intéresse par d'autres ressources fortement interalliées avec celle-ci. Dans le cas des artistes, si deux artistes sont liés par la même ville dans laquelle ils ont commencé leur carrière musicale et par les mêmes collaborateurs, il semble raisonnable de penser qu'ils pourraient intéresser le même utilisateur. Le système de recommandation dbrec, construit autour de cette mesure, est évalué face à un système de recommandation basé sur le filtrage collaboratif (last.fm). Les résultats de dbrec ont été particulièrement appréciés par des utilisateurs de par la capacité de faciliter la découverte des artistes inconnus mais qui semblent intéressants.

L'application des mêmes approches de recommandation dans le domaine des livres a également été testée avec succès [Passant et al., 2009].

Une approche très similaire est employée dans [Waitelonis & Sack, 2009] pour concevoir un système de recherche de films qui permettrait aux utilisateurs d'explorer des films similaires à un film donné. Des heuristiques différentes sont utilisées pour rapprocher deux films. Par exemple, le système favorise des ressources liées au travers de certaines propriétés qui sont jugées pertinentes pour un cas donné.

Dans le domaine de la formation en ligne, des systèmes utilisant des représentations structurées de ressources éducatives ont été proposés [Shabir & Clarke, 2009]. Cependant la richesse sémantique de ce type de représentation est principalement utilisée pour l'intégration de données hétérogènes. Les mesures proposées pour le rapprochement de ressources exploitent à peine la richesse du système.

Dans les chapitres suivants, nous comparons nos propres travaux aux systèmes de rapprochement décrits ci-dessus .

3.5.2. Approches pour la découverte de liens dans le Web de données

Avec l'augmentation de la pratique de publication de données structurées, il est devenu évident que les liens entre les ressources représentent l'atout le plus important de la structure de données. Sous l'hypothèse qu'une source qui publie des données inclut rarement tous les liens possibles entre ses ressources et celles déjà publiées sur le LOD, un certain nombre d'approches a donc été proposé pour essayer de découvrir de nouveaux liens potentiels.

Une branche d'approches plutôt proéminente émerge de la communauté qui s'intéresse aux alignements des ontologies. Cette communauté traite, entre autres, la question de l'équivalence des classes dans des ontologies différentes, ou bien d'autres relations qui peuvent être automatiquement inférées pour faciliter l'intégration des données décrites par celles-ci. Suite à l'émergence du LOD, cette communauté commence à mettre en valeur leur savoir-faire dans le contexte d'alignement des ressources du graphe Web. Si la problématique traitée par ces approches est différente de la nôtre, elle est sans doute liée à l'idée de mise en correspondance des ressources communes.

Un exemple est présenté dans [Tang & Zhao, 2011]. Ce système générique intègre plusieurs approches d'alignement de ressources du LOD, qui sont d'ailleurs reprises des travaux d'alignement des ontologies:

- Techniques d'alignement basées sur la morphologie : ces techniques s'appuient sur des comparaisons simples de similarité des mots décrivant les ressources (au sein des URIs, figurant dans des labels, commentaires associés aux ressources).

- Techniques linguistiques qui s'appuient sur les approches de traitement de langue naturelle pour effectuer des alignements.
- Techniques basées sur le graphe et qui s'appuient sur la similarité des structures dans le graphe pour faire des hypothèses sur les possibles alignements.

Dans [Kopcke et al., 2010] nous trouvons une approche permettant de choisir, à l'aide de l'apprentissage automatique, les meilleures techniques d'alignement et de les combiner pour mieux servir un cas d'usage spécifique. Plus spécifiquement, dans le cadre de traitement de données culturelles issues de sources hétérogènes, nous trouvons une approche d'alignement des instances proposé par Wang et al. [Wang et al., 2012].

Enfin, dans [Scharfe & Saint-Martin, 2010] il est possible de trouver une architecture générale d'alignement de ressources focalisée notamment sur la relation d'équivalence. Une récapitulation des scénarios d'usage des telles approches y est présentée ainsi qu'une revue des systèmes existants.

Une approche très flexible et encore plus générale est mise en place dans le cadre du Silk Framework [Jentzsch et al, 2010]. Ce projet fournit un langage qui permet de spécifier de manière formelle les conditions que les ressources doivent satisfaire pour être éligibles pour l'introduction un lien entre eux. Ce langage est complété par des composants logiciels permettant de parcourir automatiquement le Web de données et d'introduire des nouveaux liens en fonction des règles spécifiées.

Toujours dans le contexte d'alignement d'ontologies, nous trouvons une approche parfaitement transposable aux alignements des instances, qui combine des techniques automatiques avec une vérification humaine faits par des experts de domaine [Hamdi et al, 2010]. Les résultats de leurs vérifications sont gardées et réutilisées pour améliorer des alignements futurs.

Si l'idée d'alignements et de découverte de nouveaux liens est conceptuellement très proche à notre notion de rapprochement, l'objectif concret visé par cette branche d'approches est souvent très différent du nôtre. Leur approche vise à corriger les imperfections du LOD et de rendre la structure plus riche et plus utile dans le cadre d'intégration des données des sources hétérogènes. Ces approches sont donc plutôt complémentaires à celles qui seront présentées dans ce mémoire.

3.5.3. Interdépendance du graphe social et graphe conceptuel

Ce n'est que récemment, et surtout avec l'émergence du Web Social Sémantique, que la communauté de recherche a commencé à s'intéresser à l'analyse commune du graphe social, du graphe conceptuel et à leurs relations réciproques. L'idée de la socialisation centrée sur l'objet que nous avons introduit dans la partie §2.2 nous a sans doute inspiré pour la plupart de ces travaux. Nous rappelons que la socialisation centrée sur l'objet représente toute relation et interaction sociale sur le Web qui ressort de l'interaction avec les mêmes ressources Web (objets). En exemple de cette pratique est l'établissement de lies sociales entre des utilisateurs qui commentent souvent des postes du même blog.

Un certain nombre d'approches traite la question de la coévolution des graphes/profils sociaux et conceptuels. Par exemple, Wang et Groth [Wang & Groth, 2010] observent ce phénomène dans les forums Web et dans les publications scientifiques. Leur analyse, effectuée à l'aide d'un modèle de régression, a permis de conclure sur une forte

influence bidirectionnelle de ces deux graphes au fil du temps. La méthode proposée par ces auteurs comprend trois phases :

1. **Création du graphe social et du graphe conceptuel** à partir de l'information issue d'un réseau social riche en contenu et créé par des utilisateurs (e.g. articles de recherche, tweets.). Dans le cas où les articles de recherche sont utilisés en tant que source, le graphe social peut être généré à partir des co-auteurs d'articles. Ces co-auteurs peuvent ensuite être mis en relation avec des concepts trouvés dans le texte de leurs articles.
2. **Calcul de métriques indicatrices de la structure de graphes.** Divers indicateurs de caractéristiques de graphes, e.g. des indices de centralité, peuvent être calculés. Par exemple le dégré sortant/entrant est une métrique qui représente le nombre de liens qui partent de/sont orientés vers un nœud du graphe. Ce type de métrique permet de caractériser la position d'un nœud (concept ou utilisateur) dans son graphe et de suivre son évolution au fil du temps.
3. **Analyse des séries temporelles**. Des analyses statistiques sont mises en place pour qualifier l'influence d'un graphe à l'autre en fonction du temps et des métriques calculées dans la phase précédente.

Au titre d'exemple, l'étude montre que la connectivité d'un concept dans le graphe de liens concept-utilisateur influence négativement la future connectivité de l'utilisateur qui est associé au concept en question. Autrement dit, apres avoir travaillé sur un sujet très populaire, la collaboration des utilisateurs autour de ce sujet diminue avec le temps. L'étude a été répliquée sur Twitter par Singer et Wagner [Singer & Wagner, 2012]. Ils ont observé une forte influence du graphe social sur le graphe conceptuel. Visiblement, ce sujet vient d'émerger dans un petit nombre d'approches et montre un grand potentiel d'extension dans l'avenir, notamment en termes d'utilisation des deux graphes pour les recommandations. Ces voies ont été jusqu'à présent très peu exploitées.

En ce qui concerne la prédiction de nouveaux liens sociaux dans un réseau, l'usage des profils sociaux et conceptuels a déjà été proposé. Ces approches utilisent majoritairement des mots-clés et ne profitent donc pas d'un vrai graphe conceptuel. Par conséquent, ils se différencient de notre méthode qui consiste à utiliser des concepts possédant des URIs. Un exemple de leurs approches est souligné dans [Brzozovski, 2011]. Les mesures de similarité décrites dans cet ouvrage sont basées sur des profils socials et conceptuels. Ils sont principalement conçus pour étudier la probabilité de connexion de deux utilisateurs. Schifanella et al. [Schifanella et al., 2010] propose un approche de prédictions de liens basé sur l'analyse de similarité de mots-clés dans les profils des utilisateurs. Les auteurs confirment la corrélation entre la similarité de mots-clés et l'existence d'un lien social. Cette approche n'exploite pas non plus la richesse du graphe de concepts.

Les rapprochements utilisant à la fois le graphe social et le graphe conceptuel dans le cadre de l'innovation restent donc un défi à relever dans nos travaux.

3.6. Conclusions

Nous avons introduit dans ce chapitre le cadre conceptuel qui généralise les travaux de thèse. Ce cadre repose sur l'idée d'un rapprochement utile entre deux éléments informationnels présents sur le Web, par rapport à une situation d'usage particulière. La notion de rapprochement permet de comprendre les fonctions standards des applications Web comme les moteurs de recherche et de recommandation. Etant donné que la notion de rapprochement est fortement liée à la structure du Web, elle permet de mettre en valeur des structures riches de données et mieux comprendre leur applicabilité. Par exemple, les données du Web Sémantique Social pourraient présenter des nombreux avantages dans des rapprochements faits par des applications Web. Si dans cette thèse nous nous intéressons principalement aux scénarios liés à l'innovation, il est clair que ce cadre conceptuel permet également de jauger la richesse, de plus en plus élevée, du graphe Web dans plein d'autres cas d'utilisation.

Nous avons également proposé un modèle d'interaction entre les utilisateurs et le Web, qui permet de comprendre la genèse et l'usage de la structure du graphe Web, ainsi que la capacité des applications Web à traiter l'information et la rendre accessible aux utilisateurs.

Bien que le modèle d'interaction a principalement servi à appuyer notre discours sur le traitement du graphe Web, nous espérons qu'il pourra servir d'inspiration pour des travaux réalisés dans le domaine de l'interaction homme-machine.

Bibliographie

[Adomavicius & Tuzhilin, 2005] Adomavicius, G., & Tuzhilin, a. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), 734-749. doi:10.1109/TKDE.2005.99

[Akiyama & Obara, 2010] Akiyama, T., & Obara, K. (2010). Proposal and Evaluation of Serendipitous Recommendation Method Using General Unexpectedness. *Proc. of Workshop on the Practical Use of Recommender Systems, Algorithms and Technologies* (PRSAT 2010). Barcelona. Retrieved March 21, 2011, from http://ir.ii.uam.es/prsat2010/papers/paper1.pdf.

[Alexander, 1979] Alexander, C. (1979). *The Timeless Way of Building*. Oxford University Press; Later printing edition. Retrieved from http://www.amazon.com/Timeless-Way-Building-Christopher-Alexander/dp/0195024028

[Andel & Bourcier, 2009] Andel, P. V., & Bourcier, D. (2009). *De la sérendipité dans la science, la technique, l'art et le droit : Leçons de l'inattendu*. Retrieved from http://www.amazon.fr/sérendipité-dans-science-technique-droit/dp/2355130183

[André & schraefel, 2009] Andre, M., & Schraefel, M. (2009). Designing for (un) serendipity. *The Biochemist*, (December), 19-22. Retrieved from http://www.biochemist.org/bio/03106/0019/031060019.pdf

[André et al., 2009] André, P., Teevan, J., & Dumais, S. T. (2009). Discovery Is Never by Chance : Designing for (Un) Serendipity. In, *ACM Creativity & Cognition* 2009.

[Brin & Page, 1998] Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *In: Seventh International World-Wide Web Conference (WWW 1998)*, April 14-18, 1998, Brisbane, Australia.

[Brzozowski, 2011] Brzozowski, M. (2011). Who Should I Follow? Recommending People in Directed Social Networks. *Proceedings of Internantional AAAI Conference on Weblogs and Social Media*. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2867/3232

[Ge et al., 2010] Ge, M., Delgado-Battenfeld, C., & Jannach, D. (2010). Beyond accuracy: evaluating recommender systems by coverage and serendipity. *Proceedings of the fourth ACM conference on Recommender systems* (p. 257–260). ACM. Retrieved March 22, 2011, from http://portal.acm.org/citation.cfm?id=1864761.

[Hamdi et al., 2010] Hamdi, F., Reynaud, C., & Safar, B. (2010). Pattern-based mapping refinement. *17th International Conference on Knowledge Engineering and Knowledge Management, EKAW* (Vol. 11). Retrieved from http://www.springerlink.com/index/21X0M5352M566733.pdf

[Iaquinta et al., 2008] Iaquinta, L., Gemmis, M. D., Lops, P., Semeraro, G., Filannino, M., & Molino, P. (2008). Introducing Serendipity in a Content-Based Recommender System. *Eighth International Conference on Hybrid Intelligent Systems*, 168-173. Ieee. doi:10.1109/HIS.2008.25

[Jentzsch et al., 2010] Jentzsch, A., Isele, R., & Bizer, C. (2010). Silk – Generating RDF Links while publishing or consuming Linked Data. In P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, et al. (Eds.), *Proceedings of the 9th International Semantic Web Conferene - ISWC 2010* (pp. 1-4). Shanghai, China: Springer Verlag.

[Kamahara et al., 2005] Kamahara, J., Asakawa, T., Shimojo, S., & Miyahara, H. (2005). A Community-based Recommendation System to Reveal Unexpected Interests. *Communities*. IEEE Computer Society. Retrieved March 23, 2011, from http://www.computer.org/portal/web/csdl/doi/10.1109/MMMC.2005.5.

[Kopcke et al. 2010] Kopcke, H., Thor, A., & Rahm, E. (2010). Approaches for Matching Web Data Entities. *Ieee Internet Computing*, 23-31.

[Mann, 2001] Mann, D. (2001) *An Introduction to TRIZ: The Theory of Inventive Problem Solving*. *Creativity and Innovation Management*, 10, 123–25.

[McNee & Riedl, 2006] McNee, S., & Riedl, J. (2006). Being accurate is not enough: how accuracy metrics have hurt recommender systems. *Human factors in computing systems*, 1097-1101. Retrieved March 21, 2011, from http://portal.acm.org/citation.cfm?id=1125659.

[Murakami et al., 2008] Murakami, T., Mori, K., & Orihara, R. (2008). Metrics for evaluating the serendipity of recommendation lists. *New Frontiers in Artificial Intelligence*, 40–46. Springer. Retrieved March 23, 2011, from http://www.springerlink.com/index/5742245730n31454.pdf.

[Passant et al., 2009] Passant, A., Heitmann, B., & Hayes, C. (2009). Using Linked Data to build Recommender Systems. *Proceedings of RecSys*, New York,. Retrieved from http://sw.deri.org/2009/06/recsys/recsys_linkeddata.pdf

[Passant, 2010] Passant, A. (2010). dbrec — Music Recommendations Using DBpedia. In P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, et al. (Eds.), *Proceedings of the 9th International Semantic Web Conferene - ISWC 2010* (Vol. 1380, pp. 1-16). Shanghai, China: Springer Verlag.

[Prié, 2011] Prié, Y. (2011) *Vers une phénoménologie des inscriptions numériques. Dynamique de l'activité et des structures informationnelles dans les systèmes d'interprétation*. HDR. Université Claude Bernard Lyon 1 - Laboratoire LIRIS

[Rao et al., 2010] Rao, N., & others. (2010). Application domain and functional classification of recommender systems—a survey. *DESIDOC Journal of Library & Information Technology,* 28(3), 17–35. Retrieved from http://publications.drdo.gov.in/ojs/index.php/djlit/article/view/174

[Russel & Norvig, 1995] Russel, S., & Norvig, P. (1995). *Artificial Intelligence: A Modern Approach* . Prentice Hall; 1st edition. Retrieved from http://www.amazon.com/Artificial-Intelligence-Approach-Stuart-Russell/dp/0131038052

[Scharfe & Saitn-Martin, 2010] Scharffe, F., & Saint-Martin, M. (2010). Méthodes et outils pour lier le web des données Web des données et interrelations Une approche générale pour relier les données du web. *Actes de RFIA* 2010.

[Schifanella et al., 2010] Schifanella, R., Barrat, A., & Cattuto, C. (2010). Folks in folksonomies: social link prediction from shared metadata. *Proceedings of the third ACM conference on Web Search and Data Mining WSDM2010* (pp. 271-280). New York, New York, USA: ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1718521

[Shabir & Clarke, 2009] Shabir, N., & Clarke, C. (2009). Using Linked Data as a basis for a Learning Resource Recommendation System. Learning in the Synergy of Multiple Disciplines, *Proceedings of the EC-TEL 2009*. Retrieved from http://eprints.soton.ac.uk/268053/

[Singer & Wagner, 2012] Singer, P., & Wagner, C. (2012). Understanding co-evolution of social and content networks on Twitter. *Making Sense of Microposts (#MSM2012) Workshop at World Wide Web Conference WWW2012* (pp. 57-60). Lyon, France. Retrieved from http://ceur-ws.org/Vol-838/paper_10.pdf

[Singer & Wagner, 2012] Singer, P., & Wagner, C. (2012). Understanding co-evolution of social and content networks on Twitter. *Making Sense of Microposts (#MSM2012) Workshop at World Wide Web Conference WWW2012* (pp. 57-60). Lyon, France. Retrieved from http://ceur-ws.org/Vol-838/paper_10.pdf

[Tang & Zhao, 2011] Tang, Y., & Zhao, G. (2011). SDD-Matcher : a Semantic-Driven Data Matching Framework. *Semantic Web Journal*.

[Waitelonis & Sack, 2009] Waitelonis, J., & Sack, H. (2009). Towards Exploratory Video Search Using Linked Data. 2009 *11th IEEE International Symposium on Multimedia*, 540-545. Ieee. doi:10.1109/ISM.2009.111

[Wang & Groth, 2010] Wang, S., & Groth, P. (2010). Measuring the dynamic bi-directional influence between content and social networks. In P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, et al. (Eds.), *Proceedings of the 9th International Semantic Web Conferene - ISWC 2010*. Shanghai, China: Springer Verlag.

[Wang et al., 2012] Shenghui Wang, Antoine Isaac, Stefan Schlobach, Lourens van der Meij, Balthasar A. C. Schopman(2012): Instance-based Semantic Interoperability in the Cultural Heritage. *Semantic Web* 3(1): 45-64

4. Rapprochement des experts et des concepts de leur expertise

Ce chapitre étudie les cas pour lesquels les rapprochements entre utilisateurs et concepts s'appliquent lors d'une recherche d'experts. Nous proposons une approche générale pour la mise en place d'un tel rapprochement (§4.1). Elle dépend en grande partie des critères d'appréciation de l'expertise, qui eux-mêmes dépendent de la situation. L'ensemble des critères utilisés dans une situation particulière est appelé « hypothèse d'expertise » (§4.2). La dernière partie du chapitre est consacrée à une approche (§4.3) et un système (§4.4) capable de suggérer des hypothèses d'expertise appropriées à une situation donnée, permettant ainsi de trouver des experts pertinents.

4.1. Rapprochement des personnes et concepts dans le cadre de la recherche d'experts

4.1.1. Scénario d'usage et les objectifs de rapprochement

L'objectif principal d'un rapprochement au service de la recherche d'experts est de rapprocher dans le graphe Web des experts et des concepts représentant leurs sujets d'expertise. Si ce rapprochement est pertinent dans le cadre de recherche d'experts, il ne faut pas le restreindre à cette utilisation dans la mesure où il peut s'avérer intéressant dans d'autres cas liés à l'innovation. Un service qui recommanderait des ressources

différentes (e.g., offres de travail, concours, appels à projets) à des experts pourrait également bénéficier d'un tel rapprochement. Des services de diffusion de matériaux publicitaires pourraient aussi en profiter. Nos travaux se focalisent sur le scénario principal, mais n'excluent pas l'hypothèse de leur réutilisation dans d'autres scénarios.

La recherche d'experts est souvent effectuée par des plateformes d'innovation sous forme de diffusion de problèmes à un public d'auteurs de solutions potentielles [Speidel, 2011], [Sloane, 2011]. Pour chaque problème, la personne chargée de la communication au sein d'une plateforme donnée se met en quête d'experts sur des sujets choisis. Chez Hypios, les experts ainsi détectés sont contactés de manière très attentive et personnalisée. En effet, les personnes chargées de la communication contactent les experts directement par e-mail, en prenant soin d'ajuster leur discours en fonction du poste, de la formation et de l'expérience du potentiel auteur de solutions, afin de le motiver à participer à la résolution de problèmes.

Si l'envoi automatique d'e-mails d'invitation est envisageable, l'expérience d'Hypios révèle qu'un travail manuel et personnalisé sur la rédaction de ces e-mails récolte souvent plus de retours de la part des experts contactés, qui, convaincus de la réalité de l'engagement d'une personne véritable dans cette communication, sont plus enclins à répondre. Cette intervention humaine dans la communication débouche souvent sur des retours d'experts qui, n'ayant pas de temps pour travailler sur une nouvelle solution, envoient des documents pré-existants : articles de recherche, brevets pertinents pour le problème et parfois même les coordonnées de collègues susceptibles de pouvoir répondre au défi proposé. Tous ces retours constituent un fonds de pistes de résolutions possible qui est d'une valeur considérable pour les clients et les utilisateurs des plateformes d'innovation. Dans le cas où les solutions d'un problème ont été trouvées, ces pistes aident à mieux choisir la bonne solution et à l'implémenter par la suite. L'importance de ces pistes est toutefois plus significative dans le cas de non résolution d'un problème. Dans ce cas, l'existence de pistes de solution permet, par exemple, d'expliquer la non résolution par un prix pas suffisamment attractif pour motiver les auteurs de solutions.

4.1.2. Processus générique de rapprochement

Dans le scénario de la recherche d'experts, le rapprochement se base principalement sur la notion de l'hypothèse d'expertise. L'hypothèse d'expertise est une règle qui définit des conditions nécessaires pour considérer quelqu'un comme expert d'un sujet précis dans une situation spécifique. La condition contient obligatoirement la mention d'**un utilisateur**, **un domaine de compétence** et **des éléments qui font le lien** entre les deux. Par exemple, une hypothèse d'expertise peut être « Chaque personne qui à rédigé au moins 2 articles de recherche et 3 billets de blog sur un sujet, peut être considéré comme expert en ce sujet ». Ces règles s'appliquent surtout sur des traces utilisateur et il faut comprendre le résultat de leur application comme étant dépendant de l'état des traces disponibles.

La littérature existante compte de nombreux exemples de systèmes de recherche d'experts fondés sur des hypothèses d'expertise différentes. Nous présentons un récapitulatif des travaux de cette nature dans la partie §4.2.1.1. Intuitivement, il est clair que certaines hypothèses sont plus recevables que d'autres dans certaines situations. Il est donc capital de bien choisir une hypothèse d'expertise appropriée à une recherche d'experts concrète. Ce processus est résumé en Figure 27.

Figure 27 Le choix et l'application des hypothèses d'expertise dans les plateformes d'innovation sur le Web

Dans la suite de ce chapitre, nous proposons nos solutions pour les différentes étapes de ce processus. Dans la partie §4.2 nous présentons différentes hypothèses d'expertise recueillies dans les approches existantes, décrites dans la littérature. Ces approches utilisent, majoritairement, des sources de données non structurées ou semi structurées, en mettant l'accent sur l'extraction de l'information et sur les différentes tâches de classification des documents par sujets. Nous nous sommes intéressés à l'applicabilité de ces hypothèses d'expertise pour la recherche d'experts directement dans le Graphe Web et plus particulièrement en utilisant le LOD. Les conclusions de cette étude ont servi à construire l'approche des choix d'hypothèses d'expertise présentée en §4.3 ainsi qu'un système de recherche d'experts présenté en §4.4.

4.2. Hypothèses d'expertise et le graphe Web

4.2.1. Hypothèses d'expertise par le type de trace concerné

L'analyse de travaux existants a révélé que l'étude des différents types de traces utilisateur (e.g, blogs, articles scientifiques) était souvent considérée. Par conséquent, les hypothèses utilisées dans ces travaux sont influencées par les types de traces qui étaient disponibles dans leur contexte. Ces types de traces peuvent être classées de la manière suivante :

- **Contenu créé par des utilisateurs**. Un billet de blog est un exemple de type de trace de cette catégorie. Le contenu, au sens de cette catégorie, est toujours la conséquence d'une activité de création du contenu et il est important de le distinguer des autres contenus qui peuvent emerger en conséquence d'autres activités ou même de manière indépendante de l'utilisateur; Un utilisateur qui clique sur un bouton « j'aime » crée un certain trace numérique de cette action, mais la création de celle-ci n'est pas l'objectif principal de l'activité de l'utilisateur comme dans le cas de la création d'un billet de blog ou d'un tweet.
- Traces d'activité des utilisateurs autres que création du contenu. Nous distinguons deux types d'activités qui sont à l'origine de la création de ce type de traces :
  - **Activités avec du contenu Web.** Le fait d'aimer, partager ou répondre à des contenus créés par d'autres utilisateurs en sont des exemples ;
  - **Activités effectuées indépendamment du Web mais qui y sont référencées**. Par exemple l'animation d'une conférence, même si elle ne donne pas lieu à un contenu particulier, peut être référencé dans le Web par le public de cette conférence;
- **Réputation et autorité de l'utilisateur**. Ce type de trace comprend des différentes évidences, témoignages de l'opinion des autres sur l'utilisateur ou bien sur son œuvre.

Des types de traces peuvent appartenir à plusieurs catégories. Par exemple, un commentaire laissé à propos d'un billet de blog et à la fois le résultat d'une activité d'interaction avec du contenu et celle de la création du contenu.

Il est important de noter que notre classification de types de traces présentée ci-dessus n'a pas le caractère exhaustif. Elle vise principalement à faciliter la présentation des nombreuses hypothèses d'expertise par la suite.

Dans la partie 4.2.2 nous présentons quelques exemples de chaque type d'hypothèse ainsi que leur utilisation dans l'état de l'art.

* * *

Le travail effectué dans cette thèse au sujet de la recherche d'experts se distingue des autres, notamment par l'utilisation de données liées. Ce choix possède deux avantages principaux :

- **La séparation des hypothèses d'expertise des données sur les experts**. La plupart des systèmes de recherche d'experts sont construits autour d'une hypothèse d'expertise particulière et le processus d'extraction de données structurées découle de cette hypothèse, rendant ainsi difficile la réutilisation des mêmes données pour illustrer d'autres hypothèses. La séparation permet notamment de varier les hypothèses d'expertise tout en utilisant les mêmes sources de données. Nous verrons par la suite que cette possibilité est très importante pour la construction de systèmes de recherche d'experts couvrant plusieurs champs de compétence et qu'elle pourrait même s'avérer déterminante pour le bon fonctionnement de ces systèmes.
- **L'appui sur une base globale de traces utilisateurs**. S'appuyer sur le graphe Web global offre la possibilité de considérer une grande variété de types d'information sur les experts potentiels, fournies par diverses sources et mises à jour, indépendamment de l'application qui effectue la recherche d'experts. Cette prise en compte de sources variées permet d'évaluer de manière plus complète les qualités d'un expert potentiel, d'analyser la distribution de ses domaines d'intérêts/compétence et d'en effectuer d'autres analyses, dont le résultat n'aurait pas la même validité si une seule source de données avait été sollicitée. Ceci est dû à la multiplicité de canaux de communication, inhérente à l'interaction dans la société actuelle.

Dans la partie §4.2.2 nous présenterons une étude du potentiel du graphe de données liées. Nous verrons plus précisément comment, dans son état actuel, il peut répondre aux besoins informationnels des hypothèses d'expertise que nous avons trouvées dans la littérature. En raison d'une croissance rapide du graphe Web notre analyse pourrait bientôt paraître dépassée par une amélioration de certains aspects du graphe. Nous avons donc essayé de rendre nos conclusions plus générales en discutant la nature des informations exigées par certaines hypothèses, ainsi que d'illustrer les mécanismes par lesquels cette information parvient dans le graphe de données liées publiques en prenant comme exemple le graphe actuel.

Dans la partie §4.2.3 nous discuterons des obstacles techniques provenant des imperfections du graphe de données liées publiques qui sont susceptibles de limiter l'applicabilité de certaines hypothèses. Cette analyse mettra en lumière les bonnes

pratiques de publication de données liées qui assureront leur usage optimal dans le cadre de la recherche d'experts.

4.2.1.1. Quelques exemples d'hypothèses d'expertise et leur utilisation dans la littérature

Dans cette partie nous présentons des hypothèses d'expertise, trouvées pour la plupart dans différents travaux existants sur des systèmes de recherche d'experts. Ces hypothèses utilisent différents types de traces utilisateur pour rapprocher des experts de concepts, sujets de leur compétence.

4.2.1.1.1. Les hypothèses d'expertise liées au contenu créé par les utilisateurs

H1: Si l'utilisateur a écrit un article sur le sujet X, alors il peut être considéré comme expert en sujet X.

Cette hypothèse est très couramment utilisée puisque le fait d'avoir un article accepté constitue une preuve de la connaissance de l'utilisateur en ce domaine. Des formes alternatives de cette hypothèse imposent certaines contraintes par rapport au nombre minimal de publications ainsi qu'au niveau de la conférence ou journal scientifique où l'article a été publié. Ces paramètres sont souvent utilisés pour effectuer le classement d'experts et évaluer leur niveau de compétence. Dans les travaux existant qui utilisent cette hypothèse, il est souvent question d'extraire des concepts – sujets de compétences pertinents pour un article donné. Par exemple, dans [Buitelaar & Eigner, 2008] une approche simple de reconnaissance de patterns lexicaux est employée pour identifier des concepts dans les textes d'articles de recherche et ensuite relier des auteurs de ces articles avec des concepts. Demartini et Niederée [Demartini & Niederée, 2008] s'appuient sur les données sémantiques présents sur des ordinateurs équipés d'un bureau sémantique (en. Semantic Desktop)[68]. Ces auteurs s'intéressent au scénario où l'utilisateur d'un système irait chercher un expert pour répondre à une question liée à un sujet de compétence. Ils se servent ensuite des articles et de tout autre type de traces qu'ils peuvent trouver dans le bureau sémantique de l'utilisateur pour retrouver les personnes potentiellement compétentes. L'approche proposée est limitée aux traces collectées dans cet ordinateur personnel et le résultat de la recherche d'experts ne serait pas le même sur des ordinateurs différents. Si cette caractéristique peut paraître utile en termes de personnalisation de l'approche, elle est également très limitée en termes de découverte d'experts inconnus. De même, l'information prise en compte dans l'évaluation de l'expertise est relativement incomplète.

H2 : Si l'utilisateur à créé du contenu Wikipedia sur le sujet X, alors il peut être considéré comme expert en sujet X.

Avec le temps et les efforts de ses contributeurs pour assurer la qualité du contenu, Wikipedia est devenu une source de connaissances fiable [Giles, 2005]. Le mécanisme de vérification de contenu est aujourd'hui rigoureux ce qui rend difficile l'insertion de fausses informations dans les textes. Il est donc raisonnable de considérer qu'un auteur de contenu Wikipedia retenu est un expert potentiel sur le sujet de son contenu. Une approche qui se sert de l'historique de changements de pages Wikipedia pour retrouver des experts peut être trouvée dans [Demartini, 2007]. Une fois identifiés, les experts

[68] Des tels systèmes d'exploitation représentent des contenus et des informations stockées dans un ordinateur sous des formats riches, notamment en RDF.

sont classés à l'aide d'algorithmes similaires à PageRank [Brin & Page, 1998] et HITS [Kleinberg, 1999] qui aident à déterminer les articles les plus influents et à considérer l'expertise de leurs auteurs comme étant la plus importante. D'autres travaux exploitant l'historique de changements de Wikipedia pour en extraire des relations entre des personnes sont également présentés dans [Orlandi & Passant, 2010] mais ils ne sont pas directement destinés à la recherche d'experts.

H3 : Si l'utilisateur à créé un certain nombre de billets de blog sur le sujet X, alors il peut être considéré comme expert en sujet X.

Avec la prolifération du blogging dans de nombreuses sphères intellectuelles, il est devenu très intéressant d'étudier les blogs influents et de considérer leurs auteurs comme experts sur le sujet de leurs billets. Plusieurs approches de recherche ont été proposées pour exploiter la sémantique des blogs pour la recherche d'experts. Kolari et al. [Kolari et al., 2008], par exemple, utilisent les blogs internes d'une organisation, plus précisément IBM, pour en extraire des sujets de compétence et évaluer l'expertise de leurs auteurs. Leur méthode est facilement généralisable pour tous les blogs dans le Web ouvert. Des approches similaires sont également proposées dans [Chua, 2007] et [Amitay et al., 2008].

H19[69] : Si l'utilisateur à créé un certain nombre de tweets sur le sujet X, alors il peut être considéré comme expert sur un sujet X.

Une version alternative de l'hypothèse H3, utilisant les tweets comme type de trace commence à être de plus en plus utilisée dans la littérature. Dans [Zoltan & Johan 2011], un système d'extraction de concepts d'ontologies à partir de tweets est utilisé pour constituer des profils d'experts. L'usage de modèles de sujets (en. *topic models*) pour la mise ne valeur des tweets dans la recherche d'experts est proposée dans [Wagner, 2010].

H20 : Si un sujet X apparaît de manière récurrente dans les tweets créés par l'utilisateur, alors celui-ci peut être considéré comme expert sur le sujet X.

Le système proposé dans [Ziamatin, 2011] effectue la constitution de profils d'experts en fonction de la persistance d'un concept dans les tweets d'un utilisateur à travers le temps. Les travaux présentés dans [Abel et al., 2011] et [Stan et al., 2011] proposent des approches similaires et confirment empiriquement une performance supérieure des approches basées sur la dynamique des sujets.

4.2.1.1.2. Les hypothèses d'expertise liées aux activités des utilisateurs

Les hypothèses d'expertise liées aux activités d'utilisateurs en ligne

H4 : Si un utilisateur répond à des questions sur un sujet X (posées par des experts en ce sujet), alors il peut être considéré comme expert du sujet X.

Cette hypothèse est souvent utilisée par des approches qui se servent des communautés de questions et réponses en ligne. Une approche de ce type, présentée dans [Adamic et al., 2008], utilise ainsi la communauté Yahoo Answers[70]. Certaines approches visant à

[69] Les hypothèses H19 et H20 ont été introduites après la publication de l'article [Stankovic et al., 2010]. Ce choix de leur numérotation est donc fait pour rester cohérent avec les numéros d'hypothèses déjà publiées.

[70] http://answers.yahoo.com/

mieux classifier des experts tiennent notamment compte du niveau d'expertise de l'utilisateur qui à posé la question pour évaluer l'expertise de celui qui y a répondu. Jurezyk et Agichten [Jurezyk & Agichten, 2007] construisent un graphe d'interaction des utilisateurs dans la communauté. Pour chaque réponse ils créent un lien orienté entre l'utilisateur qui à posé la question et celui qui à propose une réponse. Ce graphe est ensuite traité par des algorithmes similaires à PageRank et HITS pour effectuer la propagation de l'expertise dans le graphe et en tirer de meilleurs experts. Une approche semblable est reprise dans [Zhang et al., 2007] sur l'exemple du Java support forum.

Certaines communautés de pratique, telle que la communauté des informaticiens développeurs, s'appuient fortement sur se type de forums pour échanger leurs connaissances professionnelles. Cette hypothèse est notamment pertinente pour la recherche d'experts parmi des développeurs (ou professionnels de métiers similaires) qui seront difficiles à trouver parmi des auteurs de publications scientifiques ou d'autre contenus.

H5 : Si un utilisateur est le premier à découvrir (et partager) des ressources importantes sur le sujet X, alors il peut être considéré comme expert en sujet X.

Noll et al. [Noll et al., 2009] se servent de marque-pages pour identifier des experts. Ces auteurs considèrent que la capacité de trouver des informations sur un sujet qui captent l'attention d'autres utilisateurs relève d'une certaine expertise. Leur approche, implémentée sur le site de filtrage collaboratif del.icio.us[71] est notamment focalisée sur le temps de marquage qui permet de tracer les experts qui ont été les premiers à identifier une ressource.

H6 : Si un utilisateur participe dans un projet de développement du logiciel collaboratif utilisant un langage de programmation, alors il peut être considéré comme expert en ce langage de programmation.

Bien que nous n'ayons pas trouvé d'exemples d'approches utilisant cette hypothèse dans la littérature scientifique, d'une pratique de plus en plus importante du développement collaboratif en ligne au sein de sites de partage du code tels que GitHub[72], Google Code[73] et SourceForge[74], il est clair qu'une telle hypothèse est très pertinente. Le niveau de participation de l'utilisateur dans le projet ainsi que son rôle peuvent être utilisés pour classer les experts.

Les hypothèses d'expertise liées aux activités et accomplissements d'utilisateurs hors ligne

H7 : Si le CV d'un utilisateur comprend des capacités liées au sujet X, alors il peut être considéré comme expert en sujet X.

Cette hypothèse est souvent utilisée dans la recherche d'experts dans le cadre de la gestion des ressources humaines. Le réseau social LinkedIn par exemple fournit une interface de recherche d'experts basée sur cette hypothèse[75].

[71] http://delicious.com/

[72] https://github.com/

[73] http://code.google.com/

[74] http://sourceforge.net/

[75] http://blog.linkedin.com/2011/02/03/linkedin-skills/

H8 : Si l'utilisateur a obtenu des bourses de recherche pour des projets liés au sujet X, alors il peut être considéré comme expert en sujet X.

Un exemple d'utilisation de cette hypothèse peut être trouvé dans le système SAGE (Searchable Answer Generated Environment) Expert Finder [Becara-Fernandez, 2006]. Même si cette hypothèse est sur le principe très logique, il se pose la question de déterminer le rôle de l'utilisateur dans le projet. Sur des projets multidisciplinaires qui comprennent la participation de plusieurs experts, et parfois même plusieurs institutions il est difficile d'évaluer l'expertise d'une personne sur l'ensemble des sujets liés au projet.

Par la suite nous présenterons quelques hypothèses H9-H16 que nous n'avons pas trouvées dans la littérature, mais qui s'imposent en conséquence de la compréhension de la nature des traces créées naturellement lors des activités des utilisateurs experts en ligne.

H9 : Si un utilisateur détient une certaine position dans une entreprise et que son travail est lié au sujet X, alors il peut être considéré comme expert sur le sujet X.

H10 : Si un utilisateur supervise quelqu'un travaillant sur le sujet X ou bien enseigne le sujet X, alors il peut être considéré comme expert sur le sujet X.

H11 : Si un utilisateur a plusieurs années d'expérience de travail sur le sujet X, alors il peut être considéré comme expert sur le sujet X.

H12 : Si un utilisateur est membre de comité de programme d'un évènement professionnel ou académique dédié au sujet X, alors il peut être considéré comme expert sur le sujet X.

H13 : Si un utilisateur est invité pour donner une conférence au sujet X lors d'un congrès, alors il peut être considéré comme expert sur le sujet X.

H14 : Si un utilisateur préside une session sur le sujet X lors d'un congrès dédié au sujet Y, alors il peut être considéré comme expert sur le sujet X. Par extension, il peut également être considéré comme expert sur le sujet du congrès Y.

H15 : Si un utilisateur donne une conférence sur le sujet X lors d'un congrès dédié au sujet Y, alors il peut être considéré comme expert sur le sujet X. Par extension, il peut également être considéré comme expert sur le sujet du congrès Y.

H16 : Si un utilisateur a été invité en tant qu'expert dans une émission télévisée (transmise sur le Web) sur le sujet X, alors il peut être considéré comme expert sur le sujet X.

4.2.1.1.3. Les hypothèses d'expertise liées à la réputation des utilisateurs

Dans la littérature il est possible de trouver certaines hypothèses d'expertise basées sur une vision limitative de la réputation, notamment fondé sur la popularité des traces ou bien sur la connectivité au sein d'un réseau social.

H17 : Si l'utilisateur rédige un billet de blog sur le sujet X qui attire un grand nombre de commentaires (atteigne une popularité), alors il peut être considéré comme expert sur le sujet X.

C'est l'approche proposée par Kolari et al. [Kolari et al., 2008] que nous avons déjà mentionnée, en lien avec l'hypothèse H3, qui se base sur la capacité de l'utilisateur

d'attirer de l'attention à ses contenus. Une telle capacité peut être comprise comme signe d'une certaine notoriété dans une communauté.

H18 : Si un utilisateur est fortement connecté avec un expert en sujet X, alors il peut lui-même être considéré comme expert en sujet X.

Cette hypothèse est utilisée pour propager l'expertise dans un réseau de connexions sociales à condition qu'un ensemble d'experts dans le réseau soit initialement identifié. L'approche proposée en [Fu et al., 2007] se sert d'un graphe social créé à partir d'échanges d'e-mails et de collaborations sur des documents pour appliquer cette hypothèse à la recherche d'experts.

4.2.2. Richesse du graphe Web en traces de différents types

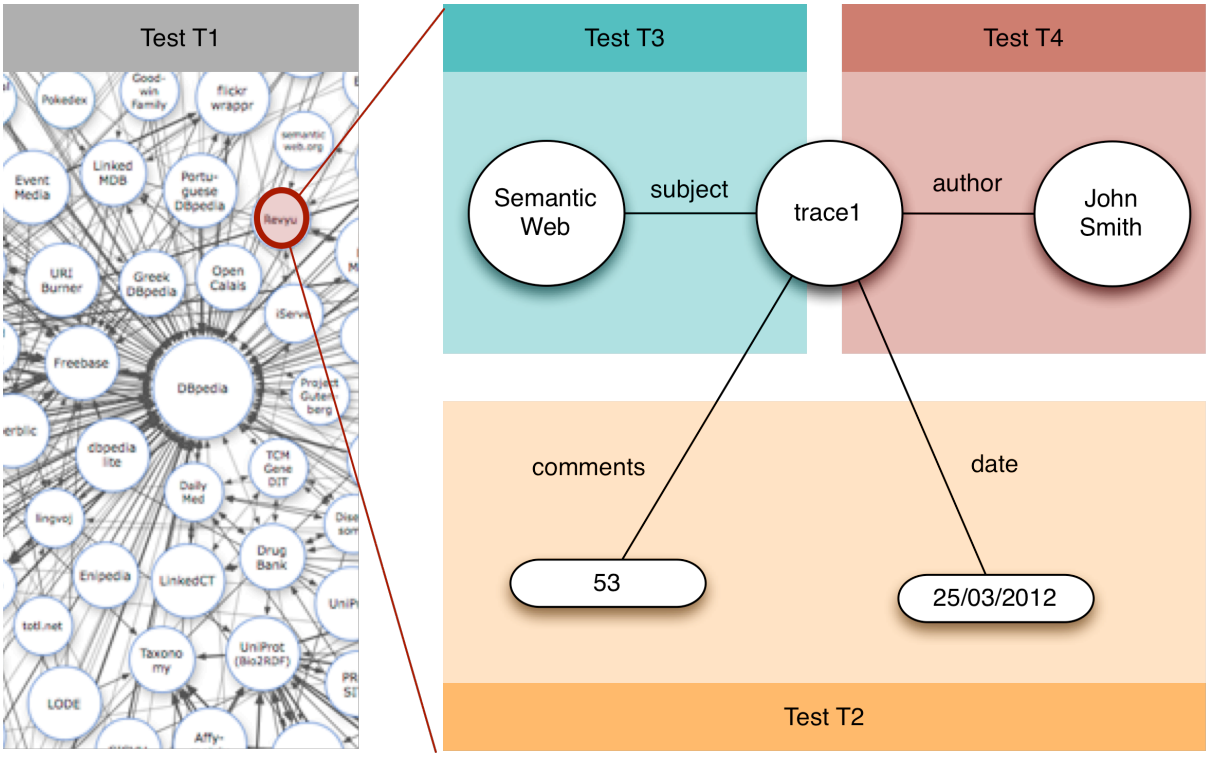


Figure 28 Représentation graphique de tests

Afin d'évaluer la capacité de graphe du LOD à répondre aux besoins des hypothèses d'expertise que nous avons isolées, nous avons effectué une étude des données disponibles dans celui-ci. Nous avons défini 4 tests permettant d'évaluer des aspects particuliers. Nous donnons ci-dessous les définitions plus précises de nos tests. Pour mieux expliquer leur nature à l'aide d'un exemple nous allons nous servir de la Figure 28 qui illustre d'une manière visuelle, les quatre tests, sur l'exemple de l'hypothèse H5. Nous rappelons que cette hypothèse concerne la capacité de l'utilisateur d'être parmi les premiers à découvrir des ressources Web pertinentes par rapport à un sujet donné.

T1 : Pour une hypothèse d'expertise donnée, est-il possible de trouver des ensembles de données structurées (EDS) dans le LOD qui soient utiles pour permettre son utilisation?

Ce test consiste à vérifier si des EDS contenant le type de trace utilisateur qui est employé par l'hypothèse d'expertise donnée existe dans LOD. Par exemple, pour une hypothèse qui utilise les articles scientifiques pour inférer l'expertise, LOD passerait le

test s'il contient au moins un EDS avec des informations sur des articles de recherche. En absence de tel EDS nous pourrons conclure que LOD ne permet pas l'application de l'hypothèse en question.

Prenons l'exemple de l'hypothèse H5. LOD passe ce test grâce à Faviki[76] qui publie des marques pages des utilisateurs sous forme de pages RDFa, en utilisant l'ontologie CommonTag[77].

Tous les autres tests que nous présentons ci-dessous dépendent du T1 i.e. une note négative sur ce test rends inutile l'examin de l'état du LOD sur les autres 3 tests.

T2 : Pour une hypothèse d'expertise donnée, les EDS utiles sont-elles assez riches en information pour rendre possible l'utilisation de l'hypothèse ?

Ce test consiste à vérifier si les EDS contenant des types de traces utiles pour une hypothèse contiennent assez de détails pour pouvoir servir à l'exécution de l'hypothèse. Par exemple, si une hypothèse utilise le temps d'enregistrement d'un marque-pages, LOD passerait le test seulement si le EDS contenant des marque-pages contient de l'information sur les dates de leur enregistrement. Dans le cas de notre hypothèse H5, LOD ne passe pas ce test, car dans Faviki, l'EDS concerné par l'hypothèse, l'information sur la date de publication des marques-pages ne figure pas.

T3 : Pour une hypothèse d'expertise donnée, les EDS utiles contiennent-ils des liens vers des concepts permettant d'identifier de manière non ambigüe des sujets de compétence ?

Le troisième test consiste à vérifier si les traces utilisateur disponibles dans des EDS utiles pour l'hypothèse sont liées à des concepts. Ce test permet de vérifier la facilité avec laquelle LOD peut directement être utilisé pour trouver des experts avec l'hypothèse donnée.

Dans le cas de notre hypothèse H5 que nous prenons en exemple, LOD passe ce test, car Faviki contient les liens vers des concepts DBPedia, utilisés au sein de cette source pour désambigüiser le sens de marques-pages.

Si ce test n'est pas passé, dans certains cas il reste possible d'améliorer le LOD pour compenser le défaut de concepts. Nous verrons par la suite des méthodes concrètes pour le faire.

T4 : Pour une hypothèse d'expertise donnée, les EDS utiles contiennent-ils des liens vers des ressources permettant d'identifier de manière non ambigüe les utilisateurs concernés par des traces?

Ce test permet de s'assurer que les traces nécessaires à l'exécution d'une hypothèse d'expertise contiennent un lien vers un utilisateur ou une autre entité potentielle porteuse de compétence. Ces liens pourront également être utiles dans l'intégration de traces liées à un même utilisateur à travers plusieurs sources.

Dans le cas de notre hypothèse H5, LOD ne passe pas ce test, car Faviki ne contient pas d'information sur des personnes, auteurs des marques-pages.

Nous avons effectué ces tests de manière manuelle. Pour le test T1 il était suffisant de regarder les descriptions d'EDS sur leurs pages Web et parfois voir un exemple de

[76] http://www.faviki.com/pages/welcome/

[77] http://www.commontag.org/Home

ressource faisant partie de cet EDS pour déterminer la nature des informations qui y sont présentées. Si lors d'une telle étude nous avons trouvé un EDS contenant les traces requises pour une hypothèse nous avons marqué « + » dans le tableau de résultats de tests (Tableau 1). Sinon nous avons marqué « - ».

Les vérifications nécessaires pour valider le test T2 ont demandé plus d'effort. Pour s'assurer que toutes les précisions nécessaires pour l'exécution des différentes hypothèses soient présentes dans les EDS ouvertes, nous avons dû vérifier le contenu de ressources particulièrement pertinentes pour chaque hypothèse dans les différents jeux. Le moteur de recherche sémantique Sindice[78] s'est avéré particulièrement utile dans cette tâche, en permettant de chercher les occurrences d'usage de certaines propriétés, porteuses de l'information sollicitée. Ce moteur de recherche a été d'une utilité cruciale dans des cas d'EDS qui n'avaient pas de point d'accès SPARQL tels que FOAF[79] et SIOC[80]. Dans les cas où cela était possible, nous avons également effectué des requêtes SPARQL DESCRIBE sur des EDS pour retrouver tous les triplets dans lesquels participe une ressource et déterminer la présence ou absence de l'information recherchée par l'hypothèse et d'autres requêtes permettant d'étudier l'usage de certaines propriétés et classes dans le EDS en question. Uniquement en cas de non aboutissement de tous ces efforts pour trouver les détails nécessaires, nous avons marqué « - » dans le Tableau 1. L'approche similaire à l'aide de Sindice et de requêtes SPARQL a été utilisée pour effectuer les tests T3 et T4.

Dans certains cas, notamment en présence de plusieurs EDS pertinentes pour des hypothèses de richesse variée, nous avons attribué une note neutre « +- ». Cette note indique que la richesse de certaines sources pertinentes rend l'exécution de l'hypothèse possible, mais que certaines des sources pertinentes pourront toujours bénéficier d'une mise au point. Cette note est également donnée dans des cas où un EDS possède des liens sollicités seulement pour une partie des ressources pertinentes.

Tableau 1 Résultats de tests T1-T4 pour les hypothèses d'expertise H1-H18

Type de trace concerné	Hypothèse d'expertise	T 1	T2	T3	T4
Contenu créé par l'utilisateur	H1: Si l'utilisateur a écrit un article sur le sujet X, alors il peut être considéré comme expert sur le sujet X.	+	+	+-	+
	H2 : Si l'utilisateur à créé du contenu Wikipedia sur le sujet X, alors il peut être considéré comme expert sur le sujet X.	+	+	+	-
	H3 : Si l'utilisateur à créé un certain nombre de billets de blog sur le sujet X, alors il peut être considéré comme expert sur le sujet X.	+	+	+-	+-

[78] http://sindice.com/

[79] http://xmlns.com/foaf/spec/

[80] http://sioc-project.org/ontology/

Type de trace concerné	Hypothèse d'expertise	T 1	T2	T3	T4
	H19[81] : Si l'utilisateur à créé un nombre de tweets sur le sujet X, alors il peut être considéré comme expert sur le sujet X.	+	+	+-	+
	H20 : Si un sujet X récure d'une manière persistante dans les tweets crées par l'utilisateur, alors celui-ci peut être considéré comme expert sur le sujet X	+	+	+	+
Activités en ligne	H4 : Si un utilisateur répond à des questions sur un sujet X (posées par des experts en ce sujet), alors il peut être considéré comme expert sur le sujet X.	+	-	-	-
	H5 : Si un utilisateur est le premier à découvrir (et partager) des ressources importantes sur le sujet X, alors il peut être considéré comme expert sur le sujet X.	+	-	+	-
	H6 : Si un utilisateur participe dans un projet de développement du logiciel collaboratif utilisant un langage de programmation, alors il peut être considéré comme expert en ce langage de programmation.	+	+	+-	+-
Activités hors ligne	H7 : Si le CV d'un utilisateur comprend des capacités liées au sujet X, alors il peut être considéré comme expert sur le sujet X.	-	-	-	-
	H8 : Si l'utilisateur a obtenu des bourses de recherche pour des projets liés au sujet X, alors il peut être considéré comme expert sur le sujet X.	+	+	-	+
	H9 : Si un utilisateur détient une certaine position dans une entreprise et que son travail est lié au sujet X, alors il peut être considéré comme expert sur le sujet X.	+	-	-	+-

[81] Si Twitter a recamment interdit la republication de ses tweets, ce qui a fait disparaître un certain nombre de sources de tweet en LOD, ce type de trace publiée ailleurs (sur Identi.ca et SMOB) reste disponible en RDF.

Type de trace concerné	Hypothèse d'expertise	T 1	T2	T3	T4
	H10 : Si un utilisateur supervise quelqu'un travaillant sur le sujet X ou bien enseigne le sujet X, alors il peut être considéré comme expert sur le sujet X.	-	-	-	-
	H11 : Si un utilisateur a plusieurs années d'expérience de travail sur le sujet X, alors il peut être considéré comme expert sur le sujet X.	-	-	-	-
	H12 : Si un utilisateur est membre de comité de programme d'un évènement professionnel ou académique dédié au sujet X, alors il peut être considéré comme expert sur le sujet X.	+	+	-	+
	H13 : Si un utilisateur est invité pour donner une conférence au sujet X lors d'un congrès, alors il peut être considéré comme expert sur le sujet X.	+	+	-	+
	H14 : Si un utilisateur préside une session sur le sujet X lors d'un congrès dédié au sujet Y, alors il peut être considéré comme expert en sujet X. Par extension, il peut également être considéré comme expert en sujet du congrès Y.	+	+	-	+
	H15 : Si un utilisateur donne une conférence sur le sujet X lors d'un congrès dédié au sujet Y, alors il peut être considéré comme expert sur le sujet X. Par extension, il peut également être considéré comme expert sur le sujet du congrès Y.	+	+	-	+
	H16 : Si un utilisateur a été invité en tant qu'expert dans une émission télévisée (transmise sur le Web) sur le sujet X, alors il peut être considéré comme expert sur le sujet X.	-	-	-	-
Réputation et autorité	H17 : Si l'utilisateur rédige un billet de blog sur le sujet X qui attire un grand nombre de commentaires, alors il peut être considéré comme expert sur le sujet X.	+	+	+-	+-

Type de trace concerné	Hypothèse d'expertise	T 1	T2	T3	T4
	H18 : Si un utilisateur est fortement connecté avec un expert en sujet X, alors il peut lui-même être considéré comme expert sur le sujet X.	+	+	+-	+-

Un résumé des résultats de nos tests est présenté dans le Tableau 1. Un rapport plus détaillé sur les résultats de notre évaluations sur chaque test et chaque hypothèse est publié dans [Stankovic et al., 2010]. Ce niveau de détail, dépasse le cadre de ce chapitre.

4.2.3. Problèmes techniques actuels et remèdes possibles

Si les résultats de notre étude montrent des capacités intéressantes du LOD, certains obstacles limitent, à l'heure actuelle son applicabilité dans le cas de certaines hypothèses. La plupart de ces limitations sont de nature technique et ne proviennent pas de défaut intrinsèque à la représentation de données sur LOD. Il est donc intéressant d'étudier leur nature afin de pouvoir remédier à des imperfections du LOD et développer des stratégies pour dépasser des limitations actuelles. Une analyse approfondie de ces obstacles nous a permis d'isoler 4 causes de non-applicabilité des hypothèses pour la recherche d'experts sur le LOD :

- Données à l'accessibilité réduite ;
- Manque de données ;
- Manque de détails dans les données ;
- Manque de liens entre les données.

Nous détaillons la nature de ces causes ainsi que quelques remèdes possibles dans les parties suivantes.

4.2.3.1. Données à l'accessibilité réduite ;

Certaines informations sont par nature confidentielles. C'est le cas d'e-mails et d'autres contenus échangés au sein des intranets des entreprises. Ces données manquent souvent de liens vers d'autres données. Par leur nature, elles sont rarement exposées en formats structurés. Dans l'état actuel il est donc difficile d'appliquer des hypothèses qui dépendent de ce type de données (e.g., blogs et wikis dans des intranets). Pourtant cette difficulté n'est pas inhérente aux données liées car celles-ci ne sont pas nécessairement publiques. Le principe de données liées n'empêche pas leur mise à disposition dans le cadre d'un accès contrôlé.

Des mécanismes de contrôle d'accès sont nécessaires pour permettre la publication de données liées contenant des informations confidentielles, qui pourront s'avérer utiles même en dehors d'un scénario de recherche d'experts, dans le cadre de différents logiciels d'analyses et d'aides mémoire. Quelques mécanismes existants tels que OAuth[82]

[82] http://oauth.net/

et Foaf+SSL[Story et al., 2009] pourront aider à affronter ces difficultés dans l'avenir. Un manque de motivation pour publier les informations confidentielles en données liées ainsi qu'un manque de compréhension de la distinction entre les données liées et les données liées publiques (souvent par erreur considérées les mêmes [Pellegrini, 2009]) s'ajoutent à la liste d'obstacles. Toutefois, les études qui sont menées sur les bénéfices de la publication des données liées, ainsi que la recherche de nouveaux modèles commerciaux [Brinker, 2010][Dodds, 2010] pour les données liées nous permettront peut être un jour de rendre utiles ces données liées privées.

4.2.3.2. Manque de données

Certains types de traces utilisateur sont simplement absents du LOD actuel. Dans nos analyses, nous avons identifié quelques types de trace qui pourront être produites par la sémantisation des sources de données existantes sur le Web.

Les sites de questions et réponses sont une source très utile de données sur l'expertise, et pourtant leurs données ne figurent pas dans LOD. Avec l'introduction d'exportations de données en utilisant l'ontologie SIOC par exemple, il est relativement facile d'imaginer l'inclusion de ces données et par conséquent une applicabilité de l'hypothèse H4 sur le LOD.

Des données sur le parcours professionnel de personnes, nécessaires pour l'application des hypothèses H7, H10 et H11 seront également utiles dans le format RDF. Dans l'optique d'une ouverture des données publiques initiée par certains gouvernements (e.g. EU, RU, France), il est raisonnable de s'attendre à une ouverture de données concernant les diplômés des écoles et des universités, élément qui contribuera à la recherche d'experts.

Des émissions télévisées transmises sur le Web dans lesquelles participent des experts, ainsi que des cours vidéo en ligne[83] pourront également être annotés avec des métadonnées en RDF, ce qui rendra certaines hypothèses (e.g., H16) applicables.

Des données concernant des évènements professionnels ne sont actuellement publiées dans le LOD que pour des évènements liés au Web Sémantique[84]. Cette pratique pourrait servir d'exemple à d'autres évènements.

Certains travaux sur la représentation de connaissances concernent les traces que nous venons d'évoquer. L'initiative ExpertFinder[85] et le travail présenté dans [Aleman-Meza et al., 2007] listent des ontologies existantes permettant de décrire des données liées à l'expertise. Bien évidemment l'ontologie FOAF, permettant de décrire les informations personnelles de base, que nous avons déjà mentionné fait partie de vocabulaires considérés. Elle est complémentée par l'ontologie DOAP (en Description of a Project) qui permet de décrire la participation des internautes aux projets thématiques. Au sein d'ExpertFinder un vocabulaire qui élargit FOAF est présenté pour permettre la descriptions des institutions et autres entités utiles pour les cas liés à l'expertise au delà des personnes.

[83] http://videolectures.net/

[84] http://data.semanticweb.org/

[85] http://expertfinder.info/

4.2.3.3. Manque de détails dans les données

Un bon exemple du manque de détails est le site Faviki qui publie de l'information sur les tags des utilisateurs. Même si les tags sont associés avec des concepts DBPedia et apportent des indications claires sur l'utilisateur, la date de leur enregistrement n'est pas mise à disposition, ce qui est très contraignant pour l'hypothèse H5. Il est clair donc, que malgré les bonnes intentions des fournisseurs de données, certains cas d'usage restent imprévus. Il serait donc utile de travailler sur les politiques de bonnes pratiques de publication de données pour rendre leur exploitation plus facile et plus diversifiée. Une initiative, Pedantic Web[86] dont les membres visent à corriger les imperfections de Web de données, constitue certainement un pas dans cette direction.

4.2.3.4. Manque de liens entre les données

Le manque de lien est certainement le type d'obstacle le plus fréquent. Les liens connectant une trace à son sujet de compétence font défaut dans la plupart des EDS pertinentes pour H1. Les EDS contenant des informations sur les projets de recherche collaboratifs et bourses scientifiques, potentiellement utiles pour l'hypothèse H8, pâtissent du même manque. Il existe en effet des EDS, tels que CORDIS, qui utilisent une taxonomie conceptuelle qui leur est propre et qui ne réfère pas aux sources de concepts généralement utilisées, comme DBPedia ou Freebase.

Un autre exemple est le EDS Semantic Web Dog Food[87], qui contient des données sur des conférences et des articles dans le domaine du Web Sémantique. Les profils des utilisateurs /experts potentiels sont référencés dans le format FOAF, mais pas les concepts.

Les données extraites des blogs, publiées à l'aide de l'ontologie SIOC, souvent par une extension pour Wordpress[88] qui effectue l'export en RDF de manière automatique, sont rarement liées à des concepts. Dans nos études, nous avons même été confrontés au cas de ressources de type « billet de blog » privées de toute indication de leur créateur.

Si le manque de liens vers des concepts peut être surmonté par des techniques d'extraction de concepts à partir du texte (ces techniques sont abordées dans la partie §2.4.1.1), les liens vers des ressources représentant les utilisateurs sont plus difficiles à établir après la publication de données. Repérer qu'une ressource dans un EDS fait référence au même utilisateur ou objet physique est un défi de recherche déjà abordé par la communauté scientifique.

Différentes méthodes de comparaison des titres de ressources ainsi que d'autres méthodes probabilistes et statistiques pour la mise en correspondance ont été proposées dans [Tang et al., 2007]. Ce défi est encore amplifié quand les données sont représentées par des ontologies différentes mais comparables [Nikolov et al., 2009] car cela impose souvent la nécessité de comprendre le rapport entre les éléments des différentes ontologies en terme de leur égalité et d'autres relations. Dans ce cas, les travaux existants sur l'alignement d'ontologies (en. ontology matching) pourront s'appliquer. Deux types d'approches sont dominantes dans les méthodes d'alignement : (1) les alignements basés sur des instances, et (2) les alignements basés sur la structure

[86] http://pedantic-web.org/

[87] http://data.semanticweb.org/

[88] http://sioc-project.org/wordpress/

de l'ontologie. Le premier type d'alignement concerne notamment les méthodes de comparaisons de similarité textuelle et linguistique de noms de classes, alors que le deuxième comprends des méthodes d'analyse de la structure du graphe et la recherche d'invariances dans cette structure. Une récapitulation très complète de différentes méthodes est donnée dans le livre « Ontology Matching » par Jérôme Euzenat et Pavel Shvaiko [Euzenat & Schvaiko, 2007].

Une approche alternative, Silk Framework [Jentzsch et al., 2010], s'appuie sur un langage déclaratif qui définit des conditions permettant de considérer deux ressources comme étant liées. Grâce à toutes ces approches nous pouvons considérer le manque de liens comme étant une carence provisoire, dû à une étape de transition vers un Web de données mieux connectées.

4.2.3.5. Bonne pratique de la publication de données liées utiles pour la recherche d'experts

Notre analyse de la capacité de LOD à servir à la recherche d'experts aboutit à l'identification de 5 principes généraux sur la bonne pratique de publication de traces utilisateurs en tant que données liées :

- Publier des traces pertinentes pour l'expertise des utilisateurs, avec tous les détails permettant d'identifier les experts et si possiblede les classer par niveau d'expertise.
- Apporter un moyen d'identifier l'utilisateur concerné par une trace utilisateur de manière non ambigüe et mettre en correspondance les ressources désignant cet utilisateur à travers plusieurs sources de données.
- Apporter un moyen de repérer des concepts identifiant des sujets de compétence ayant un lien avec une trace utilisateur. S'appuyer sur des concepts fréquemment utilisés dont le sens est partagé et compris parmi une communauté de pratique.
- Apporter un moyen d'intégrer les informations concernant une trace utilisateur à travers plusieurs EDS qui peuvent en contenir.
- Contrôler l'accès aux données liées et fournir un accès protégé pour les données sensibles et confidentielles.

Nous espérons qu'avec une adhésion plus forte à ces principes ainsi qu'avec les méthodes identifiées dans cette partie pour surmonter des imperfections dans le LOD, ce dernier poursuivra son évolution vers un ressource de plus en plus adaptée pour la recherche d'experts permettant à chaque utilisateur de mettre en valeur ses compétences et ses qualités.

4.3. Choix d'hypothèse d'expertise

Face à la pluralité des hypothèses d'expertise utilisables dans une démarche de recherche d'experts, la question de choix d'hypothèse s'impose à l'utilisateur. Ce choix peut s'avérer difficile, surtout pour un utilisateur qui n'est pas lui-même expert en sujet de la recherche d'experts. Dans cette partie nous proposons une approche pour suggérer des hypothèses d'expertise les plus susceptibles de donner des résultats satisfaisants.

4.3.1. Principe général

Notre approche concernant la recommandation d'hypothèses d'expertise est fondée sur l'observation suivante : différentes communautés de pratique, partageant des connaissances sur différents sujets, utilisent différents canaux de communication pour leurs échanges. Cette hypothèse est principalement fondée sur nos propres observations de la nature des échanges dans la communauté de recherche travaillant sur le Web Sémantique et dans d'autres communautés. On peut observer une différence aussi fondamentale entre la littérature étudiante et la communication scientifique. Les études focalisées sur les échanges scientifiques liés aux sujets proches de la biologie s'intéressent à la création de blogs ou de contenus encyclopédiques en ligne comme les principaux modes d'usage du Web Social pour la dissémination des messages scientifiques dans cette communauté [CCNESVS,2010] ; à l'inverse, les études focalisées sur la communauté du Web Sémantique présentent des analyses d'usage de Twitter en tant qu'outil puissant dans la dissémination des messages scientifiques [Letierce et al., 2010]. En partant de cette hypothèse d'une utilisation différente des canaux de communication selon les communautés, nous nous proposons d'effectuer un choix de type de trace le plus approprié pour un sujet d'expertise. Pour faire ce choix, nous définissons des métriques qui s'appliquent aux EDS liées afin de détecter les types de traces dominantes pour des sujets particuliers. Le type de trace le plus approprié permet de limiter le choix d'hypothèse d'expertise à quelques hypothèses basées sur ce type de trace parmi lesquelles l'utilisateur pourrait lui-même faire un tri final. Notre approche est notamment motivée par les objectifs suivants :

- Profiter du nombre croissant de sources de traces utilisateur publiées en tant que données liées pour construire un système qui pourrait facilement intégrer de nouvelles sources émergeant de traces utilisateur.
- Fonder des recommandations d'hypothèse d'expertise sur les métadonnées, au lieu de se baser sur une simple analyse de la pratique actuelle de communication scientifique afin de construire un système évolutif. Un système flexible et facilement adaptable aux potentiels changements de pratique de communication dans des communautés de pratique qui pourrait même, potentiellement s'adapter à l'émergence de nouveaux canaux de communication.
- Prévoir la nécessité de prendre des décisions concernant le choix d'hypothèse d'expertise dans les situations où la source de données est encore inconnue. Il est nécessaire donc de concevoir des métriques capables de prendre en compte la composition du LOD entier ou de travailler avec d'autres estimations de la pratique de communication globale. Si l'on recommande de cette manière les types de traces, il devrait être possible d'effectuer aussi un choix de source de données qui pourrait correspondre au mieux à une hypothèse d'expertise choisie.

Par la suite, en §4.3.2 nous présenterons un nombre de métriques que nous avons conçues pour sonder la composition de données disponibles et choisir un type de trace appropriée. En §4.3.3, nous présenterons nos expériences conduites afin de vérifier si nos métriques sont bien utiles pour indiquer les types de traces opportuns. Finalement, en §4.3.4 nous discuterons des différents contextes d'application de métriques i.e. des différentes situations dans lesquelles elles peuvent être calculées (sur le LOD entier, sur des sources particulières, sur un ensemble de sources) ainsi que des conclusions concrètes que l'on peut tirer grâce à nos métriques dans chacune de ces situations.

4.3.2. Conception de métriques

Les différentes métriques que nous présentons dans cette partie permettent de quantifier l'intensité de la relation entre un type de trace et un concept. Si nous avons initialement défini un grand nombre de métriques potentielles, présentées dans [Stankovic, 2010], dans la pratique leur calcul s'est montré trop lourd et peu adapté à des applications concrètes. Nous avons donc décidé de ne retenir que des métriques pouvant être calculées à partir des descriptions statistiques des EDS sans devoir effectuer des requêtes SPARQL pour obtenir des données supplémentaires. Si nous appuyons nos travaux sur des descriptions statistiques et pragmatiques, c'est en partie dû à la mise à disposition de descriptions d'EDS à l'aide des ontologies SCOVO[89] et VoID[90], une pratique de plus en plus courante dans la communauté de fournisseurs de données liées. VoID est un vocabulaire conçu pour la représentation des EDS liées, fournissant notamment des éléments pour décrire le type de contenu présent dans un EDS, les ontologies utilisées et des liens vers d'autres EDS. SCOVO complémente l'expressivité du VoID en terme des données statistiques relatives à la composition d'un EDS. Nous donnons un exemple de description d'un EDS, fait en utilisant ces deux ontologies, sur la Figure 29.

```
@prefix scovo: <http://purl.org/NET/scovo#>.
@prefix scovot2: <http://milstan.net/ontologies/scovot2#>.
@prefix void: <http://rdfs.org/ns/void#>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix dcterms: <http://purl.org/dc/terms/>.
<http://example.com/testDataset1> a void:Dataset ;
        dcterms:comment "Milstan's test dataset";
        dcterms:creator <http://example.org/creator#me>;
        dcterms:license <http://creativecommons.org/licenses/by/3.0/>;
        void:sparqlEndpoint <http://api.talis.com/stores/milstan-
dev1/services/sparql>;
        void:statItem :a, :b, :c .
         :a scovot2:topic <http://dbpedia.org/resource/Semantic_Web>;
        scovot2:traceType <http://rdfs.org/sioc/ns#BlogPost>;
        scovo:dimension void:numOfTriples;
        rdf:value 150.  :b scovot2:traceType
<http://rdfs.org/sioc/ns#BlogPost>;
        scovo:dimension void:numOfTriples;
        rdf:value 67890.  :c scovot2:topic
<http://dbpedia.org/resource/Semantic_Web>;
        scovo:dimension void:numOfTriples;
        rdf:value 400.
```

Figure 29 Exemple de description d'EDS en VoID et SCOVO

Un atout supplémentaire pour l'usage de statistiques d'EDS est certainement l'émergence d'approches de résumés de données (en. data summaries) tels que celles présentées dans [Katchadourian & Consens, 2010][Harth et al., 2010]. Ces approches proposent d'ajouter des index très riches aux référentiels de données stockant des EDS. Ces indexes regroupent des ressources en fonction de différentes critères et conditions qu'elles remplissent. Une telle pratique permettra le calcul rapide de nos métriques, basé sur les informations statistiques que ces index contiendront.

Nous proposons deux types de métriques :

[89] http://vocab.deri.ie/scovo/

[90] http://vocab.deri.ie/void/guide

- **Métriques basées sur la quantité de données**. La métrique la plus simple est certainement la quantité de traces d'un certain type. Nous définissons donc, Q_t comme la quantité de traces d'un certain type t disponible. La question qui se pose naturellement est : où cette quantité est-elle disponible : dans un EDS, LOD entier, etc. ? Toutes nos métriques peuvent être appliquées à différents niveaux d'espace de données, et nous discuterons leur applicabilité dans la partie §4.3.4. Une autre métrique qui s'impose naturellement est la quantité de traces d'un certain type qui remplissent une certaine condition. Dans notre cas il est intéressant d'étudier si les traces sont associées à un concept (par exemple à travers la propriété dct:subject). Nous définissons donc $Q_{t,C}$ comme le nombre de traces de type t qui sont associées à l'un des concepts de l'ensemble C. Ces deux simples métriques, bien qu'elles soient utiles en soi, ne constituent que la base nécessaire à la définition de métriques plus complexes.
- **Métriques basées sur la distribution de sujets.** Ce type de métrique permet d'évaluer, dans la totalité des traces disponibles, l'importance de la cooccurrence d'un type de trace et d'un concept. Nous définissons donc l'homogénéité de sujets (en. subject homegenity) $SH_{t,s}$ comme le nombre de traces utilisateur de type t associées avec le concept s, divisé par le nombre total de traces de type t. Cette métrique démontre le degré de présence du concept s dans les traces de type t. Nous définissons également l'homogénéité de type (en. type homogenity) $TH_{t,s}$ comme le nombre de traces utilisateur de type t associées avec le concept s, divisé par le nombre total de traces disponibles associées avec le sujet s. Cette métrique démontre la présence d'un type de trace dans l'ensemble des traces associées avec un sujet. En même temps, la valeur de cette métrique représente la limite supérieure de rappel pour toute recherche d'experts basée sur les hypothèses utilisant ce type de trace. Les formules présentant ces deux métriques en fonction des métriques simples Q_t et $Q_{t,s}$ sont:

$$SH_{t,s} = \frac{Q_{t,s}}{Q_t} \quad TH_{t,s} = \frac{Q_{t,s}}{Q_{out:UserTrace,s}} \tag{5}$$

4.3.3. Expérimentation

Après avoir défini des métriques il est nécessaire de s'assurer qu'elles sont en effet des indicateurs des vraies phénomènes de la communication dans une communauté et qu'elles peuvent servir pour la recommandation de types de traces les plus appropriées, et donc pour la recommandation d'hypothèses d'expertise. Pour ce faire, nous avons conduit une expérience empirique visant à déterminer si une corrélation existe entre les valeurs de nos métriques et les performances de la recherche d'experts. Nous avons d'abord construit un EDS d'exemple, décrit en §4.3.3.1, servant à simuler les recherches d'experts et à permettre l'exécution des hypothèses d'expertise ainsi que le calcul des métriques. Les mesures de performances pour la recherche d'experts que nous avons utilisées dans nos études sont présentées en §4.3.3.2. Afin de pouvoir étudier la corrélation entre les performances de la recherche d'experts et les métriques il fallait d'abord obtenir des informations fiables sur toutes les personnes vraiment expertes, dont les traces utilisateurs étaient disponibles dans le EDS. Nous avons produit cette

information à l'aide des utilisateurs capables de juger l'expertise des autres sur un sujet dont ils sont eux-mêmes experts, ce qui nous a permis de produire un « étalon-or » (décrit en §4.3.3.3). Tous ces éléments se croisent finalement dans un test de corrélation dont les résultats sont un indicateur de l'applicabilité de certaines de nos métriques dans certaines situations particulières (décrit en §4.3.3.4).

Dans notre expérimentation, nous nous sommes focalisés sur la communauté de recherche du Web Sémantique, plus précisément sur les traces et personnes qui y sont associées. La raison principale de ce choix était l'accessibilité des utilisateurs experts, membres de cette communauté, qui nous ont aidé à effectuer nos évaluations de manière fiable.

4.3.3.1. Jeu de données expérimental

Pour créer un jeu de données expérimentales, assez représentatives de la communauté de pratique du Web Sémantique, nous avons intégré les données de plusieurs sources, déjà existantes dans le LOD. À l'époque de notre étude, le LOD contenait deux sources principales de traces liées aux Web Sémantique : des billets de blog (instances de sioct :BlogPost) et des articles de recherche disponibles dans Semantic Web Dog Food. Nous avons importé des données de ces deux EDS dans notre référentiel local, ce qui a produit un total de 837 instances de type sioct :BlogPost et 1437 instances de type swrc :Publication (toutes les publications de conférences et ateliers majeurs organisés au sein de la communauté du Web Sémantique). Nous avons traité ces données par un logiciel de collecte et normalisation de données, décrit en §4.4.2. Ce logiciel nous a permis de corriger les imperfections éventuelles, telles que le manque de liens vers des concepts.

Ces données ont été complétées par la transformation des tweets récupérés sur le site Twitter.com ainsi que des diaporamas de site Slideshare.com en RDF. Les données de Twitter correspondaient aux dernières conférences organisées par la communauté, et les diaporamas pertinents ont été récupérés en effectuant des recherches par mot-clé sur le site SlideShare. La transformation a été effectuée par le même logiciel de collecte et de normalisation de données qui a produit un total de 6631 instances de sioc :MicroblogPost et 1657 instances de bibo :SlideShow. Construit de cette manière, ce jeu de données contient presque toutes les traces de ces quatre types pertinents pour des sujets d'intérêt de la communauté du Web Sémantique, que l'on peut rencontrer sur le Web avec un effort raisonnable à la date de sa création i.e. 15-06-2010.

4.3.3.2. Mesures de performance

Dans notre définition des mesures de performance de la recherche d'experts, i.e. précision et rappel (en. precision et recall) nous suivons une approche assez standard avec certaines adaptations mineures. Comme beaucoup d'autres auteurs, nous entendons par « précision » le rapport entre les vrais experts trouvés par le système et le nombre total d'utilisateurs trouvés et présentés en tant qu'experts. Pour mémoire, le rappel est souvent entendu comme étant le rapport entre les vrais experts trouvés par le système et le nombre total de vrais experts. Dans notre cas, il est très difficile de déterminer tous les experts existants sur un sujet ce qui remet en question notre capacité à calculer une telle mesure. En effet, il est bien plus pertinent d'évaluer la capacité de ce système à trouver les vrais experts au sein des données dont il dispose. Dans notre calcul de rappel nous nous limitons donc au rapport entre le nombre de vrais

experts trouvés par le système et le nombre total de vrais experts identifiables dans le EDS (ou bien dans plusieurs EDS) avec lequel travaille le système. Nous considérons qu'un expert n'est pas identifiable dans un EDS s'il n'existe aucun chemin dans le graphe entre la ressource représentant l'expert et la ressource (concept) représentant le sujet de son expertise. Dans notre cas le jeu de données expérimental est utilisé comme EDS de calcul.

Finalement nous utilisons également la mesure F' qui représente la moyenne harmonique de la précision et du rappel, dont la formule est donnée par la formule 6, et qui permet de favoriser un équilibre entre ces deux mesures.

$$F' = 2\frac{précision \bullet rappel}{précsion + rappel} \tag{6}$$

4.3.3.3. Création de l'étalon-or

Afin de construire un étalon-or qui sera ensuite utile pour mesurer la performance des différentes hypothèses d'expertise dans la recherche d'experts, nous avons recueilli des données des utilisateurs experts (évaluateurs) sur plusieurs sujets qui rentrent dans le cadre du travail de la communauté du Web Sémantique. Ces évaluateurs nous ont communiqué leur jugement sur qui sont des vrais experts dans notre jeu de données expérimental.

Nous avons choisi trois sujets pour cette étude : Linked Data, SPARQL et Open Data. Ces trois sujets ont été bien représentés dans notre jeu de données expérimental et sont, selon nous, assez différents pour que les ensembles de vrais experts qui leur correspondent puissent se distinguer. Il fallait d'abord trouver, pour chacun de ces sujets, toutes les ressources représentant des personnes pour lesquelles il existe un chemin dans les jeux de données expérimentales entre le sujet et la personne. Une liste de ces ressources a été créée pour chacun des 3 sujets. Ensuite, nous avons demandé à nos experts d'évaluation de marquer dans ces listes toutes les personnes qu'ils considéraient comme experts.

La question qui s'impose naturellement dans toutes les tâches de classement par plusieurs évaluateurs est le niveau d'accord entre eux. Ce niveau est le plus souvent calculé par la métrique kappa (k) définie dans [Fleiss, 1981] et reprise dans notre formule 7. Cette métrique utilise le nombre de cas d'accord et de désaccord mutuel. Chaque type d'accord/désaccord est labellisé dans le Tableau 2. Par *a* nous identifions le nombre de cas où les deux experts d'évaluation sont d'accord sur une note positive, par *b* le nombre de cas d'accord sur une note négative. Les cas de désaccord sont marqués par *c* et *d*.

$$\kappa = \frac{2(ad - bc)}{(a + c)(c + d) + (b + d)(a + b)} \tag{7}$$

Tableau 2 Matrice de confusion pour le calcul de la métrique kappa

		Evaluateur 1	
		Positive	Négative
Evaluateur 2	Positive	a	b
	Négative	c	d

Le niveau d'accord moyen n'était pas suffisant (moins de 0.6 est considéré inacceptable) après la première itération d'évaluations. Nous avons donc appliqué une méthode permettant d'atteindre l'accord entre les examinateurs en permettant des discussions contrôlées dans lesquelles les experts d'évaluation ont le droit de changer, de manière itérative, leurs notes. Cette méthode, développée en s'inspirant de la méthode Delphi [Rowe & Wright, 2009] est utilisée dans des processus de décision collaborative, représente un des apports méthodologiques de cette thèse. Elle est décrite plus en détails dans un document instructif pour son application qui est publié sur une page Web[91].

Les résultats de deux itérations qui ont conduit à un niveau d'accord admissible sur les 3 sujets sont présentés dans le Tableau 3.

Tableau 3 Le taux d'accord des évaluateurs par sujet, dans les deux itérations

	First Round	Second Round
Linked Data	0.4215	**0.6878**
SPARQL	0.4673	**0.6482**
Open Data	0.4673	**0.7228**

Nous avons finalement pris toutes les personnes marquées en tant qu'experts par au moins deux évaluateurs. Ces experts constituent notre étalon-or.

4.3.3.4. Test de corrélation

Pour chaque type de traces présent dans notre jeu de données expérimentales, nous avons calculé la précision et le rappel. Ensuite nous avons calculé les valeurs de nos 4 métriques définies dans §4.3.3.2. Le calcul est basé sur le fait que l'hypothèse la moins restrictive (celle qui exige le moins de traces d'un type pour conclure que l'utilisateur est expert) a été utilisée. Toutes les valeurs pour chacun des 3 sujets sont présentées dans le Tableau 4.

Le test de corrélation de Pearson sur ces valeurs montre l'existence d'une corrélation positive entre la métrique TH et le rappel (coefficient de la corrélation r=0.846), entre TH et F' (r=0.778) et entre SH et la précision (r=0.619). Tous les coefficients de corrélation étant au-dessus du seuil relatif à la taille de notre échantillon (r=0.576), nous pouvons conclure que le résultat est statistiquement signifiant. Cette conclusion nous permet de fonder la recommandation des types de traces, ainsi que des hypothèses

[91] http://milstan.net/stankovic_rowe_methodology/

utilisant ces types de traces, les plus adaptées pour un sujet donné, sur les métriques TH et SH, en fonction de la performance désirée (rappel ou précision).

Tableau 4 Valeurs des métriques et la performance de la recherche d'experts

t	Q_t	Q_{tc}	SH	TH	precision	relative recall	F'
Linked Data							
sioc:BlogPost	837	10	0.083	0.024	0.800	0.050	0.094
sioc:MicroblogPost	6631	96	0.014	0.236	0.469	0.296	0.363
bibo:Slideshow	1657	55	0.033	0.135	0.673	0.243	0.357
swrc:Publication	1436	86	0.060	0.211	0.721	0.408	0.521
SPARQL							
sioc:BlogPost	837	27	0.030	0.218	0.259	0.206	0.229
sioc:MicroblogPost	6631	13	0.002	0.105	0.385	0.147	0.213
bibo:Slideshow	1657	33	0.020	0.266	0.303	0.294	0.298
swrc:Publication	1436	29	0.020	0.234	0.414	0.353	0.381
Open Data							
sioc:BlogPost	837	5	0.006	0.013	0.600	0.025	0.048
sioc:MicroblogPost	6631	45	0.007	0.116	0.511	0.192	0.279
bibo:Slideshow	1657	80	0.048	0.207	0.300	0.200	0.240
swrc:Publication	1436	150	0.105	0.399	0.666	0.583	0.518

En ce qui concerne les métriques simples, s'il n'était possible de démontrer une corrélation assez importante entre Q_t et les mesures de performance, $Q_{t,C}$ montre des résultats similaires à TH, avec un coefficient de corrélation plus faible (r=0.777 avec le rappel et r=0.761 avec F'). L'usage de cette métrique est donc également envisageable.

4.3.4. Niveaux d'applicabilité des métriques

La question de l'applicabilité des métriques a déjà été évoquée. Il est clair que le résultat de l'application des métriques dépend fortement du choix d'EDS sur lequel elles sont appliquées. D'un autre coté, certaines situations peuvent restreindre ce choix. Nous présenterons ici trois niveaux d'applicabilité de ces métriques et nous discuterons leurs particularités.

4.3.4.1. Calcul des métriques sur un seul EDS

Le calcul des métriques sur un seul EDS, que nous contrôlons, a certainement l'avantage de faciliter l'obtention de statistiques et la possibilité de s'assurer de leur exactitude. Parfois ce choix est aussi le seul envisageable, lorsqu'il est impossible de trouver ailleurs les traces liées relatives à notre cas d'utilisation. Néanmoins, il peut être très risqué de n'appliquer les métriques qu'à un unique EDS, les données du jeu en question n'étant pas toujours représentatives du sujet qui nous intéresse. Un remède possible

consiste certainement à estimer sa représentativité par rapport à la structure générale du LOD entier. Il est ensuite possible, si nécessaire, de composer plusieurs EDS pour atteindre un niveau satisfaisant de similarité de la structure de l'EDS composé avec la structure du LOD.

4.3.4.2. Calcul de métriques sur un ensemble d'EDS

Une approche possible pour calculer les métriques sur plusieurs EDS consiste à intégrer leurs descriptions statistiques éventuellement existantes dans le format SCOVO et VoID. Cette approche comporte un risque lié à la qualité, parfois médiocre, des données dans le LOD, qui se répercute forcement sur la qualité de descriptions statistiques d'EDS. De nombreux auteurs, e.g. [Hogan et al., 2010][Jain et al., 2010] ont témoigné des imperfections des données dans le LOD. Deux types d'imperfections sont notamment sensibles dans notre cas d'utilisation :

- Le manque de déclaration explicite des types de ressources ;
- Le manque de liens entre des traces et des concepts.

Si l'on peut remédier à ces imperfections au niveau des EDS que nous contrôlons, par des techniques que nous utilisons dans notre processus de normalisation de données décrit en §4.4.2, il est néanmoins difficile de surmonter des problèmes de ce genre dès qu'on considère les descriptions statistiques déjà calculées sur des données imparfaites. Il est donc clairement hasardeux de s'appuyer seulement sur des descriptions statistiques, qui mettraient en péril l'exactitude de nos calculs et rendraient nos résultats approximatifs. Des approches de classement d'EDS [Toupikov et al., 2009] et de filtrage par la qualité [Bizer & Cyganiak, 2009] offrent toutefois la possibilité de juger de la fiabilité de certains EDS. Cet indicateur de fiabilité contribue à estimer le risque d'inexactitude des données utilisées.

4.3.4.3. Calcul de métriques sur le LOD entier

Dans certains cas les EDS à utiliser pour la recherche d'experts ne sont pas choisis à l'avance. Il peut donc être utile d'utiliser des métriques capables de choisir les EDS les plus adaptés au sujet d'expertise pour lequel la recherche d'experts est sollicitée. La stratégie déjà décrite qui utilise des descriptions statistiques pourrait également être appliquée ici. Des systèmes [www.hpi.uni-potsdam.de/naumann/sites/bte2010] qui génèrent automatiquement des descriptions statistiques d'EDS liées publiques pourront s'avérer utiles dans cette tâche. Une stratégie alternative consiste à se servir d'un moteur de recherche de Web Sémantique, tel que Sindice pour obtenir des statistiques sur le nombre d'instances de traces de type particulier et d'autres statistiques nécessaires. Il est toutefois important de garder à l'esprit que cette approche est tout aussi approximative, étant donné la nature incomplète des index des moteurs de recherche.

4.4. Hy.SemEx – Système de recherche d'experts sur les EDS du LOD

Afin de mettre en œuvre notre approche de suggestion d'hypothèse d'expertise dans un cadre réel d'un scénario de recherche d'experts, nous avons développé un système,

HySemEx. Ce système est notamment conçu pour profiter de traces utilisateurs disponibles dans le LOD. Il n'est spécialisé pour aucun sujet de compétence, mais vise plutôt une flexibilité importante qui émerge en conséquence de nos choix de conceptions, décrits par la suite.

4.4.1. HySem.Ex dans le contexte de son usage

Le système hy.SemEx est centré sur la notion de l'hypothèse d'expertise. Contrairement à la plupart des systèmes, qui ne dévoilent pas la manière dont ils sélectionnent et classent les experts, hy.SemEx met en évidence la notion d'hypothèse d'expertise et permet à l'utilisateur de choisir une hypothèse avant de s'engager dans la recherche d'experts. Si ce pas supplémentaire introduit certainement un niveau de complexité sur le plan de l'interaction avec l'utilisateur, il lui donne aussi plus de contrôle. Le fait que hy.SemEx soit conçu pour la recherche d'experts à partir des larges masses de données provenant du LOD, amplifié par une variété de concepts de connaissances humaines avec lesquelles le système doit pouvoir travailler, permet une certaine hétérogénéité de situations d'usage. C'est cette hétérogénéité qui rend nécessaire, même indispensable, un degré de contrôle de la part de l'utilisateur. Ce contrôle est implémenté sous la forme de différents choix qui lui sont proposés lors de son interaction avec le système.

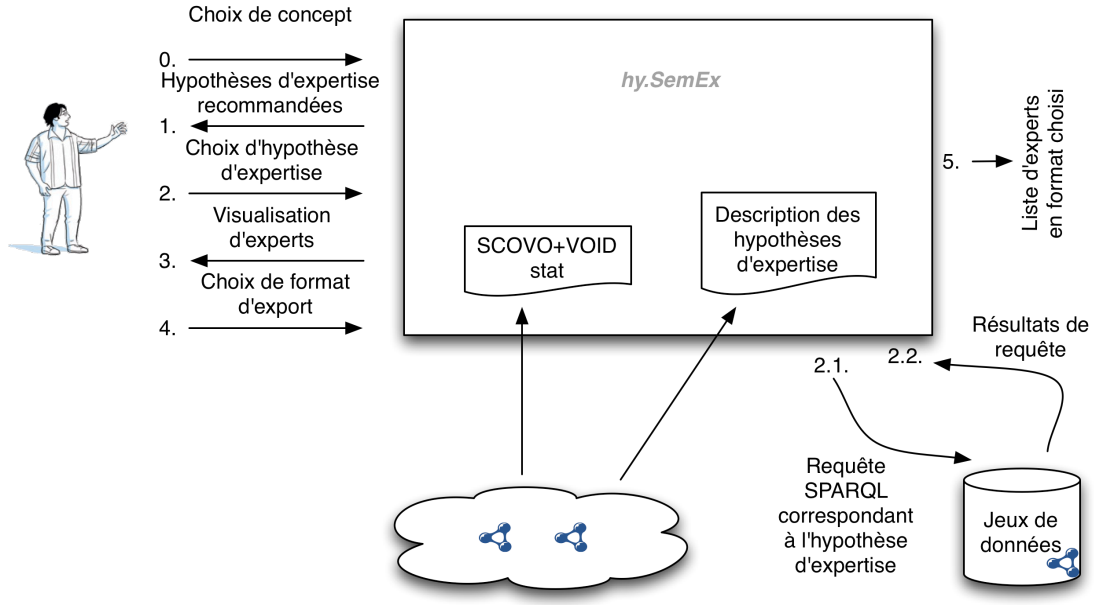


Figure 30 Mode d'interaction avec le système Hy.SemEx

Le mode d'interaction de l'utilisateur avec le système est décrit en Figure 30. Pour commencer (0), l'utilisateur fournit d'abord un concept. Les URIs DBPedia sont utilisés pour désambigüiser des concepts au lieu de mots-clés textuels utilisés couramment dans d'autres systèmes. Ensuite, hy.SemEx analyse les données qui lui sont disponibles afin d'effectuer une recommandation d'hypothèse d'expertise appropriée pour le sujet donné. Pour ce faire, le système consulte les descriptions statistiques des EDS avec lesquelles il peut travailler et qui peuvent lui être fournies dans le format RDF en utilisant les ontologies SCOVO et VoID. Un exemple d'une telle description a été déjà donné sur la Figure 29. L'utilisateur peut également spécifier une mesure de performance qu'il souhaite favoriser dans sa recherche d'experts (e.g., précision, rappel). Ces données statistiques peuvent être fournies dans le cadre de la configuration

du système. Mais grâce à la disponibilité croissante des données statistiques concernant la composition d'EDS en RDF, un système d'inclusion automatique de ces données directement à partir du Web est aussi envisageable.

En fonction de ces paramètres (mesure de performance et données statistiques) ainsi que des hypothèses d'expertise qui sont connus au système, il recommande les hypothèses (1) qui sont les mieux adaptées à la recherche en question. Le mécanisme de recommandation, ainsi que d'autres aspects de fonctionnement technique du système seront décrits plus en détails plus tard. Parmi quelques hypothèses recommandées, l'utilisateur choisit celle (2) qu'il juge la plus appropriée. Une capture d'écran du système lors de cette phase d'interaction est donnée dans la

Figure 31. Ensuite, une requête SPARQL correspondant à l'hypothèse d'expertise est lancée sur les EDS RDF préexistants, dans lesquels sont intégrées des traces utilisateur de plusieurs sources (2.1). Comme résultat de cette requête (2.1) une liste d'experts est montrée à l'utilisateur (3) qu'il peut ensuite exporter (4 et 5) en différents formats (e.g. RDF, XML, CSV). Une capture d'écran avec une liste d'experts est consultable à la Figure 32. Des listes de ce genre sont ensuite utilisées chez Hypios dans un processus de diffusion de problèmes par un contact direct avec les auteurs potentiels de solutions.

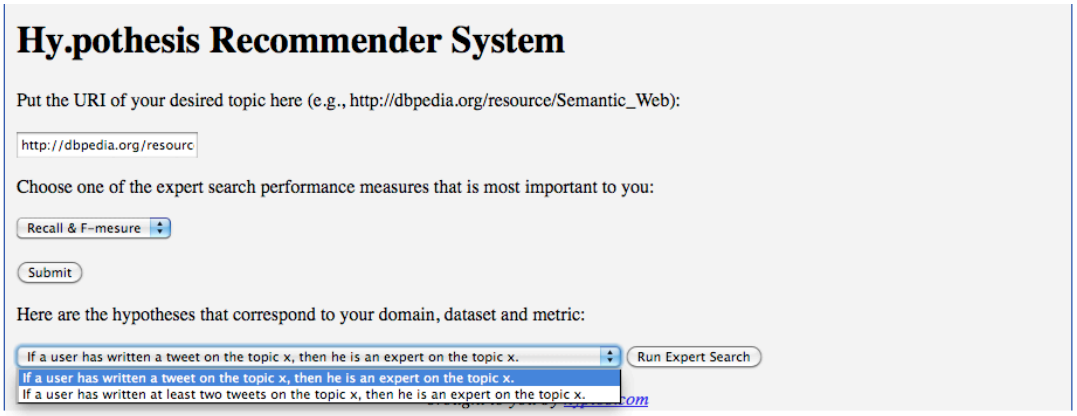


Figure 31 Capture d'écran de Hy.SemEx - choix d'hypothèse d'expertise

URI	First Name	Last Name	Full Name	Postal Address	Telnr	Fax	Type	Unique ID	Select
http://data.hypios.com/tweets/user-digikim			digikim					http://data.hypios.com/tweets/user-digikim	true
http://data.hypios.com/tweets/user-lysander07			lysander07					http://data.hypios.com/tweets/user-lysander07	true
http://data.hypios.com/tweets/user-terraces			terraces					http://data.hypios.com/tweets/user-terraces	true
http://data.hypios.com/tweets/user-juansequeda			juansequeda					http://data.hypios.com/tweets/user-juansequeda	true
http://data.hypios.com/tweets/user-skruk			skruk					http://data.hypios.com/tweets/user-skruk	true
http://data.hypios.com/tweets/user-rszeno								http://data.hypios.com/tweets/user-rszeno	true
http://data.hypios.com/tweets/user-Ozelin			Ozelin					http://data.hypios.com/tweets/user-Ozelin	true
http://data.hypios.com/tweets/user-milstan			milstan					http://data.hypios.com/tweets/user-milstan	true
http://data.hypios.com/tweets/user-AxelPolleres			AxelPolleres					http://data.hypios.com/tweets/user-AxelPolleres	true
http://data.hypios.com/tweets/user-toniher			toniher					http://data.hypios.com/tweets/user-toniher	true
http://data.hypios.com/tweets/user-zbeauvais			zbeauvais					http://data.hypios.com/tweets/user-zbeauvais	true
http://data.hypios.com/tweets/user-ullrich								http://data.hypios.com/tweets/user-ullrich	true
http://data.hypios.com/tweets/user-johnbreslin								http://data.hypios.com/tweets/user-johnbreslin	true
http://data.hypios.com/tweets/user-purbon								http://data.hypios.com/tweets/user-purbon	true
http://data.hypios.com/tweets/user-pimecca								http://data.hypios.com/tweets/user-pimecca	true
http://data.hypios.com/tweets/user-yovisto			yovisto					http://data.hypios.com/tweets/user-yovisto	true
http://data.hypios.com/tweets/user-webr3			webr3					http://data.hypios.com/tweets/user-webr3	true
http://data.hypios.com/tweets/user-rgaidot								http://data.hypios.com/tweets/user-rgaidot	true
http://data.hypios.com/tweets/user-secoresearch			secoresearch					http://data.hypios.com/tweets/user-secoresearch	true
http://data.hypios.com/tweets/user-tomayac								http://data.hypios.com/tweets/user-tomayac	true
http://data.hypios.com/tweets/user-cgueret			cgueret					http://data.hypios.com/tweets/user-cgueret	true
http://data.hypios.com/tweets/user-clauwa			clauwa					http://data.hypios.com/tweets/user-clauwa	true

Figure 32 Capture d'écran de Hy.SemEx - Liste d'experts

4.4.2. Architecture du logiciel

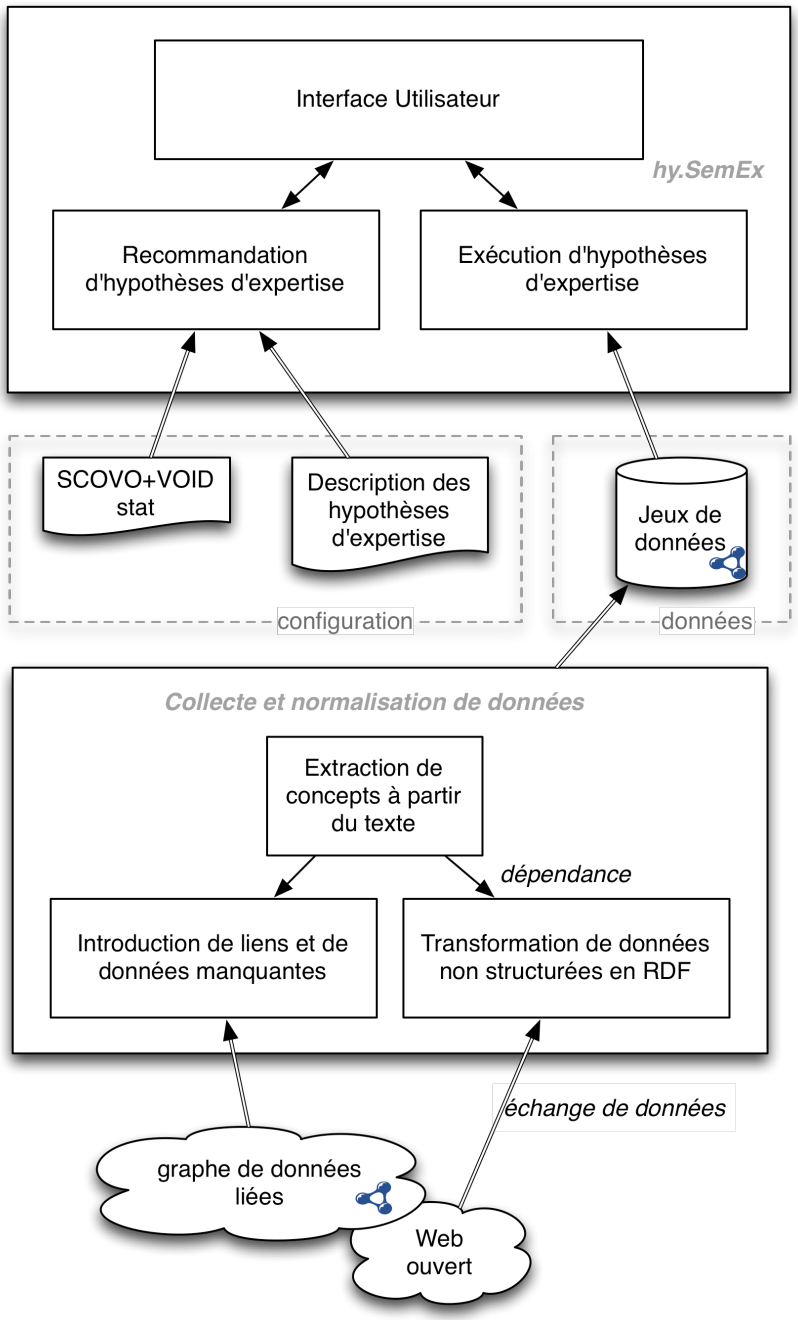


Figure 33 Principaux composants du logiciel Hy.SemEx et du logiciel complémentaire pour la collecte et la normalisation de données

Hy.Semex repose sur une interface utilisateur écrite en Java et s'appuie sur le framework de développement d'applications Web Tapestry[92], et sur deux composants fonctionnels principaux :

1. Le module de recommandation des hypothèses d'expertise ;
2. Le module d'exécution des hypothèses d'expertise.

[92] http://tapestry.apache.org/

Le premier prend en entrée un concept donné par l'utilisateur, ainsi qu'une éventuelle préférence pour la précision ou le rappel. L'utilisateur n'est pas obligé de connaître les notions de précision et de rappel. Il choisit plutôt s'il veut favoriser la sélection des meilleurs experts (précision plus importante) ou bien favoriser la capacité de retrouver le plus d'experts possible (rappel). Ce module se sert des métriques définies SH (en cas de préférence pour la précision) et TH (en cas de préférence pour le rappel), pour choisir le type de trace utilisateur le plus approprié pour le sujet donné (ces métriques ont été auparavant définies dans la partie §4.3.2). Ensuite, le module sélectionne les hypothèses d'expertise basées sur ce type de trace. C'est à l'utilisateur de trier parmi ces hypothèses et choisir celle qui lui paraît la plus pertinente pour sa recherche.

Ce module utilise des descriptions de la distribution statistique des traces utilisateurs disponibles dans les EDS, qui peuvent lui être fournies en format RDF à l'aide des ontologies VoID et SCOVO, ainsi qu'un vocabulaire spécifique pour la description de traces utilisateurs que nous avons conçu spécifiquement pour cette tâche, OUT[93] (en. Ontology of User Traces). Les hypothèses d'expertise avec lesquelles le système va travailler doivent également lui être fournies. Pour représenter leurs différentes caractéristiques, nous avons spécifiquement conçu une ontologie EHO[94] (en. Expertise Hypothesis Ontology). Celle-ci est décrite plus en détails dans le chapitre 8, qui reprend la totalité de notre travail de conception d'ontologies pour les scénarios de l'innovation sur le Web.

Le deuxième module prend en entrée une hypothèse d'expertise que l'utilisateur a choisie suivant les recommandations du premier module. Il exécute la requête SPARQL associée à l'hypothèse sur un EDS (contenant des traces utilisateurs) rassemblées au préalable. Il est développé à l'aide de Jena[95], une librairie Java pour le Web Sémantique. Ce module est construit pour répondre à une certaine flexibilité. Les requêtes SPARQL sont stockées avec les descriptions des hypothèses d'expertise ce qui implique la capacité du système à intégrer de nouvelles hypothèses d'expertise, non définies au moment de la création du logiciel, et de travailler avec elles sans modification du code.

La question qui se pose naturellement est celle de la nature et de la provenance des EDS avec lesquels travaille le système. Une flexibilité est introduite à ce niveau permettant de facilement configurer le système afin qu'il puisse travailler avec des EDS choisis par l'utilisateur est spécifié dans un fichier de configuration. Pour pouvoir tester les performances de notre système il a fallu avoir un EDS avec des traces utilisateur de types différents afin de simuler une hétérogénéité de traces, intrinsèque aux situations d'usage réelles. Suite aux problèmes de composition actuelle du LOD en terme de traces utilisateur identifiées précédemment dans la partie §4.2.3, nous avons décidé de construire un EDS qui contient à la fois des données existantes de LOD enrichies, pour y réparer des défauts communs, et les nouvelles données structurées construites d'après les contenus Web textuels (tel que les blogs par exemple). La stratégie d'avoir un jeu de donnée local dans lequel l'on intègre plusieurs sources fut nécessaire pour combattre des imperfections dans le LOD public. Dans l'avenir avec une amélioration de la qualité de données liées publiques, qui, nous l'espérons, viendra prochainement, il sera également intéressant de considérer des stratégies d'exécution distribuées de requêtes

[93] http://ontologies.hypios.com/out

[94] http://ontologies.hypios.com/eho

[95] http://jena.apache.org/

SPARQL. Une telle approche permettrait d'interroger plusieurs EDS sans devoir les intégrer localement. Des approches pour effectuer des requêtes distribuées ont été proposés, entre autres, dans [Quilitz & Leser, 1008], [Langegger et al., 2008].

Notre logiciel de collecte et normalisation de données est constitué, lui aussi de deux modules principaux :

- Le module d'introduction de liens et de données manquant dans des données liées ;
- Le module de transformation de données non structurées en RDF ;

Ces deux modules s'appuient sur un module accessoire qui sert à extraire des concepts à partir du texte. Ce dernier fait appel à Zemanta une API Web dont l'une des fonctions est d'effectuer l'extraction de URL Wikipedia pour des concepts mentionnés dans un texte fourni. La transformation de ces URLs en URI DBPedia est assez simple et non ambigüe.

Le premier module prend des ressources du LOD déjà importées dans le EDS créé et essaye d'identifier le maximum de textes pouvant lui être associé, pour le fournir à Zemanta et en obtenir des concepts. Ces concepts sont ensuite liés aux ressources existantes, dans le cas où celles-ci n'ont pas déjà des concepts associés. Notre module travaille actuellement avec deux types de traces utilisateur :

- **Publications académiques**. Ces traces sont directement importées par simple téléchargement d'EDS dans le référentiel local de Jena. Ensuite, la partie du contenu textuel utilisée est obtenu en combinant les propriétés représentant le titre (dc :title), l'abstract(swrc :abstract) et d'autres éléments textuels (dc :description, sioc :content) de l'article.
- **Billets de blog**. Ces traces sont importées en cherchant toutes les instances du type sioc :BlogPost sur le moteur de recherche de Web Sémantique, Sindice.[96] Si à l'heure actuelle ce moteur de recherche de Web de données est le plus utilisé, il est parfaitement envisageable d'utiliser d'autres outils pour trouver des instances de billets de blog (décrits par SIOC ou par une autre ontologie). Dans le cas d'instances sans concepts associés nous avons dû effectuer une extraction de concepts à partir du texte. Si les éléments textuels (dc :title, dc :description, sioc :content) étaient présents, nous les avons utilisés pour cette tâche. Dans d'autres cas, nous avons dû chercher l'adresse URL du billet pour effectuer l'extraction de concepts à partir du contenu textuel de sa page Web. Cette tâche n'était pas toujours évidente car l'URL n'était pas toujours présente parmi les informations disponibles. Pour pallier ce problème nous avons dû concevoir des heuristiques devinant l'URL envisageable à partir du URI de la ressource décrivant le billet de blog. L'information sur le créateur du billet du blog était également souvent manquante. Dans des cas de ce type, nous avons créé des ressources de type foaf :Person pour décrire l'auteur du billet.

Le deuxième module prend en entrée des mots-clés (noms de concepts) en fonction desquels il peut sonder le Web à la recherche de différents types de traces non structurées et les convertir en RDF. Ce module travaille avec deux types de traces :

[96] Nous avons créé un module spécifique pour effectuer les recherches sur Sindice à partir du code Java. Ce module a été publié en tant que librairie à source ouverte Sindice4J, http://code.google.com/p/sindice4j/

- **Diaporamas en ligne**. Les diaporamas en ligne, notamment ceux partagés dans le site d'échange de diaporamas Slideshare.net sont une source de contenu de référence sur une multitude de sujets de connaissance humaine. Nous avons donc développé une fonctionnalité qui prend des URLs de diaporamas pertinents pour des mots-clés donnés, extrait des concepts à partir du contenu textuel du diaporama présent sur sa page Web et nous avons construit des données RDF le décrivant. Les données sur l'auteur du diaporama sont extraites de son profil Slideshare.
- **Tweets (billets de microblog)**. Notre approche pour l'extraction des informations à partir des microblogs, qui se base principalement sur les données de Twitter est similaire à celui du traitement des diaporamas. La description détaillée de cette approche est donnée dans le chapitre 6 qui concerne, entre autre, le traitement de tweets et leur représentation en RDF.

La conception logicielle de ces modules suit le principe de bonne pratique (en. Design Patterns) ce qui permet d'ajouter de nouveaux extracteurs relativement facilement sans trop d'intervention sur le code existant.

4.4.3. Retour d'utilisateurs

Afin de s'assurer de la capacité de notre système à répondre à un vrai besoin des utilisateurs et à les satisfaire par son fonctionnement, nous avons conduit une enquête de satisfaction. Dix utilisateurs ont répondu à notre questionnaire et trois parmi eux ont également accepté de participer à des entretiens détaillés qui nous ont permis d'obtenir une vision plus précise des points forts et point faibles de notre système Hy.SemEx.

Les questionnaires avaient pour objectif principal d'évaluer si les hypothèses d'expertise suggérées par le système étaient appropriées. Il était donc très important que les évaluateurs soient eux-mêmes experts du sujet pour lequel le système était utilisé, afin qu'ils puissent juger si le type de trace suggéré était vraiment susceptible d'identifier les bons experts pour le sujet en question. Chaque participant a testé le système avec 7-10 sujets de son domaine d'expertise. Les participants ont noté les suggestions faites par le système sur une échelle de 1 à 5 (1 : les hypothèses utilisant ce type de trace ne devraient pas être utilisées pour la recherche d'experts sur se sujet ; 5 : ce type de trace est le meilleur pour la recherche d'experts sur ce sujet). Pour chaque sujet, les participants ont traité le cas où la précision est favorisée et le cas où le rappel est favorisé. Cela a permis de générer 176 évaluations au total.

Dans le cas de focalisation sur la précision, la satisfaction moyenne était de 4.234±0.857[97]. Dans le cas de focalisation sur le rappel, de 3.947±0.751. Ce résultat indique que les suggestions émises par notre système ont suscité une satisfaction raisonnable. Dans l'analyse, il est important de tenir compte du fait que le système a suggéré des hypothèses correspondant à un seul type de trace, qui était le mieux classé par des métriques SH ou TH. Ce résultat ne témoigne donc pas seulement d'une aptitude générale de ces métriques à identifier de bons types de trace, mais aussi d'une capacité à identifier comme meilleur le type de trace qui sera jugé pertinent pour des experts.

Après avoir reçu les suggestions de notre système, les participants ont pu choisir une hypothèse parmi celles suggérées ou dans la liste des autres hypothèses. Nous avons

[97] moyenne±écart type

remarqué une certaine régularité dans leur choix d'hypothèses. Par exemple, dans le cas de focalisation sur la précision, les utilisateurs ont préféré l'hypothèse considérant comme experts des personnes ayant écrit au moins 5 billets de blog sur le sujet parmi des hypothèses utilisant des billets de blog comme type de trace, alors que dans le cas de focalisation sur le rappel, il était suffisant d'être l'auteur de 2 billets de blog sur le sujet. Lors des entretiens, les participants ont expliqué que le caractère restrictif d'une hypothèse était très important dans leur choix. Cette propriété pourrait en effet être utile non seulement dans la suggestion de la classe d'hypothèses utilisant le même type de trace, mais aussi d'une hypothèse particulière. Les utilisateurs ont également indiqué le besoin de pouvoir changer les paramètres numériques des hypothèses, et ont, en général, fait l'éloge du contrôle de processus et de la transparence offerte par le système.

4.5. Conclusions

Dans ce chapitre, nous avons abordé la problématique du rapprochement entre utilisateurs et concepts. Dans le cadre de la recherche d'experts, les concepts correspondent aux sujets d'expertise des utilisateurs. Le rapprochement repose sur une notion clé, celle de l'hypothèse d'expertise. C'est une règle qui définit les conditions nécessaires pour considérer un utilisateur comme spécialiste en la matière L'application de cette règle dépend des traces, i.e. de des informations sur les utilisateurs dont le Web dispose. Plusieurs facteurs sont à l'origine de la richesse des traces laissées sur le Web : la richesse de la communication humaine sur le Web, le type d'échanges et les canaux de communication. C'est grâce à une telle richesse que nous avons pu concevoir une approche sémantique aussi souple. De plus, la nature structurée des traces sur le Web Sémantique Social nous a permis de proposer des métriques, SH et HT, quantifiant l'importance de certains types de trace pour un certain concept. L'étude que nous avons conduit à l'aide d'un « étalon-or » nous a amené à conclure qu'une corrélation existe entre les valeurs de nos métriques et les performances de la recherche d'experts. Plus les valeurs calculées pour un couple <type de trace, concept> était élevées, plus la recherche d'experts était utile. Les deux métriques sont corrélées différemment avec différentes mesures de performances. Il est donc possible, en utilisant une métrique particulière, de guider la recherche d'experts vers une précision ou un rappel optimal.

Grâce à nos métriques, il est possible de suggérer de meilleures hypothèses d'expertise à appliquer dans une recherche d'experts, sur des parties du graphe Web particulières. Nous avons implémenté cette approche de suggestion d'hypothèses et de recherche d'experts sous forme d'un prototype. Les premiers retours de la part des utilisateurs qui ont testé ce système ont été très favorables. Ils ont notamment apprécié la souplesse induite par les suggestions d'hypothèses d'expertise données par le système.

Bibliographie

[Abel et al., 2011] Abel, F., Gao, Q., Houben, G. J., & Tao, K. (2011). Analyzing Temporal Dynamics in Twitter Profils for Personalized Recommendations in the Social Web. *Web Science Conference*. Koblenz.

[Adamic et al., 2008] Adamic, L., Zhang, J., Bakshy, E., & Ackerman, M. (2008). Knowledge sharing and yahoo answers: everyone knows something. *In Proceedings of the 17th international*

conference on World Wide Web (pp. 665-674). Beijing, China: ACM New York, NY, USA. Retrieved from http://portal.acm.org/citation.cfm?id=1367497.1367587.

[Aleman-Meza et al., 2007] Aleman-Meza, B., Bojars, U., Boley, H., Breslin, J., Mochol, M., Nixon, L., et al. (2007). Combining RDF vocabularies for expert finding. *Lecture Notes in Computer Science*, 4519, 235. Springer. Retrieved from http://www.springerlink.com/index/p6u10781711xp102.pdf.

[Amitay et al., 2008] Amitay, E., Carmel, D., Golbandi, N., Har'El, N., Ofek-Koifman, S., Yogev, S., et al. (2008). Finding people and documents, using web 2.0 data. *In Proceedings of the SIGIR 2008 Workshop on Future Challenges in Expertise Retrieval* (pp. 1-6). Retrieved from http://ilps.science.uva.nl/fCHER/files/slides/fcher.harel.pdf.

[Becara-Fernandez, 2006] Becerra-Fernandez, I. "Searching for experts on the web: a review of contemporary expertise locator systems," *ACM Trans. on Internet Technology*, vol. 6, no. 4, pp. 333-355, Nov. 2006.

[Bizer & Cyganiak, 2009] Christian Bizer and Richard Cyganiak. Quality-driven information filtering using the wiqa policy framework. *Journal of Web Semantics:* Science, Services and Agents on the World Wide Web, 7(1):1–10, 2009.

[Brin & Page, 1998] Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In: *Seventh International World-Wide Web Conference (WWW 1998)*, April 14-18, 1998, Brisbane, Australia.

[Brinker, 2010] Brinker, S. 2010. 7 business models for linked data. [Online]. Available at: http://www.chiefmartec.com/2010/01/7-business-models-for-linked-data.html

[Buitelaar & Eigner, 2008] Buitelaar, P., &Eigner, T. (2008). Topic Extraction from Scientific Literature for Competency Management. *The 7th International Semantic Web Conference*. Karlsruhe, Germany.

[CCNESVS, 2010] Comité Consultatif National d'Éthique pour les Sciences de la Vie et de la Santé (France), Communication d'informations scientifiques et médicales et société: enjeux éthiques.- Mars 2010. Retrieved from: http://www.upf.edu/pcstacademy/_docs/CCNE-Avis109.pdf

[Chua, 2007] Chua, S. J. (2007). Using web 2.0 to locate expertise. *IBM Centre for Advanced Studies Conference*. Retrieved from http://portal.acm.org/citation.cfm?id=1321250.

[Demartini & Niederée, 2008] Demartini, G., &Niederée, C. (2008). Finding experts on the semantic desktop. *The 7th International Semantic Web Conference*. Karlsruhe, Germany. Retrieved from http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-403/ISWC_PICKME08.pdf#page=23.

[Demartini, 2007] Demartini, G. (2007). Finding experts using wikipedia. In Proceedings of the Workshop on Finding Experts on the Web with Semantics (FEWS2007) at *ISWC/ASWC2007*, (pp. 33-41). Busan, South Korea, Retrieved from http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-290/paper03.pdf.

[Dodds, 2010] Dodds, L. 2010. Thoughts on Linked Data Business Models. [Online]. Available at: http://www.ldodds.com/blog/2010 /01/thoughts-on-linked-data-business-models/

[Euzenat & Schvaiko, 2007] Euzenat, J., & Shvaiko, P. (2007). *Ontology Matching*. Berlin-Heidelberg: Springer-Verlag. Retrieved from http://book.ontologymatching.org/

[Fleiss, 1981] Fleiss, J. L. (1981) *Statistical methods for rates and proportions*. 2nd ed. (NY: John Wiley)

[Fu et al., 2007] Fu, Y., Xiang, R., Liu, Y., & Zhang, M. (2007). Finding Experts Using Social Network Analysis. *IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)*. IEEE. doi: 10.1109/WI.2007.14.

[Giles, 2005] Giles, J. (2005). "Internet encyclopaedias go head to head".*Nature* 438 (7070): 900–1. Bibcode 2005Natur.438..900G.DOI:10.1038/438900a.

[Harth et al., 2010] Harth, A., Hose, K., Karnstedt, M., Polleres, A., & Sattler, K. (2010). Data Summaries for On-Demand Queries over Linked Data. *In Proceedings of the 17th international conference on World Wide Web, WWW2010* (pp. 411-420). Raleigh, NC, USA: ACM Press.

[Hogan et al., 2010] Aidan Hogan, Andreas Harth, Alexandre Passant, Stefan Decker, and Axel Polleres. Weaving the Pedantic Web. *In 3rd International Workshop on Linked Data on the Web (LDOW2010) at WWW2010*, volume 628 of CEUR Workshop Proceedings. CEUR-ws.org, 2010.

[Jain et al., 2010] Prateek Jain, Pascal Hitzler, Peter Z. Yeh, Kunal Verma, and Amit P.Sheth, Linked Data Is Merely More Data. In: Dan Brickley, Vinay K. Chaudhri, Harry Halpin, and Deborah McGuinness: Linked Data Meets Artificial Intelligence. Technical Report SS-10-07, AAAI Press, Menlo Park, California, 2010, pp. 82-86.

[Jentzsch et al., 2010] Jentzsch, A., Isele, R., & Bizer, C. (2010). Silk – Generating RDF Links while publishing or consuming Linked Data. In P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, et al. (Eds.), *Proceedings of the 9th International Semantic Web Conferene - ISWC 2010* (pp. 1-4). Shanghai, China: Springer Verlag.

[Jurezyk & Agichten, 2007] Jurczyk, P., &Agichtein, E. (2007). Discovering authorities in question answer communities by using link analysis. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management - CIKM '07*. New York, New York, USA: ACM Press. doi: 10.1145/1321440.1321575.

[Katchadourian & Consens, 2010] Khatchadourian, S., & Consens, M. (2010). Exploring RDF Usage and Interlinking in the Linked Open Data Cloud using ExpLOD. *In Proc. Linked Data on the Web Workshop 2010* on the WWW2010 (Vol. 1430). Raleigh, NC.

[Kleinberg, 1999] J. Kleinberg. Authoritative sources in a hyperlinked environment. J. ACM, 46(5):604–632, 1999.

[Kolari et al., 2008] Kolari, P., Finin, T., Lyons, K., &Yesha, Y. (2008). Expert Search using Internal Corporate Blogs. *In Workshop on Future Challenges in Expertise Retrieval, SIGIR 2008* (pp. 2-5). Retrieved from http://ilps.science.uva.nl/fCHER/files/fCHER_proceedings.pdf#page=9.

[Langegger et al., 2008] Langegger, A., Woss, W, Blochl, M.: A semantic web middleware for virtual data integration on the web. In: *Proceedings of the 5th European Semantic Web Conference (ESWC)*. Volume 5021 of Lecture Notes in Computer Science., Springer Verlag (June 2008) 493–507

[Letierce et al. 2010] Letierce, J., Passant, A., Breslin, J., & Decker, S. Understanding how Twitter is used to spread scientific messages. *Proc. of the Web Science Conference, 26-27 April 2010,* Raleigh, NC, USA

[Nikolov et al., 2009] Nikolov, A., Uren, V., Motta. E. 2009. Towards Data Fusion in a Multi-ontology Environment. *Proceedings of the WWW2009 Workshop on Linked Data on the Web, Madrid, Spain*. Avaialble at: http://ceur-ws.org/Vol-538/ldow2009_paper15.pdf

[Noll et al., 2009] Noll, M. G., Yeung, C. A., Gibbins, N., Meinel, C., & Shadbolt, N. (2009). Telling Experts from Spammers: Expertise Ranking in Folksonomies. *In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. Boston, MA: ACM, New York, USA.

[Orlandi & Passant, 2010] Orlandi, F., & Passant, A. (2010). Semantic Search on Heterogeneous Wiki Systems. WikiSym '10: *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*: (pp. 1-10). ACM.

[Pellegrini, 2009] Pellegrini, T. 2009. Linked Data Flows: A new picture to illustrate the "openness" we mean. [Online]. Available at: http://blog.semantic-web.at/2009/10/28/linked-data-flows-a-new-picture-to-illustrate-the-openness-we-mean/

[Quilitz & Leser, 1008], Quilitz, B., Leser, U.: Querying distributed RDF data sources with SPARQL. In: *Proceedings of the 5th European SemanticWeb Conference (ESWC)*. Volume 5021 of Lecture Notes in Computer Science., Springer Verlag (June 2008) 524–538

[Rowe & Wright, 2009] Rowe & Wright (1999): The Delphi technique as a forecasting tool: issues and analysis. *International Journal of Forecasting*, Volume 15, Issue 4, October 1999.

[Sloane, 2011] Sloane, P. (2011). *A Guide to Open Innovation and Crowdsourcing: Advice From Leading Experts*. Kogan Page. Retrieved from http://www.amazon.com/Guide-Open-Innovation-Crowdsourcing-Leading/dp/0749463074

[Speidel, 2011] Klaus-Peter Speidel, "Problem-Description in Open Problem-Solving. How to overcome Cognitive and Psychological Roadblocks", *A Guide to Open Innovation and Crowdsourcing. Advice from Leading Experts*, edited by Paul Sloane, KoganPage, London, 2011

[Stan et al. 2011] Stan, J.; Viet-Hung Do; Maret, P., "Semantic User Interaction Profils for Better People Recommendation," Advances in Social Networks Analysis and Mining (ASONAM), 2011, M Rowe, S Angeletou and H Alani. Predicting Discussions on the Social Semantic Web. *In Proceedings of Extended Semantic Web Conference 2011*. Heraklion, Crete. (2011)

[Stankovic et al., 2010] Stankovic, M., Wagner, C., Jovanovic, J., & Laublet, P. (2010). Looking for Experts ? What can Linked Data do for You ? *Proceedings of Linked Data on the Web 2010, on WWW2010*. Raleigh, NC. Retrieved from http://events.linkeddata.org/ldow2010/papers/ldow2010_paper19.pdf

[Stankovic, 2010] Stankovic, M., Open Innovation and Semantic Web: Problem Solver Search on Linked Data. In *Proc. of International Semantic Web Conference 2010*; 7th-11th Nov., 2010, Shanghai, China

[Story et al., 2009] Story, H., Harbulot, B., Jacobi, I., & Jones, M. (2009). FOAF+TLS: RESTful Authentication for the Social Web. *Proceedings of the SPOT workshop at ESWC2009*.

[Tang et al., 2007] Tang, J., Zhang, J., Zhang, D., Yao, L., Zhu, C., Li, J., et al. (2007). ArnetMiner: An Expertise Oriented Search System for Web Community. *In Proceedings of Semantic Web Challenge'2007*.

[Toupikov et al., 2009] Toupikov, N., Umbrich, J., Delbru, R., Hausenblas, M., & Tummarello, G. (2009). DING! Dataset Ranking using Formal Descriptions. *WWW 2009 Workshop: Linked Data on the Web (LDOW2009)*, Madrid, Spain. Retrieved from: http://sw-app.org/pub/ldow09-ding.pdf.

[Wagner, 2010] Wagner, C. (2010). Exploring the Wisdom of the Tweets: Towards Knowledge Acquisition from Social Awareness Streams. *PhD Symposium at 7th Extended Semantic Web Conference (ESWC2010)* (pp. 493-497). Heraklion, Crete, Grece: Springer.

[Zhang et al., 2007] Zhang, J., Ackerman, M. S., &Adamic, L. (2007). Expertise Networks in Online Communities: Structure and Algorithms. In USA, 2004. ACM Press. WWW '07: *Proceedings of the 16th international conference on World Wide Web* (pp. 221-230). New York, NY, USA: ACM Press.

[Ziamatin, 2011] Ziaimatin, H. (2011). DC Proposal : Capturing Knowledge Evolution and Expertise in Community-driven Knowledge Curation Platforms. *International Semantic Web Conference*. Bonn, Germany.

[Zoltan & Johan, 2011] Zoltan, K.; Johann, S.;, "Semantic analysis of microposts for efficient people to people interactions," *Roedunet International Conference (RoEduNet),* 2011 10th , vol., no., pp.1-4, 23-25

5. Rapprochement de concepts

Ce chapitre aborde la question du rapprochement de concepts. La première partie (§6.1) met en avant le contexte des rapprochements et souligne les besoins de telles approches dans la vie réelle. La partie §6.2 expose les approches issues de l'état de l'art. La partie §6.3 est consacrée à l'étude de notre approche de rapprochement et de ses versions alternatives et les comparaisons de leurs performances. La partie §6.4 présente les résultats des comparaisons de performance entre notre approche et celles issues de l'état de l'art. Nous verrons ensuite, dans la partie §6.5, le prototype logiciel fonctionnel que nous avons développé dans le cadre de la thèse et qui met en œuvre notre approche. Ce prototype (HyProximity Prototype) a été utilisé au sein de la compagnie Hypios. Il nous a entres autres permis d'effectuer des études de cas afin d'identifier et d'appréhender son utilité dans la pratique. Ces études sont présentées dans la partie §6.6.

5.1. Rapprochement des concepts au service de la découverte et de la latéralité dans la résolution de problèmes

5.1.1. Contexte de rapprochement

Dans un grand nombre de scénarios d'interactions avec le Web, les concepts identifiant des sujets d'intérêt sont des éléments clé pour l'appréciation de la pertinence des éléments Web. Ces concepts désignent, d'une manière non ambigüe et déterminante, des sujets de documents et d'autres éléments Web. Par exemple, grâce aux concepts d'une trace utilisateur laissée sur le Web il est possible d'identifier à la fois les thématiques de cette trace ainsi que des intérêts des utilisateurs, consommateurs et auteurs de celle-ci. Il est donc notamment important d'étudier les concepts et la nature de leurs liens pour concevoir des notions de pertinence utiles pour différents rapprochements dans des scénarios liés à l'innovation.

Pour illustrer l'importance des concepts, dans le cadre de scénarios liés à l'innovation, imaginons un problème associé à des concepts t1 et t2. Si nous devions effectuer une recherche d'experts correspondant à ce problème, et que nous devions nous limiter à la recherche d'experts dont les profils contiennent ces deux concepts, il est probable que nous manquerions certains experts susceptibles de proposer des solutions. Ceci peut provenir de plusieurs facteurs, comme par exemple, et de manière non exhaustive:

- **Les données incomplètes sur le Web**. Le Web ne garantit pas le caractère fini et suffisant de l'information présente. Même si un expert travaille avec le concept t1 au quotidien, il se peut que le Web soit privé de cette information et que le lien entre l'expert et ce concept ne soit pas visible.
- **Les différents niveaux de précision**. Si l'auteur du problème n'est pas expert dans le domaine (ce qui est souvent le cas), il est susceptible d'utiliser un langage parfois imprécis, voire inapproprié. Si les concepts sont associés à une mauvaise description, il existe un risque de passer à côté de certains sujets d'intérêt, écartant ainsi des communautés d'experts potentiellement compétents.
- **Langages différents.** Le manque d'expertise ou de précision de l'auteur de la description du problème n'est pas la seule source d'une faible richesse de concepts associés au problème. Elle peut aussi provenir du fait que cette personne utilise un vocabulaire spécifique à une communauté de pratique et n'a pas l'habitude de se servir des notions similaires, utilisées spécifiquement pour identifier la même problématique dans d'autres communautés.

Il peut donc être utile d'élargir l'ensemble des concepts initiaux pour limiter les conséquences de leur éventuelle pauvreté. Dans ce chapitre nous nous intéresserons donc à l'élargissement d'un ensemble de concepts, appelé ensemble de concepts initiaux ECI. Comme présenté dans la Figure 34, cet ensemble peut représenter des concepts liés à différents types d'entités telles que : traces utilisateurs, problèmes, experts. Le résultat de cet élargissement devrait être un nouvel ensemble élargi de concepts EEC. Les éléments d'EEC sont potentiellement liés aux autres entités (e.g. traces, problèmes,

utilisateurs) qui peuvent être pertinentes par rapport aux entités initiales dans différentes situations.

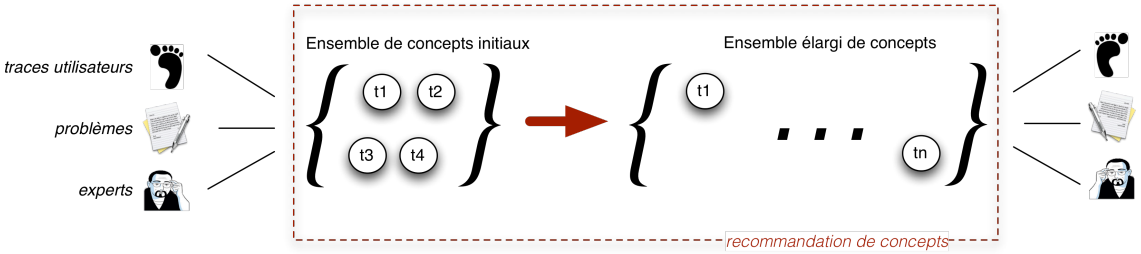


Figure 34 Enrichissement de l'ensemble de concepts initiaux

Cet élargissement n'est pas en soit une nouvelle idée. Dans le cadre de la recherche d'information, des approches d'élargissement et d'enrichissement de requêtes ont été proposées pour faire face à des problèmes similaires ayant pour but d'aider l'utilisateur à effectuer de meilleures recherches. Nous présenterons plus loin quelques exemples de ces approches. Cette problématique, même si elle n'est pas nouvelle, émerge dans un nouveau contexte : celui des scénarios liés à l'innovation. Celui-ci impose de nouveaux défis comme par exemple le besoin d'une notion de pertinence particulière, favorisant l'inclusion des connaissances non conventionnelles dans le processus de la résolution de problèmes. Nous détaillons ces défis ci-dessous.

5.1.2. Qualités désirées d'une approche de découverte de concepts

Nous avons déjà décrit quelques spécificités des scénarios liés à l'innovation qui contraignent en quelque sorte nos travaux. Décrites dans §3.4.3.1, ces spécificités se reflètent également dans la recommandation de concepts. Quelle que soit la situation concrète dans le cadre de laquelle une recommandation de concepts est effectuée, si celle-ci est liée à la résolution de problèmes dans un scénario d'innovation ouverte, il est souhaitable que les concepts recommandés répondent aux caractéristiques suivantes :

- **Pertinence dans le sens classique**. Comme pour chaque système de recommandation, la pertinence des concepts proposés par rapport aux concepts initiaux est indispensable. Notre notion de pertinence ne diffère pas de celle adoptée par des approches classiques de la recherche d'information, comme nous le verrons par la suite dans les parties dédiées à l'évaluation de notre approche.
- **Caractère inattendu**. Pour les scénarios liés à l'innovation, il est très important de savoir attirer des idées et engager des connaissances provenant de communautés de pratique qui ne sont habituellement pas concernées par un problème. Pour ce faire il est nécessaire d'identifier des concepts à la fois inattendus et pertinents qui pourraient être au cœur de l'indentification des communautés de pratique et d'experts susceptibles de faire émerger de telles solutions.

Ces deux caractères dépendent fortement de l'appréciation de l'utilisateur et seront, pour des mêmes concepts, jugés différemment selon les utilisateurs et les circonstances. Ce jugement subjectif aurait sans doute un impact sur l'utilité qu'un concept peut avoir pour un utilisateur dans une situation particulière, car ce sont les concepts jugés à la fois pertinents et inattendus qui auront la plus grande valeur pour l'utilisateur. Les concepts

privés de caractère inattendu seraient d'une utilité limitée, car malgré leur éventuelle pertinence, ils peuvent paraître évidents et redondants. Similairement le manque de pertinence rend même des concepts inattendus peu utiles.

Bien évidemment, les concepts perçus comme pertinents et inattendus sont les plus utiles pour l'utilisateur car ils constituent pour lui une découverte. Une telle perception peut être due au caractère latéral de la pertinence de ces concepts, qui ne viennent pas à l'esprit par un chemin de pensée direct et habituel mais uniquement par l'exploration de pistes créatives, et indirectes. Les concepts répondant à ces deux caractéristiques sont les plus susceptibles de conduire à la sérendipité, si appréciée en innovation.

5.2. Etat de l'art

5.2.1. Types d'approches

Dans la littérature nous trouvons une multitude d'approches pertinentes pour la découverte de concepts. Certaines de ces approches analysent les mots et non les concepts, mais utilisent des méthodes transposables au traitement de ceux-ci. Nous les considérons donc également pertinentes. Ces différentes approches ont été développées pour être utilisables dans différentes situations telles que : la recommandation de tags ou de mots-clés pour la recherche Web.. Ces techniques ont également été élaborées en s'appuyant sur différentes sources pour découvrir ces concepts (ou mots-clés). Nous distinguons particulièrement trois types d'approches selon le type de notion de proximité sur lesquelles s'appuient:

- **Proximité par le contenu**. Cette notion de proximité nous permet de considérer comme étant proches deux concepts qui se retrouvent régulièrement dans les mêmes unités de contenu, e.g. ressources Web. C'est par la fréquence de co-usage que l'on conclut à une « complicité » des notions et donc à une proximité de deux concepts.
- **Proximité sociale**. Cette notion de proximité est dérivée de la cooccurrence dans des profils utilisateur. En effet si deux concepts ou termes font souvent l'objet de l'intérêt des mêmes utilisateurs il est raisonnable de les considérer proches dans un sens.
- **Proximité sémantique**. Deux concepts ou termes sont considérés comme proches par cette notion s'ils sont connectés d'une manière précise dans un thesaurus ou dans une base de connaissances. Cette notion exprime une proximité par le sens qui n'est pas forcement manifesté dans le co-usage.

Deux types de méthodes sont notamment utilisés pour calculer ces mesures de proximité :

- **Approches basées sur la cooccurrence**. Ces approches exploitent la cooccurrence des concepts ou termes dans différents contextes afin de calculer une notion de proximité. Ils servent souvent à calculer la proximité par le contenu ou la proximité sociale.
- **Approches basées sur l'exploitation du graphe**. Ce type d'approche exploite les distances dans les structures de graphe pour calculer une notion de proximité, le plus souvent une proximité sémantique.

Nous présenterons par la suite des approches choisies dans l'état de l'art, organisées en deux parties selon le type de méthode qu'ils utilisent.

5.2.2. Approches basées sur la cooccurrence

La plupart des approches basées sur la cooccurrence mesurent différentes caractéristiques de distribution des mots dans de larges corpus textuels. Tout texte, tel qu'un document, est d'abord représenté en tant que vecteur de mots ou groupe de mots. Il est ensuite possible d'y appliquer des mesures de similarité de vecteurs standards, telle que la similarité de Jaccard [Jaccard, 1901], Dice [Salton, 1983] ou la similarité cosinus [Salton, 1983]. Des mesures plus complexes utilisant des vecteurs ont également été développées dans la communauté de recherche d'information. Ainsi, la mesure TF-IDF [Salton, 1983] tient compte de l'importance ou de l'absence de certains mots dans les textes soumis à examen. Elle est utilisée par exemple dans [Matos et al., 2010] pour calculer une notion de proximité à partir de la cooccurrence de mots dans des articles de recherche. Similairement une mesure basée sur la cooccurrence dans les tags est proposée dans [Sigurbjörnsson & Zwol, 2008].

C'est une pratique très courante de s'appuyer sur la cooccurrence de mots dans des résultats de recherche Web i.e. dans des documents du Web indexés par des moteurs de recherche. Un exemple très connu d'une telle mesure est *Normalized Google Distance (NGD)* présenté dans [Cilibrasi & Vitanyi, 2007]. Cette mesure présente un rapport entre le nombre de documents trouvés par Google dans lesquels deux mots coïncidents et le nombre d'occurrences individuelles de ces mots. Une évaluation de la même mesure utilisant différents moteurs de recherches a été présentée dans [Gracia & Mepa, 2008]. Selon l'avis des utilisateurs, le moteur *Exalead* serait la meilleure source pour le calcul des proximités de mots, suivi de près par Yahoo !, Altavista, Ask et Google. La mesure NGD s'inspire du principe de *Pointwise Mutual Information*, proposé par Fano en 1961 [Fano, 1961], qui utilise les probabilités d'occurrence de mots, individuelles et mutuelles. Simplifié, NGD est aujourd'hui très utilisée grâce aux moteurs de recherches qui facilitent son calcul sans qu'il soit nécessaire de créer des indexes où de calculer des probabilités.

Le pseudo retour de pertinence (*pseudo relevance feedback*) est également une technique très répandue. Cette technique trouve des concepts proches à un ensemble de concepts de départ basé sur des corrélations dans les résultats d'une recherche initiale, faite en utilisant ces mêmes concepts de départ. Les résultats de recherche sont utilisés en tant qu'approximation de l'avis des utilisateurs (*en. relevance feedback*). Cette approche considère en effet que des mots trouvés suite à une recherche qui coïncideraient avec les mots-clés de la requête sont sémantiquement proches et sont donc utilisés pour enrichir cette requête. Si elle est utilisée avec succès dans la recherche Web, cette méthode se révèle inappropriée dès lors qu'on a affaire à une recherche d'experts précise (pour trouver par exemple des experts très spécialisés) [Macdonald & Ounis, 2007]. En effet, les profils d'experts contiennent souvent des sujets très divers, ce qui peut avoir un impact négatif sur la pertinence des résultats dans le cadre de l'enrichissement de requêtes.

Il est important de noter que la majorité de ces mesures, basées sur des vecteurs, peut être utilisée directement sur différents ensembles d'entités vectorielles, même si celles-ci ne correspondent pas à des textes.

5.2.3. Approches basées sur l'exploitation du graphe

Certaines approches exploitent la structure des graphes pour calculer une mesure de proximité. Ces graphes représentent souvent des relations d'hyponymie et de méronymie ou bien d'autres types de relations basées sur le sens ou la catégorisation des concepts – nœuds dans le graphe. En mesurant la distance que l'on peut observer entre deux concepts dans le graphe, ces approches concluent qu'il existe un degré de proximité (en fonction de la nature des relations présentes dans le graphe) de concepts.

Différentes mesures de distance peuvent s'appliquer à l'analyse de chemins dans le graphe. La distance de Rada [Rada et al., 1989] est la mesure la plus simple. Elle prend comme distance la longueur du plus court chemin entre deux concepts dans une taxonomie. Une version normalisée de la distance de Rada (divisée par la distance maximale dans le graphe) est proposée dans [Leacock & Chodorow, 1998]. Resnik [Resnik, 1995] propose d'utiliser le contenu informationnel de la catégorie spécifique la plus commune pour exprimer la proximité de deux concepts. Il s'appuie notamment sur l'hypothèse suivante : plus les deux concepts partagent d'information, plus ils sont proches. Une autre possibilité, dans le même esprit, est de tenir compte du contenu informationnel des concepts eux-mêmes ainsi que celui de leur catégorie spécifique la plus commune et d'en estimer la différence [Jiang & Conrath, 1997]. Wu et Palmer [Wu & Palmer, 1994] utilisent la longueur du plus court chemin entre des concepts et leur dénominateur commun le plus spécifique pour estimer leur proximité. Une mesure qui tient compte de tous les chemins entre deux concepts et leur poids respectif est présentée dans [Sussna, 1993].

Différents graphes ont été utilisés pour calculer les mesures de proximité. Le graphe du Wordnet[98] est par exemple utilisé dans [Burton-Jones et al., 2003]. L'approche présentée dans [Safar et al., 2004] s'appuie sur un treillis de Gallois alors que la taxonomie des catégories de sites Web ODP[99] est utilisée dans [Ziegler et al., 2006]. Similairement, certaines approches comme celle présentée dans [Strube & Ponzetto, 2006] s'appuient sur le graphe de catégories de Wikipedia. Zetch et Gurevych [Zetch & Gurevych, 2007] utilisent également ce dernier pour calculer la similarité de mots en utilisant des mesures de proximité dans le graphe basées sur (1) la longueur des chemins entre concepts et (2) lecontenu informationnel des catégories. Des mesures de ce dernier type montrent des performances supérieures et confirment que le graphe de catégories peut être utilisé à la place de WordNet, en fournissant une couverture plus large de concepts et de langues [Zetch & Gurevych, 2007].

Le graphe de liens entre des pages Wikipedia a également inspiré un bon nombre d'approches. Par exemple, dans [Olivier & Senellart, 2007] nous trouvons une mesure de proximité dans le graphe de liens basée sur les chaines de Markov. Cette mesure montre de meilleurs résultats en termes de pertinence des articles suggérés comme proches en comparaison avec la mesure basée sur cosinus et une mesure inspirée par PageRank™[Brin & Page, 1998]. Cette dernière, comme la mesure HITS [Kleinberg, 1999], même si elle fait partie des mesures basées sur le graphe, vise principalement à établir une notion d'importance des connexions d'un nœud dans le graphe et n'est pas orientée vers le calcul des proximités. Son focus sur les nœuds les plus proéminents dans le graphe la rend notamment inadaptée à la recherche de concepts inattendus.

[98] http://wordnet.princeton.edu/

[99] http://www.dmoz.org/

5.2.3.1. Approches basées sur les données liées

Rares sont les approches qui utilisent le graphe de données liées pour calculer la proximité de concepts ; quoiqu'il soit possible de trouver des approches dans le domaine de la recommandation de musique et des vidéos. Un système de recommandation de musiciens similaires, DBRec, a été proposé dans [Passant, 2010]. Ce système exploite le graphe de DBPédia et quelques autres EDS liés à la musique, en se focalisant uniquement sur des liens établis à travers des propriétés définies pour les concepts du type Artist. Cette approche repose sur l'hypothèse que des concepts pour lesquels la similarité est calculée ont un type bien défini permettant d'identifier des propriétés utiles, et il est donc d'une applicabilité limitée dans des cas d'hétérogénéité de concepts.

Parallèlement, un système pour la recherche exploratoire (en. *exploratory search*) de vidéos utilisant DBPedia a été proposé [Waitelonis & Sack, 2009]. L'information structurée de DBpédia permet d'organiser des vidéos en catégories - facettes de la recherche exploratoire - afin de permettre à l'utilisateur de mieux naviguer dans un espace de vidéos. Si le résultat n'est pas une mesure de similarité dans le sens propre du terme, l'objectif de cette approche est très semblable à celui de ces mesures, si ce n'est le mode d'interaction qui est différent.

Une approche hybride, essayant de transposer des mesures de similarité utilisées sur des textes au traitement des graphes de données liées, a été proposée par Damljanovic et al. [Damljanovic et al., 2011]. Cette approche s'appuie notamment sur la méthode Random Indexing (RI) [Sahlgren, 2005], utilisée pour calculer de manière efficace les similarités de mots en se basant sur de larges corpus de texte, rendant des vecteurs représentant des mots très larges. Elle dispose de meilleures performances computationnelles que les méthodes vectorielles classiques, telles que *Latent Semantic Analysis*[Deerwester et al., 1990], une méthode d'application des métriques de proximité vectorielle aux grands corpus, sans sacrifier la qualité de ses résultats [Karlgren & Sahlgren, 2001][Cohen, 2008]. Pour appliquer RI dans le contexte de graphes de données liées, il est d'abord nécessaire de représenter les informations associées aux ressources (e.g. concepts) en format similaire à celui utilisé pour représenter des documents textuels dans le cadre de RI. Cela rend cette méthode applicable sur de grands EDS tels que DBpedia. Cette approche fut développée en parallèle de la nôtre et fait partie des approches avec lesquelles nous comparons notre méthode dans la partie Evaluation.

5.2.4. Ce qui motive le développement d'une approche basée sur de données liées

Les mesures basées sur la cooccurrence ont tendance à expliquer la pertinence à travers le co-usage et négligent les relations entre concepts ou termes établies à travers leur sens. Il est donc peu réaliste de s'attendre à que ces mesures aboutissent à produire des recommandations inattendues, définies comme désirables pour notre cas d'utilisation dans la partie §5.1.2.

Il est donc raisonnable d'envisager une approche basée sur le graphe, visant à aider la découverte de concepts liés par leur sens. Notre intuition pour la construction d'une telle approche était que la suggestion de concepts pertinents et liés par leur sens serait susceptible d'atteindre les objectifs visés en §5.1.2. En partant de l'hypothèse que nous trouverions des concepts pertinents, le fait que leur lien avec les concepts initiaux soit

sémantique et non pas basé sur leur co-usage habituel laisse un espoir qu'ils soient aussi inattendus, car ils sont potentiellement peu utilisés dans des contextes similaires.

Les méthodes basées sur le graphe que nous avons présentées s'appuient souvent sur les taxonomies et les ontologies de domaines. Ces sources ne couvrent le plus souvent que des thèmes très spécifiques qui intéressent des communautés de pratiques particulières. Pour qu'elle puisse servir à des scénarios liés à l'innovation, une telle approche devrait permettre la découverte de concepts à travers des communautés de pratique et être utile pour des problèmes d'innovation, quel que soit le champ de compétence visé. Cette nécessité nous a motivé à poursuivre une approche adaptée au graphe de données liées. Seul DBpedia contient plus de 3.5 million de concepts – sujets de connaissance humaine ce qui dépasse largement la taille du Wordnet (206941 paires mot-sens)[100] et ODP (~ 1 million de catégories)[101]. Croissante et interconnectée avec d'autres jeux de données, DBpédia s'impose comme un choix prudent. Malgré ces avantages DBpedia n'est qu'une source de données, et nous avons donc également cherché une généralité dans la conception de notre approche pour permettre son application sur les concepts du graphe de données, quel que soit l'EDS dont ils font partie. Nous avons donc construit notre approche autour d'une structure de la connaissance couramment utilisée dans la représentation des concepts, notamment dans DBpedia et Freebase.

5.3. Conception d'une approche basée sur des données liées

Le cœur de notre approche s'articule autour de l'idée que différents types de liens dans le graphe de données liées ont une nature différente et méritent donc un traitement différent dans le calcul des similarités. Cette approche est donc motivée par l'idée de tirer profit de la richesse du graphe de données liées, qui n'est pas un simple graphe de nœud et liens, mais un graphe typé.

Afin de construire une approche adaptée à la nature du graphe, et notamment à ses parties contenant des concepts, nous avons étudié le graphe du DBpedia[102] pour mieux comprendre sa composition. Suite à cette étude, il apparait comme évident que les propriétés caractérisant les concepts pouvaient être réparties en deux ensembles :

- **Propriétés hiérarchiques - qui construisent une catégorisation de concepts.** Ces propriétés permettent de relier un concept à une catégorie, ou bien de relier deux catégories entre elles. Elles permettent de représenter des connaissances taxonomiques.
- **Propriétés transversales - qui établissent des liens entre des concepts sans vocation de créer une hiérarchie.** Ces propriétés servent plutôt à représenter des connaissances factuelles, sans construire une taxonomie explicite.

[100] http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html

[101] http://www.dmoz.org/

[102] Dans le cadre de nos travaux futurs nous prevoyons d'étudier la transposabiltié de cette analyse à la structure du Freebase qui est très similaire.

Dans la Figure 35 nous présentons un exemple de graphe de données liées qui nous permettra d'illustrer la différence entre ces deux groupes de propriétés. Dans ce graphe nous avons 3 concepts t1, t2 et t3. Les propriétés de type *l3* et *l4* sont les propriétés appartenant au premier groupe. Une telle propriété connecte t1 et t2 avec leur catégorie C1 ainsi que t3 avec sa catégorie C2. Dans des graphes réels, dct:subject est le plus souvent utilisé pour exprimer la sémantique du *l3*. Les liens de ce type sont orientés du concept vers la catégorie. La catégorie C2 est ensuite déclarée comme étant la sous-catégorie d'une autre catégorie C2. Dans notre exemple, ce lien est nommé *l4*, et dans les graphes réels la propriété skos:broader est utilisés pour marquer ce type de lien dans la direction de C1 vers C2, et skos:narrower dans le sens inverse. Quelle que soit la manière de l'exprimer, pour des raisons de simplicité nous nous référons à ce type de lien comme étant dirigé de sous-catégories vers des supercatégories.

Les propriétés transversales sont illustrées par les liens *l1* et *l2*. *L1* établit un lien direct entre t2 et t3, alors que l2 est une propriété exprimant une valeur primitive (e.g. numérique, textuelle). Comme t1 et t2 partagent une même valeur de la même propriété *l2*, il est possible de considérer qu'entre eux il existe un lien implicite. Ces liens implicites sont rarement utilisés dans la littérature et largement négligés. Notre étude que nous détaillerons plus tard, a montré que dans des cas liés à l'innovation ils ont une importance comparable aux liens directs. Nous les considérons donc comme étant des liens de même niveau que des liens directs.

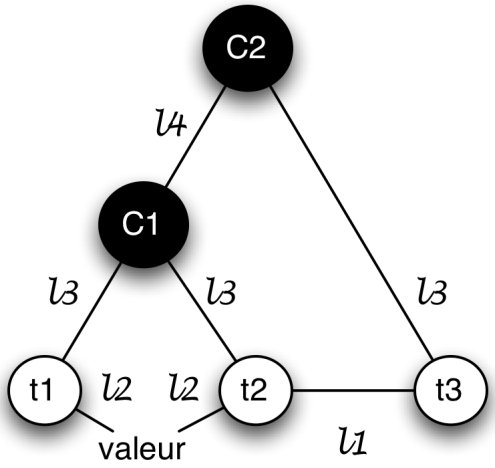


Figure 35 Structure d'un exemple de graphe de données liées

Nous avons conçu et évalué des mesures de proximité différentes pour estimer la proximité de concepts à travers les liens de ces différents groupes. Nous présentons tout d'abord une approche générique, indépendante du type de lien et de la mesure utilisée, pour ensuite donner les détails des différentes mesures et de leur utilisation.

5.3.1. Approche générale

Notre approche a pour objectif de récolter, dans un graphe de données liées, des concepts liés à un ensemble de concepts initiaux (ECI) et de les classer par pertinence. S'agissant d'un ensemble de concepts, notre métrique générale exprime la proximité d'un concept candidat avec l'ensemble de concepts initiaux.

$$hyP(ECI,c) = \sum_{c_i \in ECI} hyP(c_i,c) \quad (8)$$

Présentée par la formule 8, notre mesure appelée hyProximity définit la proximité entre un concept *c* et l'ECI comme la somme des proximités de ce concept avec chaque concept c_i, membre d'ECI. L'intuition motivant le choix de la somme comme fonction agrégante est dans l'intention de privilégier des concepts qui sont proches de plusieurs concepts initiaux plutôt que ceux qui sont (même plus) proches d'un seul concept. Comme les choix précédents, ce choix est également fait en tenant compte da la nature du scénario d'innovation et de la résolution de problèmes. Une entité de départ (e.g. problème, utilisateur, trace) de la description dont nous avons tiré l'ECI, est souvent définie par l'ensemble de ses concepts et non par chacun considéré isolément. Il est donc nécessaire de tenir compte du fait que l'ECI ne représente cette entité que par synergie de ses membres. Par analogie, la proximité d'un concept à ECI reflète la synergie des proximités unitaires qu'il a avec tous les membres d'ECI.

Les fonctions de proximité unitaires dépendent de :

- La distance entre c_i et c, calculée par une métrique de distance choisie ;
- La fonction de pondération choisie.

Différentes alternatives de ces fonctions seront présentées par la suite.

Une mesure sans algorithme permettant de la calculer de manière efficace dans des conditions réelles serait peu utile. Un possible algorithme est donné par la suite dans un langage de pseudocode.

Algorithme 1

```
Initialise Résultats, distance actuelle d=0 ;
Tant que d++ < maxDistance
    Pour chaque concept c_i ∈ ECI
      Obtiens l'ensemble des concept C_d tels que
            distance(c_i,c_d)=d, ∀c_d∈C_d
      Pour chaque concept c_d∈C_d
            Obtiens valeur precedente(c_d) de Résultats
      Enregistre dans Résultats
                  < c_d, valeur precedente(c_d)+ hyP(c_i,c_d)>
      Classe Résultats par ordre décroissant de valeur
Fin.
```

L'algorithme commence par l'initialisation de la variable Résultats qui sert à garder les valeurs intermédiaires de hyP pour des concepts candidats. Le calcul est effectué pas à pas en incrémentant de 0 à un niveau prédéfini (maxDistance), le niveau de distance de concepts en cours d'observation. À chaque niveau, pour chaque concept initial, nous explorons des concepts dont la distance de celui-ci correspond à ce niveau. Cette distance peut être exprimée par différentes métriques que nous détaillerons par la suite. Les concepts sont généralement obtenus en effectuant des requêtes SPARQL sur le graphe de données liées. Pour chacun de ces concepts, nous calculons la valeur de hyP entre lui et le concept initial et nous rajoutons cette valeur à une valeur éventuellement déjà calculée pour ces concepts lors du traitement des autres concepts initiaux. C'est de cette manière que nous assurons que la valeur hyP entre un concept et l'ECI correspond à la somme des valeurs unitaires que ces concepts génèrent avec chaque membre d'ECI. Toutes ces valeurs cumulatives permettent *in fine* de classer les concepts retrouvés.

L'algorithme est conçu pour permettre une souplesse d'exécution particulièrement importante dans toute manipulation du graphe de données liées. D'une manière générale, chaque opération d'interrogation des EDS, sources de concepts, est coûteuse

en terme de temps pour l'exécution de requêtes SPARQL sur des graphes larges, d'autant plus que ces requêtes utilisent généralement le protocole HTTP entre différents serveurs. Dans de telles conditions, la possibilité d'arrêter l'exécution de l'algorithme au moment où un nombre suffisant de concepts a été récolté peut être nécessaire. À condition que les fonctions de proximité unitaires $hyP(c_i,c_d)$ soient fortement décroissantes avec la croissance de distance, il est raisonnable de s'attendre à ce que les résultats obtenus avant un certain niveau de distance maxDistance aient des valeurs de hyP plutôt supérieures à celles des concepts potentiellement trouvables après maxDistance. Dans la version de l'algorithme présentée ici, le niveau maxDistance est fixé par avance, mais il est également envisageable de le définir en fonction du nombre de concepts collectés dans Résultats.

5.3.1.1. Traitement des liens hiérarchiques

La structure hiérarchique des catégories dans des graphes de données liées, par sa nature, rappelle les taxonomies. Cela amène naturellement à penser que les métriques de distances utilisées pour les taxonomies, présentées dans §5.2.3, pourront s'appliquer directement. Néanmoins, en étudiant les données existantes, il devient clair que certaines différences existent entre leurs organisations de connaissance qui rendent nécessaire la conception de méthodes adaptées à la donnée liée. Dans la structure de DBPedia nous trouvons par exemple quelque cycles, i.e. une catégorie A, connectée à une catégorie B à travers plusieurs liens de type skos :broader où B est également connecté à A avec le même type de lien. Ce phénomène, peu imaginable dans des taxonomies classiques, est même rendu légitime dans DBPedia par l'interprétation de la relation skos : broader qui n'est pas transitive. Si d'un point de vue A est plus large que B, d'un autre point du vue, B pourrait être plus large qu'A. Si cette question de cycles peut paraître plus évidente dans DBPedia, cette question ne lui est pas exclusivement réservée. Il s'agit d'un principe fondateur qui est au cœur de la conception du graphe de données liées, et même du Web : le principe que toute information présente, tout triplet, corresponde à un certain point de vue avec lequel différents acteurs peuvent être ou ne pas être d'accord. Il est important de noter qu'à cause (ou grâce à) ce principe, l'existence d'un cycle ne casse pas la nature hiérarchique de la structure créée par skos : broader, qui reste hiérarchique pour des points de vue qui n'acceptent qu'un sous-ensemble de liens skos : broader.

Il est donc clair qu'une métrique prenant la structure créée par des liens hiérarchiques comme étant une hiérarchie standard (sans cycles) ne pourrait pas s'appliquer directement sur le graphe de données liées. La structure étant en effet un graphe et non pas une arborescence, les méthodes classiques de traitement des distances dans les taxonomies ne peuvent pas s'appliquer. Les méthodes de détection des boucles pourront en effet permettre d'isoler la source du problème, mais selon l'hypothèse que tout point de vue est légitime, il serait injuste de procéder à une élimination des boucles afin de rendre ces métriques applicables.

Le deuxième élément de à considérer dans la construction de notre approche est la profondeur de la hiérarchie. Les analyses de DBPedia [Lehamnn et al., 2007] ont montré que les ressources y sont densément connectées, créant dont une structure peu profonde. Cette densité affecte aussi bien les liens hiérarchiques que transversaux. La Figure 36 montre le nombre de concepts, trouvable en moyenne à différentes distances (longueur du plus court chemin) à partir d'un concept de départ. L'histogramme montre

que la quasi-totalité des concepts peut être trouvée à une distance inferieure à 10. Le nombre de concepts à la distance 4 s'approche de 5000 et à distance 3 de 2000. Ces quantités de concepts excèdent largement les besoins de notre cas d'utilisation. La structure dans laquelle nous trouverons un nombre de concepts pertinents à proposer est donc assez peu profonde.

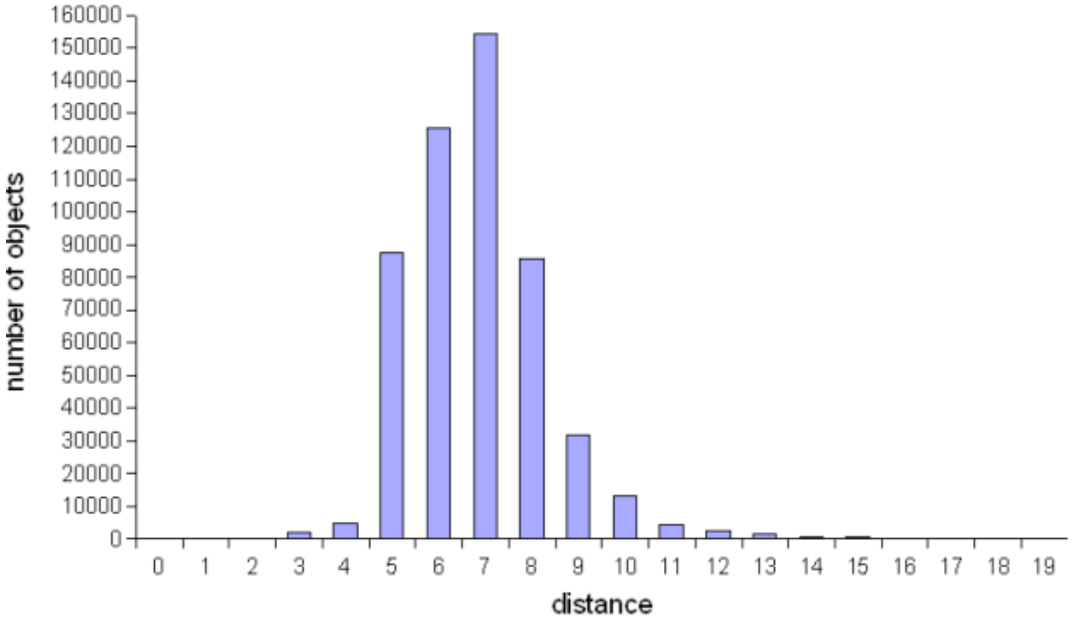


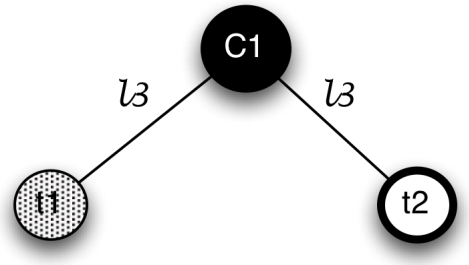
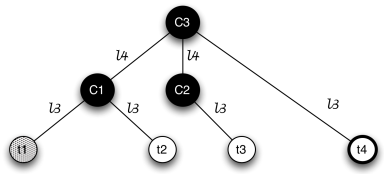
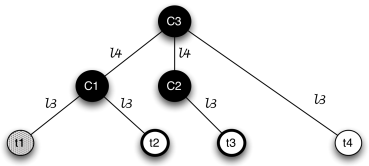
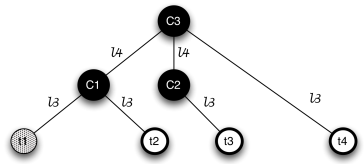
Figure 36 Distribution des distances dans DBPedia repris de [Lehamnn et al., 2007]

$$hyP(c_1,c_2) = \frac{p(c_1,c_2)}{d(c_1,c_2)} \quad (9)$$

Notre conception des métriques de distance reflète ces deux spécificités du graphe : l'existence des boucles et le manque de profondeur. La quantification de la proximité de deux concepts c_1, et c_2, établie à travers des liens hiérarchiques est faite à l'aide d'une fonction de distance et d'une fonction de pondération. La formule 9 montre leur rapport. La valeur de la fonction hyP unitaire entre deux concepts est proportionnelle à la valeur de la fonction de pondération et inversement proportionnelle à leur distance. La fonction de pondération peut dépendre de la distance ou bien d'autres paramètres comme par exemple l'importance accordée à la spécificité des catégories qui connectent les deux concepts. Définie si généralement, la fonction hyP aurait la même forme quand les distances sont établies à travers des liens transversaux. Pour mieux distinguer, nous allons dénoter $hyP^{hyérarchique}$ et $hyP^{transversalle}$.

En s'inspirant des spécificités du graphe de données liées, nous avons conçu trois fonctions de distance différentes (d1, d2, et d3). Dans l'esprit d'une approche orientée graphe, la définition de ces fonctions est donnée de manière graphique, en précisant les patterns de graphes utilisés pour caractériser une distance particulière. Elles sont présentées dans le Tableau 5.

Tableau 5 Fonctions de distance hiérarchiques

$d1(c_1, c_1)$	$d1(c_1, c_1)$	$d1(c_1, c_1)$
Nous considérons qu'un concept est à distance 0 de lui-même selon toutes les fonctions de distance.		
 *Les concepts à distance 1 sont trouvés d'une même manière pour toutes les fonctions. Si deux concepts appartiennent à une même catégorie, alors leurs distance mesurée par des fonctions d1, d2 et d3 est égale 1. La figure ci-dessus montre un concept t2, étant à distance 1 du concept initial t1.*		
 La fonction d1 considère comme étant à distance 2 du concept initial *t1* les concepts attachés **directement** (par un lien du type *l3*) à des catégories de deuxième degré auxquelles appartient *t1*. Les catégories de deuxième degré sont connectées par un lien de type l4 à une catégorie à laquelle appartient directement le concept *t1* (par lien du type *l3*). Dans notre exemple de graphe ci-dessus, seul le concept *t4* est considéré comme étant à distance 2.	 La fonction d2 considère comme étant à distance 2 du concept initial *t1* les concepts attachés à toute sous-catégorie (par un lien du type *l3*) des catégories de deuxième degré auxquelles appartient *t1*. Dans notre exemple de graphe ci-dessus, les concepts *t2* et *t3* sont considérés comme étant à distance 2.	 La fonction d3 considère comme étant à distance 2 du concept initial *t1* les concepts attachés directement (par un lien du type *l3*) à toutes les catégories de premier et de deuxième degré auxquelles appartient *t1*. Dans notre exemple de graphe ci-dessus, les concepts *t2*, *t3* et *t4* sont considérés comme étant à distance 2.

Par analogie il est possible de dériver des patterns de graphes à utiliser pour des niveaux supérieurs à 2, en respectant les mêmes principes que ceux de ces trois fonctions, mais en partant des catégories de degré n (donc connectés au concept de départ par un chemin comprenant un lien de type *l3* et n-1 liens de type *l4*). Les

concepts qui n'auraient été trouvés par aucun des patterns de sélection auraient de fait une valeur de distance infinie. Le nombre de concepts trouvables aux distances 1 et 2 étant souvent dans les milliers, dans nos expériences nous n'avons pas été amenés à utiliser des niveaux supérieurs à 2.

Par leur conception, en considérant des parties limitées du graphe, nos fonctions ne risquent pas de tomber dans des cycles et assurent un calcul simple, à l'aide des requêtes SPARQL peu complexes (donc exécutables assez rapidement) et s'accordent bien avec notre approche algorithmique permettant l'arrêt du processus à tout moment.

Nous avons également conçu deux fonctions de pondération avec le but de traiter distinctement différents concepts, en fonction de leur position par rapport au concept initial.

La première fonction p1 décroit exponentiellement l'importance des concepts trouvés plus loin par rapport à ceux trouvés plus près du concept initial. Cette fonction est donnée par la formule 10 (dans nos expériences nous utilisons 0.3 comme la valeur de λ, mais ce paramètre peut être configuré pour ajuster la « vitesse » de décroisement).

$$p1(c_1,c_2) = e^{-\lambda^{d(c_1,c_2)}} \quad (10)$$

Une approche plus classique consisterait à utiliser le contenu informationnel de la catégorie qui est commune aux deux concepts liés. Une fonction inspirée par cette idée est donnée par la formule 11.

$$p2(c_1,c_2) = -\log(P(categorie_commune(c_1,c_2)) \bullet d(c_1,c_2) \quad (11)$$

Cette fonction est proportionnelle au logarithme de la probabilité P (catégorie_commune(c_1,c_2)) de retrouver la catégorie commune qui relie les deux concepts c_1 et c_2 en partant d'un concept aléatoire et en suivant les liens hiérarchiques du bas vers le haut. La valeur de ce logarithme est multipliée par la distance afin de neutraliser l'effet de division par la distance dans la formule hyP. Cette neutralisation est nécessaire car la fonction du contenu informationnel est déjà décroissante par rapport à la distance et une nouvelle prise en compte de la distance serait exagérée.

Nous présentons ci-dessous une comparaison de la performance de ces distances.

5.3.1.1.1. Comparaison des versions alternatives de l'approche

Afin de comparer nos différentes fonctions de distance nous avons conduit une évaluation en utilisant comme repère des anciens problèmes et leurs solutions reçus par la plate-forme Hypios. Celle-ci repose sur l'hypothèse qu'une bonne fonction devrait, au moins, permettre pour un ensemble de concepts extraits d'un problème d'innovation, de trouver des concepts qui apparaissent dans les solutions du problème. Une bonne fonction de distance considèrera comme proches des concepts susceptibles d'être associés à des axes de solution directes ou latérales.

Nous avons donc construit un jeu de données de vérification qui provient de l'activité de la plateforme Hypios. D'un côté nous avons récupéré les textes de problèmes, soumis à la plateforme depuis sa conception, tandis que de l'autre côté, nous avons récupéré les solutions proposées pour ces problèmes. Ensuite, une extraction de concepts (via le service Zemanta) a été effectuée depuis les textes des problèmes et les solutions. Cette opération a abouti à 27 paires de jeux de concepts <concepts du problème, concepts des solutions du problème>. Nous avons considéré que des concepts apparaissant dans les

solutions d'un problème sont ceux qu'une bonne mesure devrait trouver proches des concepts du problème. Ce jeu de données de vérification sera utilisé ensuite à plusieurs reprises pour évaluer différentes variations de notre approche et les comparer aux approches concurrentes.

Pour mesurer la performance, nous utilisons trois mesures, reprises de la littérature de recherche d'information : précision, rappel et la mesure F.

$$précision_n = \frac{|suggestions_n \cap concepts_de_solutions|}{n} \quad (12)$$

Par précision, nous entendons le rapport entre le nombre des suggestions d'un système considérées proches d'un ensemble de concepts représentant un problème d'innovation qui figurent aussi dans les solutions proposées à ce problème et le nombre total de suggestions proposés par ce système. Nos approches étant capables de proposer des milliers de concepts classés en ordre décroissant de proximité, la question du choix du nombre de concepts pris en compte dans le calcul de la précision s'impose naturellement. Nous mesurons donc la précision à différents niveaux, en prenant les *n* concepts classés comme les plus proches par un système ou une fonction de proximité. La précision permet d'évaluer le rapport des « bonnes » suggestions dans la masse de suggestions proposées par une approche/fonction/système.

De la même façon, nous définissons une mesure de rappel dans les n premiers concepts. La fonction rappel représente le rapport entre le nombre de « bons » concepts figurant dans les suggestions et le nombre de bons concepts trouvables pour le problème en question.

$$rappel_n = \frac{|suggestions_n \cap concepts_de_solutions|}{|concepts_de_solutions|} \quad (13)$$

Nous utilisons également une mesure composite favorisant des fonctions qui ont un bon rapport de précision et de rappel, la mesure F définie par la formule 14.

$$F = 2 \bullet \frac{précision \bullet rappel}{précision + rappel} \quad (14)$$

Il est important de noter que l'interprétation de ces deux mesures dépend des concepts présents dans les solutions, qui ne correspondent pas forcement à la totalité des concepts pertinents pour la résolution d'un problème. Avec les données qui sont à notre disposition, nous ne sommes pas en mesure d'assurer que des concepts n'appartenant pas aux solutions connues ne sont pas susceptibles d'être associés à des solutions possibles, car la complétude des solutions soumises à un problème n'est pas garantie. Comme toute évaluation, la nôtre n'est qu'une approximation de la réalité, mais nous croyons qu'elle est une approximation raisonnable. Afin de mieux approcher la vraie performance des différentes mesures de proximité, nous allons compléter l'étude basée sur les anciens problèmes et leurs solutions avec des études utilisateur, qui à leur tour souffrent d'incomplétudes. C'est grâce à la synergie qui émerge de la combinaison des différentes techniques d'évaluation que nous nous permettons de tirer des conclusions finales dans notre comparaison des différentes fonctions et systèmes.

Notre expérience de comparaison repose sur le calcul des proximités, par l'algorithme présenté en §5.3.1 afin de produire une liste de concepts proches. Dans le calcul de hyP nous avons utilisé les trois fonctions de distance d1, d2 et d3, toutes avec la fonction de

pondération p1. Les valeurs de la précision et du rappel, obtenues en prenant un nombre différent de concepts marqués en tant que les plus proches (représenté en axe X) sont présentées aux figures Figure 37 et Figure 38.

Ces résultats montrent une performance supérieure de la fonction de distance d2. Une des raisons possibles d'une telle différence est que cette fonction écarte les concepts attachés directement aux catégories plus générales. En effet, en raison des possibles erreurs dans DBPedia, une explication plausible est que l'association directe des concepts avec des catégories plus générales représente une erreur potentielle, et en conséquence de quoi, elle n'auront pas d'effet positif.

Dans les deux cas, pour la précision et pour le rappel, le test T montre que la différence entre des valeurs est signifiante entre chaque deux paires de fonctions ($p<0.001$).

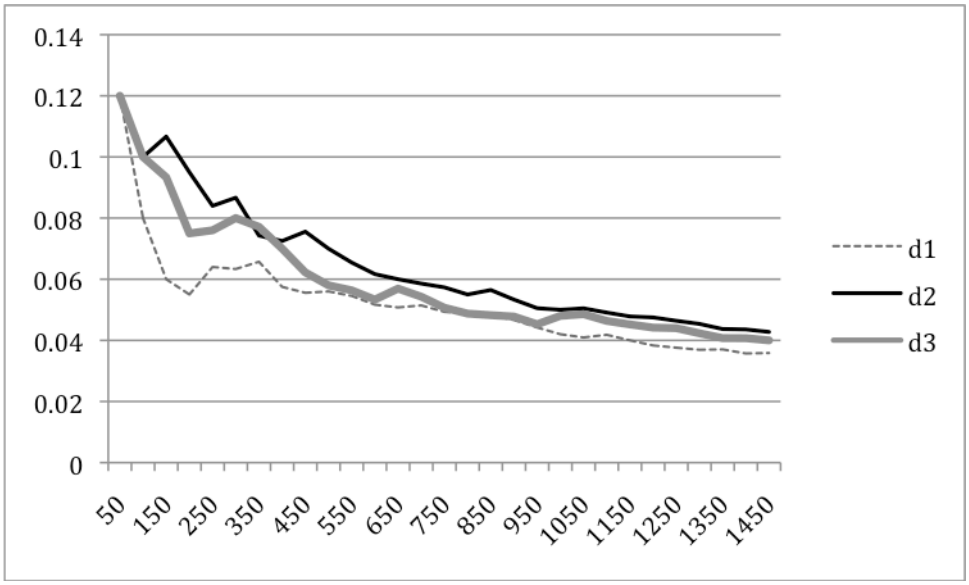


Figure 37 Précision pour des concepts obtenus par hyP^{hyérarchique} en utilisant différentes fonctions de distance et p1 en tant que fonction de pondération

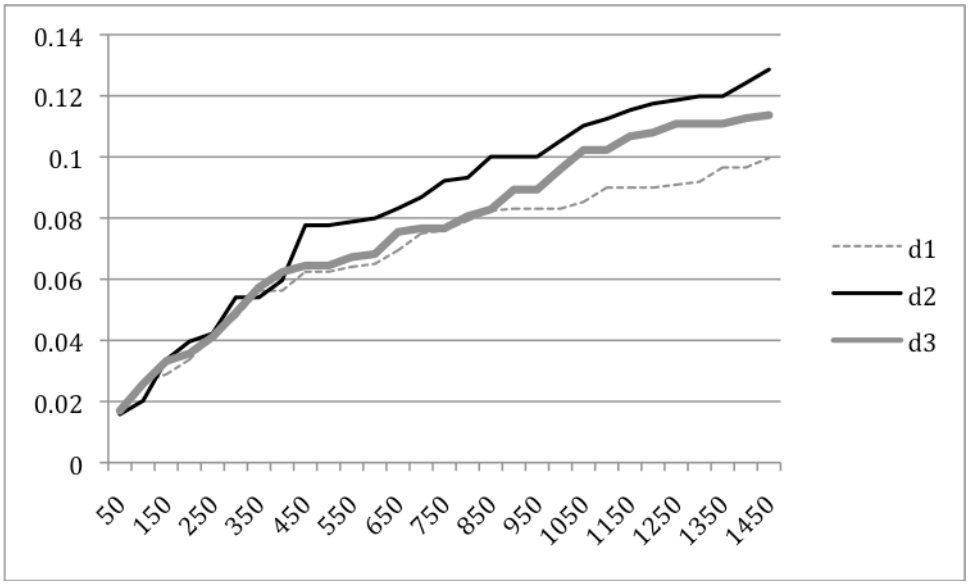


Figure 38 Rappel pour des concepts obtenus par hyP^{hyérarchique} en utilisant différentes fonctions de distance et p1 en tant que fonction de pondération

Si l'usage d'une même fonction de pondération nous a permis de comparer des fonctions de distance, il est également nécessaire de vérifier si une différence existe entre des fonctions de pondération. Nous avons donc essayé les deux fonctions de pondération sur la même fonction de distance – la fonction d2 qui a déjà montré ses qualités. Les valeurs de la précision et du rappel pour les deux fonctions p1 et p2 sont données aux figures Figure 39 et Figure 40. La fonction p2, basée sur le contenu informationnel des catégories à travers lesquelles deux concepts sont liés, montre les meilleurs résultats, surtout au niveau de la précision. Néanmoins, cette fonction demande des informations supplémentaires liées à la composition des catégories au moment de l'exécution de l'algorithme, ce qui rend les requêtes SPARQL plus complexes et plus couteuses en termes de temps d'exécution. La fonction p1, dépend uniquement de la distance – laquelle est constante dans le cadre de chaque étape de notre algorithme – ce qui la rend très facile à calculer et constitue donc un choix valide dans des situations réelles.

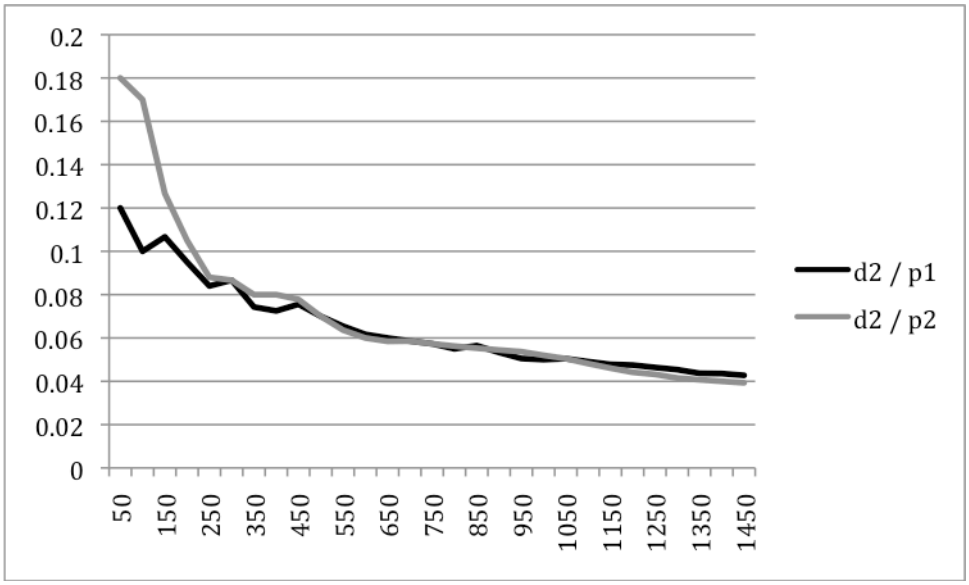


Figure 39 Précision obtenue en utilisant des fonctions de pondération p1 et p2, toujours avec la même fonction de distance d2

Il est important de noter que les valeurs faibles de la précision et du rappel sont dues à la nature de la tâche, qui est en effet assez difficile. Nous présenterons par la suite une comparaison de notre système avec d'autres méthodes issues de l'état de l'art, et nous pourrons observer que les niveaux que nous avons obtenus sont plutôt bons par rapport aux méthodes alternatives. C'est notre définition de concepts pertinents, qui est due au caractère restreint des données disponibles, qui influence ce niveau de précision et de rappel.

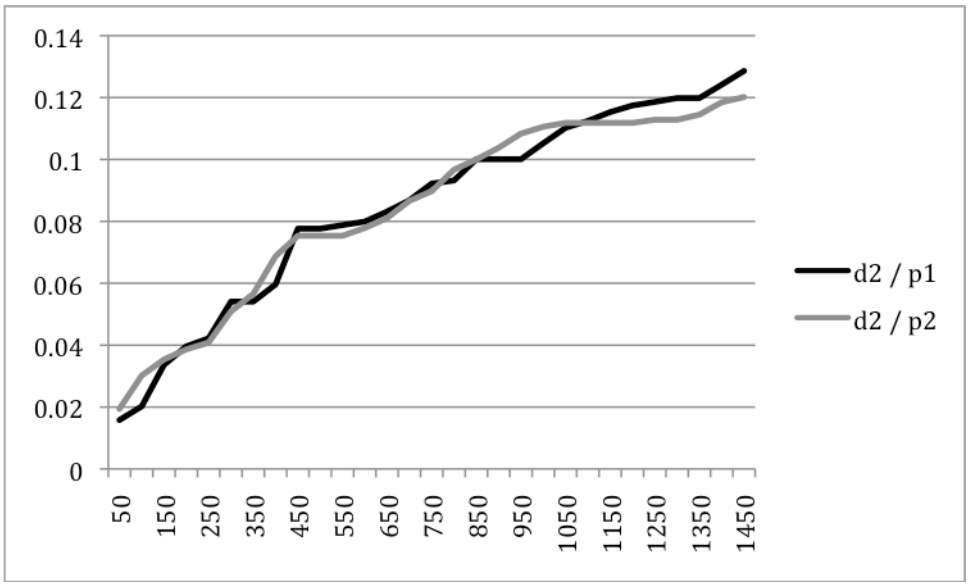


Figure 40 Rappel obtenu en utilisant des fonctions de pondération p1 et p2, toujours avec la même fonction de distance d2

5.3.1.2. Traitement des liens transversaux

Pour traiter des liens transversaux, directs (de type *l1*) ou indirects (de type *l2*), nous avons conçu une mesure de distance simple accompagnée d'une mesure de pondération inspirée par *p2*. Il est important de noter qu'un lien implicite établi à travers une valeur primitive est considéré comme un seul lien entre deux concepts.

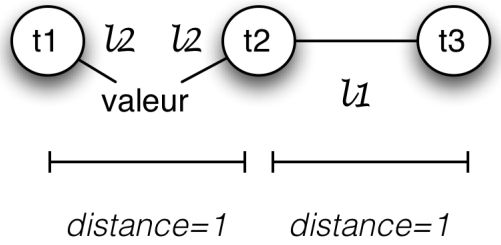


Figure 41 Distances transversales

Notre fonction de distance transversale, nommée d^t, prend comme étant à distance 1 tous les concepts liés par un lien direct ou indirect au concept de départ. Les concepts à distance 2 seront ceux qui sont liés de la même façon à un concept étant lié au concept de départ. Dans l'exemple du graphe de Figure 41, le concept *t2* est à distance 1 du concept *t1*, et concept *t3* est à distance 2 du concept *t1*.

Comme les liens de type l1 et l2 peuvent être établis en utilisant une multitude de propriétés, il est utile, comme nous le verrons par la suite, de pouvoir sélectionner certaines propriétés d'intérêt et d'écarter toutes les autres. Si K est un ensemble de propriétés d'intérêt, nous dénoterons par d^t_K une fonction de distance transversale valorisant uniquement les liens établis à travers des propriétés $k_1 \ldots k_n \in K$, et n'accordant une valeur de distance infinie aux concepts liés qu'à travers d'autres liens.

La fonction de pondération est conçue en appliquant la théorie d'information au cas des liens (elle s'applique habituellement à la valorisation du contenu informationnel des nœuds dans un graphe). Cette logique se traduit par l'hypothèse que les liens connectant

les concepts initiaux à beaucoup d'autres concepts comportent moins de poids informationnel que ceux les connectant à quelques rares concepts. Notre fonction est donnée par la formule 15. C'est une fonction logarithmique du rapport entre le nombre de liens établis depuis le concept initial c_1 avec n'importe quel autre concept à travers le même type de lien k qui le relie à c_2 et le nombre total de liens dans le graphe M. Dans les calculs réels, M peut être remplacé par un nombre, certainement supérieur au nombre de liens dans le graphe, sans changer le classement qui résulte de l'application de cette fonction. Si c_1 et c_2 sont liés par plusieurs propriétés, la valeur maximale du p^t_K est prise.

$$p^t_K(c_1,c_2) = -\log\left(\max_{k\in K}\frac{nombre_de_liens_k(c_1)}{M}\right), \exists \text{ lien de type k entre } c_1 \text{ et } c_2 \quad (15)$$

5.3.1.2.1. Adaptation de l'approche par rapport au cas d'utilisation

La densité relativement élevée du graphe de données liées, ainsi que la quantité de propriétés diverses utilisées dans des graphes riches de concepts[103], remet en question une approche qui utilise toutes les propriétés comme étant pertinentes pour le calcul des proximités. Nous nous sommes posés la question de savoir s'il était possible de faire un choix de propriétés les plus pertinentes pour un cas d'utilisation et se focaliser sur elles.

Si intuitivement il paraît plausible de sélectionner ces propriétés pertinentes de manière manuelle, en analysant leur sens tel qu'il est décrit dans les différentes ontologies, l'hétérogénéité de types concepts, et par conséquent des propriétés utilisées pour les représenter rend cette tâche très difficile, voire infaisable dans le temps imparti d'une thèse.

Dans [Passant, 2010] il est possible de trouver une approche pour la sélection des propriétés pertinentes. Celui-ci s'appuie sur l'analyse des propriétés utilisées dans les descriptions des concepts d'un type particulier bien défini. Notre approche visant à trouver des concepts de type différent et même ceux sans type explicite, il était difficile de répliquer cette même démarche.

Dans des cas comme le nôtre, il est possible de prendre quelques exemples représentatifs des concepts de départ et des concepts qui pourront être considérés comme proches, pour analyser les liens établis entre ces ensembles et trouver les propriétés utilisées pour établir ces liens. Pour établir une proximité de concepts dans un contexte particulier, il est possible de faire appel au jugement humain, ou bien de la déduire d'une source légitime (légitime par rapport à ce contexte).

Nous avons étudié des ensembles de concepts issus des anciens problèmes d'Hypios ainsi que l'ensemble des concepts extraits des solutions correspondant à ces problèmes. Nous avons ensuite analysé tous les chemins (d'une longueur maximale inférieure à 5) qui relient ces concepts. Notre analyse a permis de dévoiler que tous les chemins connectant des concepts de problèmes aux concepts de solutions étaient faits à travers seulement 5 propriétés transversales, présentées dans le Tableau 6. Ce nombre très faible de propriétés pertinentes pour un cas d'utilisation indique une vraie utilité de l'exercice de sélection des propriétés. Il est également important de noter qu'en raison

[103] A titre d'exemple, nous avons déterminé que DBpédia utilisait 46112 propriétés distinctes pour décrire ses concepts

des valeurs littérales souvent présentes dans le co-domaine de certaines propriétés, les liens indirects jouent un rôle très important dans les chemins que nous avons étudiés.

Tableau 6 Propriétés pertinentes

Propriétés transversales pertinentes pour le cas d'utilisation de la résolution de problème
http://dbpedia.org/ontology/product
http://dbpedia.org/property/products
http://dbpedia.org/ontology/industry
http://dbpedia.org/ontology/service
http://dbpedia.org/ontology/genre

Un nombre non négligeable de liens étaient des liens implicites, établis au travers de valeurs textuelles attribuées à l'une de ces 5 propriétés.

La nature de ces propriétés ne surprend pas. Ce sont en effet des propriétés susceptibles de relier des concepts dans un contexte industriel, lié aux produits et services. Notre fonction $hyP^{transversale}$ utilisera donc uniquement ces 5 propriétés dans le calcul des distances et des pondérations. Nous nous référons par la suite à cet ensemble de propriétés en tant que K_5.

5.3.1.2.2. Valeur ajoutée par l'utilisation de la fonction de pondération

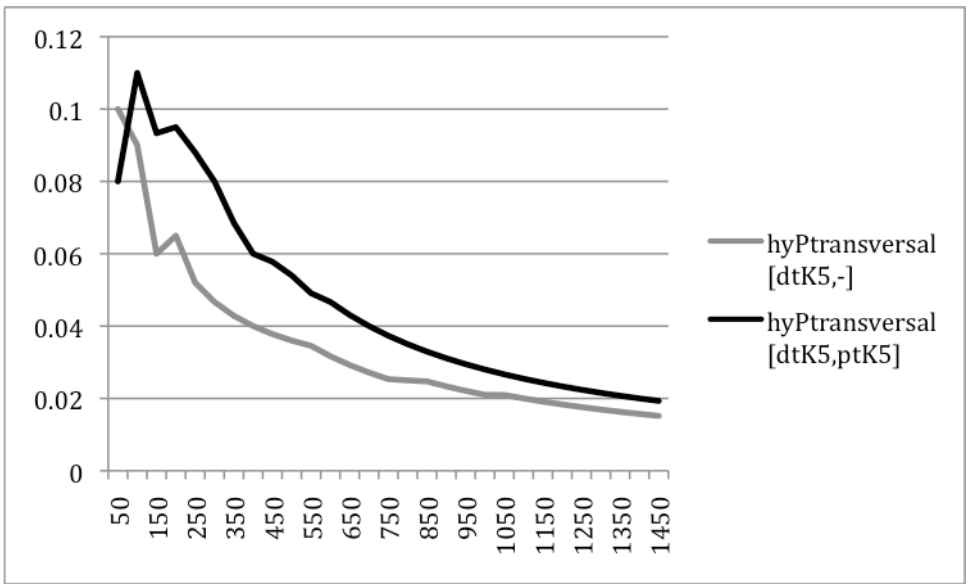


Figure 42 Précision de la fonction de distance transversale sans et avec la fonction de pondération

Une question qui s'impose naturellement est la valeur ajoutée par l'application de cette fonction de pondération. Une simple analyse des graphiques de précision et de rappel obtenus en utilisant la même approche de comparaison employée dans la partie §5.3.1.1.1 sur les deux cas :

- usage de la fonction d^t_{K5} sans aucune fonction de pondération i.e. en utilisant la constante 1 pour pondérer tous les concepts ;

. Usage de la même fonction de distance avec la fonction de pondération $p^t_{K5.}$

La comparaison présentée aux figures Figure 42 et Figure 43 indique clairement que l'usage de la fonction de pondération a un impact positif sur la performance de la fonction transversale.

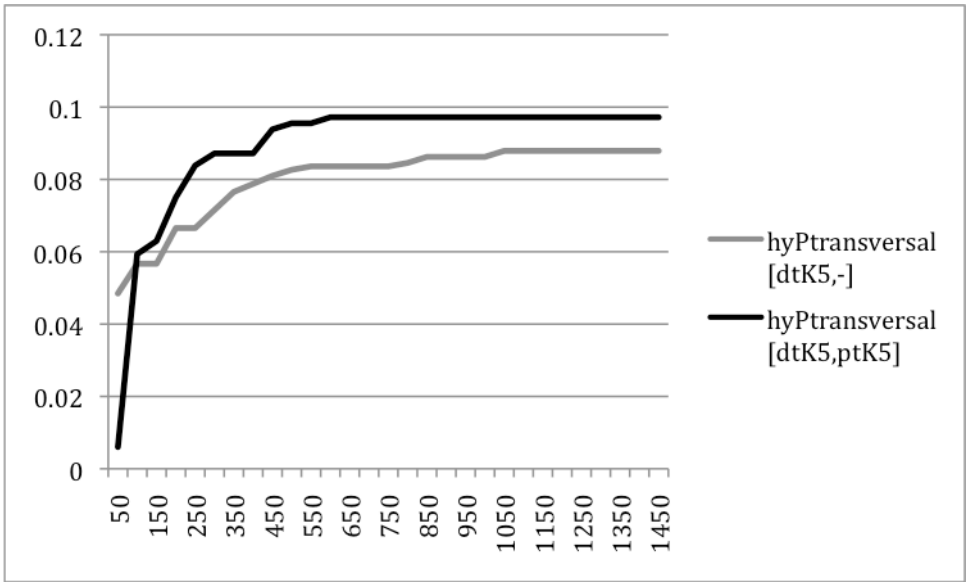


Figure 43 Rappel de la fonction de distance transversale sans et avec la fonction de pondération

5.3.1.3. Approche mixte

L'approche mixte représente une combinaison des approches hiérarchique et transversale. La formule 16 représente la fonction mixte, combinaison linéaire de ces deux fonctions. La nature additionnelle de la fonction impose que des fonctions composantes doivent avoir des valeurs comparables, i.e. dans les mêmes limites afin d'équilibrer leur influence sur la valeur composée. L'usage des fonctions de pondérations basées sur le contenu informationnel peut fournir un tel équilibre.

$$hyP^{mixte}(c_1,c_2) = hyP^{hiérarchique}(c_1,c_2) + hyP^{transversale}(c_1,c_2) \quad (16)$$

Plusieurs fonctions de distance et de pondération, hiérarchiques et transversales, peuvent figurer dans le calcul hyP^{mixte}. Inspirés par les analyses de performance de différentes fonctions, nous allons notamment utiliser la version présentée par la formule 17.

$$hyP^{mixte}(c_1,c_2) = \frac{p2(c_1,c_2)}{d2(c_1,c_2)} + \frac{p^t_{K_5}(c_1,c_2)}{d^t_{K_5}(c_1,c_2)} \quad (17)$$

5.3.1.4. Comparaison des approches

Dans cette partie nous présentons une comparaison des approches : hiérarchique, transversale et mixte, basée sur les mêmes principes que la comparaison des fonctions de distance hiérarchiques. Plus précisément nous comparons une mesure $hyP^{hiérarchique}$ s'appuyant sur la mesure de distance d2 et la mesure de pondération p2 ; la mesure

$hyP^{tranversale}$ utilisant la fonction de distance transversale d^{t}_{K5} et la fonction de pondération p^{t}_{K5} et la mesure mixte utilisant ces deux fonctions $hyP^{hiérarchique}$ et $hyP^{tranversale}$.

Au niveau de la précision, la fonction hyP^{mixte} se montre la meilleure. Un test T ($p<0.001$) a prouvé que les différences entre chacune des deux fonctions étaient significatives.

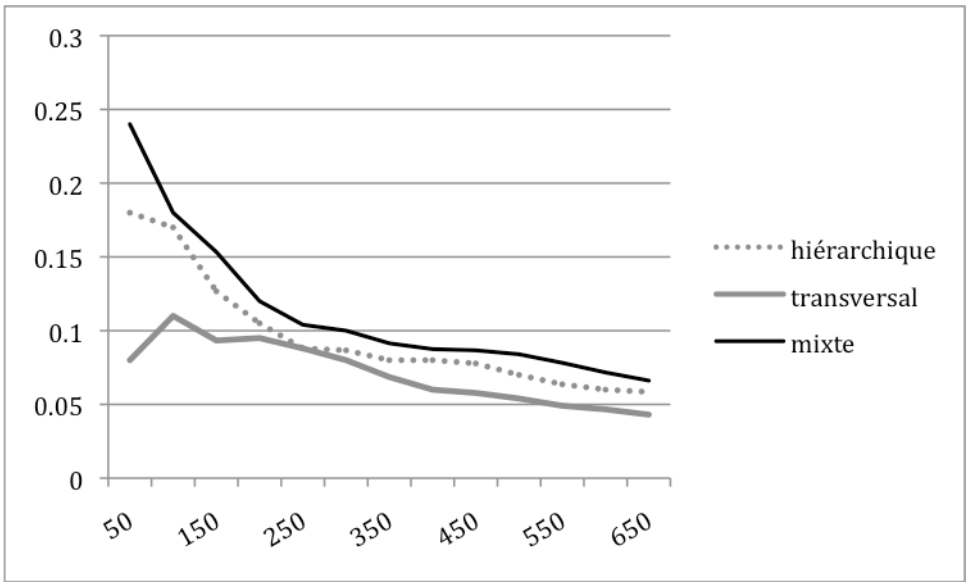


Figure 44 Précision de $hyP^{hiérarchique}$, $hyP^{tranversale}$, et hyP^{mixte}

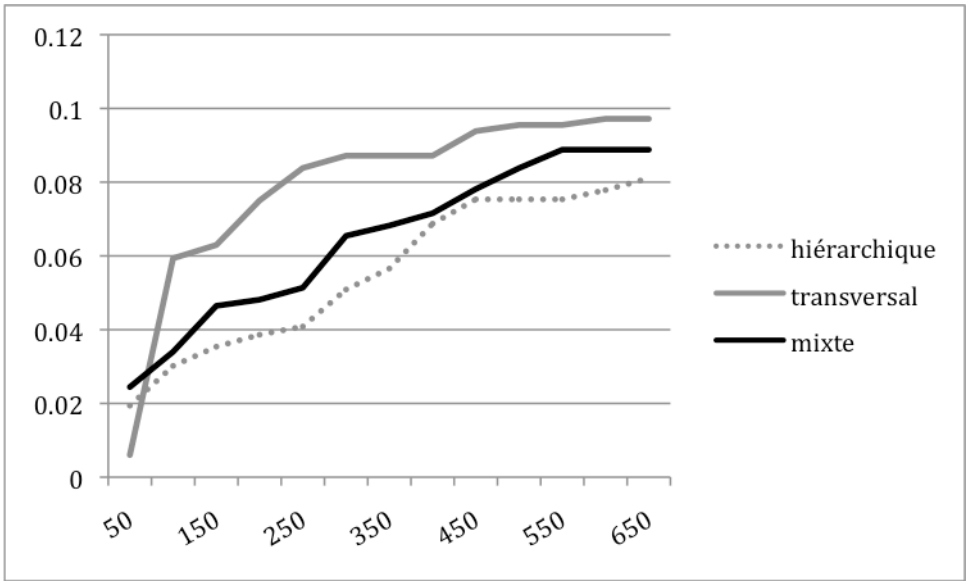


Figure 45 Rappel de $hyP^{hiérarchique}$, $hyP^{tranversale}$, et hyP^{mixte}

En comparant le rappel, nous pouvons observer que la fonction $hyP^{transversal}$ domine par sa performance les autres fonctions pour des jeux de résultats d'une taille supérieure à 100 concepts. Les différences ont été confirmées comme signifiantes par un test T entre la fonction mixte et la fonction hiérarchique ($p<0.001$), mais ce test n'a pas pu appuyer la signifiance des différences entre les autres fonctions dans ce cas.

L'instabilité des performances de ces fonctions également confirmée par la mesure F (Figure 46). Pour des ensembles de concepts suggérés d'une taille inférieure à 100 concepts, la mesure mixte est dominante. Pour une taille de 100 à 350, la dominance de

la fonction transversale est éclatante, alors que la fonction mixte reprend pour des ensembles de suggestions d'une taille supérieure à 350 concepts. Il est donc possible de choisir la fonction mixte ou hiérarchique en fonction de la quantité de concepts sollicitée dans un cas d'utilisation particulier. En fonction des besoins d'un cas d'utilisation, si la précision est beaucoup plus importante que le rappel, il pourrait être utile d'utiliser la fonction mixte quelle que soit la quantité de résultats désirée.

La fonction hiérarchique est complètement dominée par la fonction mixte, dans le cas de la mesure F aussi bien que dans les cas d'autres mesures de performances. Cela nous permettra d'exclure la fonction hiérarchique des analyses que nous présenterons par la suite.

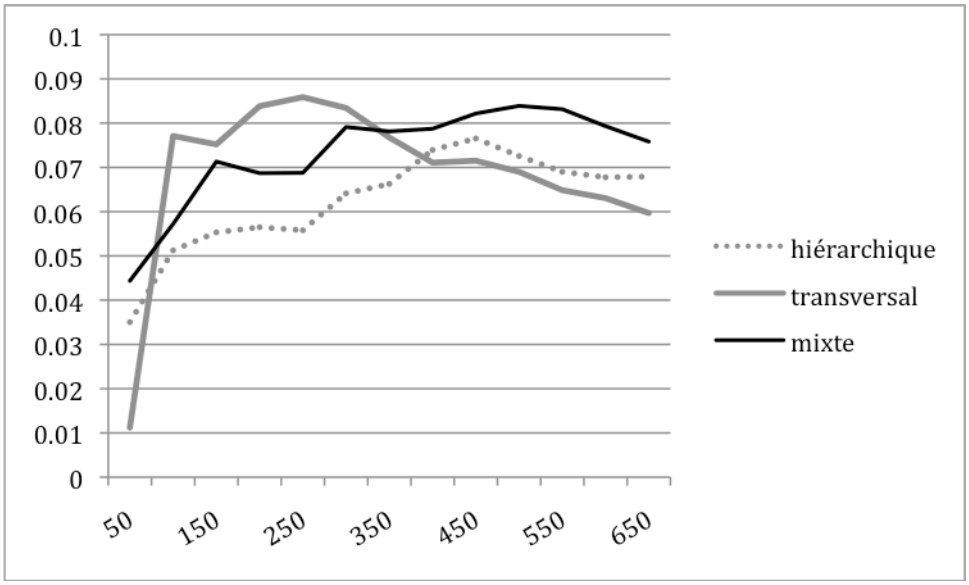


Figure 46 Mesure F de $hyP^{hiérarchique}$, $hyP^{tranversale}$, et hyP^{mixte}

5.4. Evaluation

5.4.1. Evaluation à l'aide de l'étalon-or

Si la conception de notre approche nous a amené à expérimenter différentes approches alternatives du parcours de graphe, certaines inspirées par l'état de l'art et adaptées au graphe de données liées, la comparaison avec des approches basées sur la cooccurrence et sur la distribution de concepts dans des corpus est restée un défi. Nous présentons dans cette section une telle comparaison avec deux systèmes concurrents :

- **Google AdWords Keyword Tool**. Ce service en ligne[104] sert notamment à suggérer des mots-clés alternatifs pour des campagnes publicitaires en ligne. Il utilise des fonctions de proximité par cooccurrence dans des requêtes de recherche pour proposer des mots-clés proches d'un ensemble de mots-clés initial. Très utilisé, ce système est un bon représentant des approches basées sur la cooccurrence dans le texte et dans les requêtes.

104 https://adwords.google.com/o/Targeting/Explorer?__c=1000000000&__u=1000000000&ideaRequestType=KEYWORD_IDEAS

- **Random Indexing**. Cette méthode à laquelle nous avons déjà fait référence dans l'état de l'art fut proposée par Damljanovic et al. [Damljanovic et al., 2011]. Celle-ci transpose des approches de corrélation dans le texte à la problématique du traitement des distances dans le graphe de données liées. Grâce à la coopération des auteurs de cette méthode, nous avons eu la chance de pouvoir comparer nos deux systèmes.

La comparaison a été effectuée de la même manière que dans les parties précédentes, sur le même jeu de données. La seule différence est la manière dont la performance du système de Google AdWords a été évaluée. Ce système propose des mots-clés ou des phrases clés qui parfois correspondent à des concepts identifiables dans DBPedia et dans d'autres sources de concepts du graphe de données liées, mais souvent représentent des combinaisons de mots difficile à associer avec un concept, ou des appellations alternatives du même concept. Pour cette raison nous avons considéré qu'il serait inapproprié d'essayer de mettre en correspondance ces mots-clés avec des concepts DBPedia ; Nous avons donc plutôt appliqué une vérification directe si les mots-clés suggérés apparaissent dans les textes des solutions reçues pour un problème. Cette vérification textuelle est une méthode mieux adaptée à la forme des résultats fournis par Google AdWords Keyword Tool et remplace la vérification si des concepts DBPedia proposés par nos méthodes et par Random Indexing apparaissent dans des concepts extraits à partir de solutions.

5.4.1.1. Comparaison

La comparaison révèle que nos méthodes basées sur le graphe obtiennent une performance largement supérieure a celles des systèmes basés sur la cooccurrence (cette domination est confirmée par test T, $p<0.001$).

En termes de rappel, Random Indexing surpasse la performance de nos mesures pour la tranche de 0-200 concepts suggérés. Après cette limite notre mesure $hyP^{transversale}$ reste dominante.

La mesure F cette fois-ci n'ajoute pas beaucoup de nouvelles informations par rapport à la précision et au rappel.

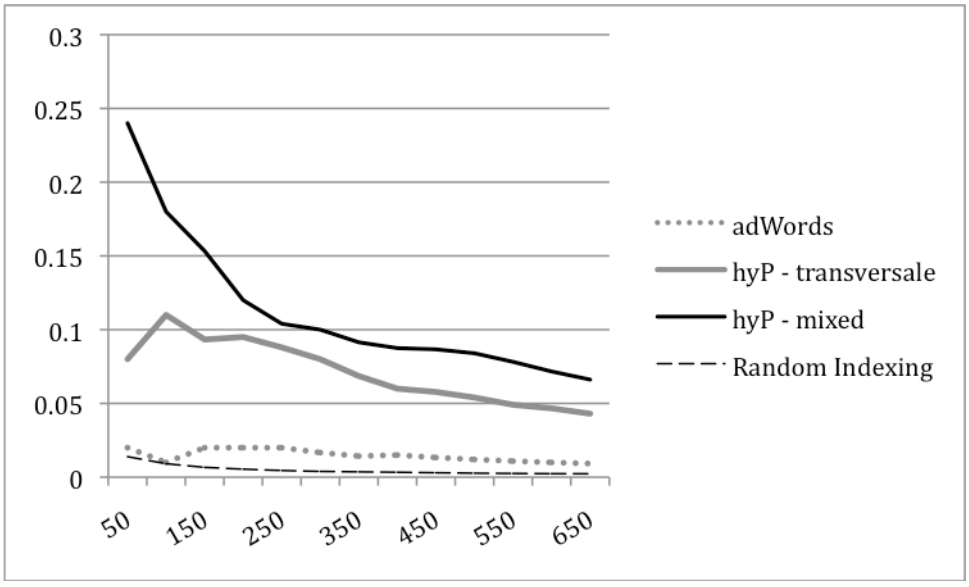


Figure 47 Précision $hyP^{hiérarchique}$, $hyP^{tranversale}$, adWords et Random Indexing

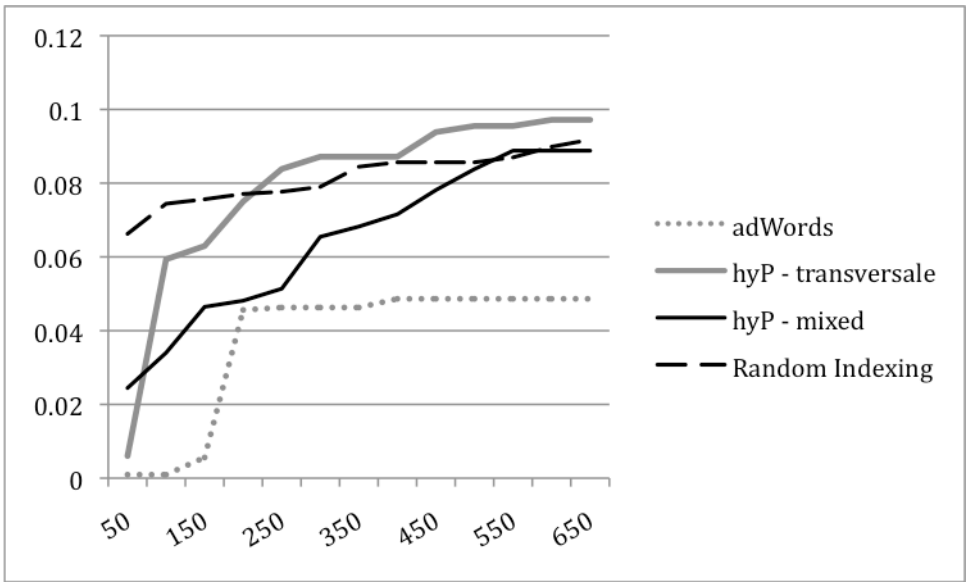


Figure 48 Rappel de hyP^{hiérarchique}, hyP^{tranversale}, adWords et Random Indexing

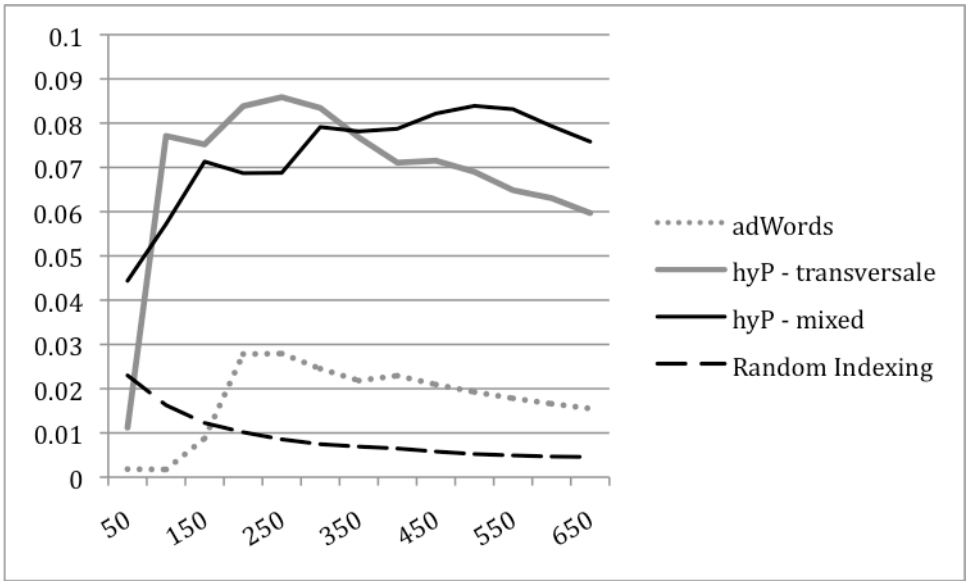


Figure 49 Mesure F de hyP^{hiérarchique}, hyP^{tranversale}, adWords et Random Indexing

5.4.2. Etude utilisateurs

Si les données issues des problèmes et solutions collectées chez Hypios nous ont permis de comparer plusieurs alternatives à notre approche, cette comparaison ne comporte pas de conclusions sur la satisfaction des utilisateurs sur les concepts recommandés. Nous avons donc conduit une étude utilisateur pour récolter des avis d'utilisateurs sur différents aspects de la qualité des recommandations proposées par différents systèmes (le nôtre et les systèmes issus de l'état de l'art).

Il est important de souligner que cette étude, aussi bien que la comparaison basée sur des anciens problèmes, souffre de certaines limitations concernant ses conclusions. Nous avons par exemple demandé aux utilisateurs de juger si les concepts suggérés étaient pertinents pour la résolution de problèmes. Si, de par leurs connaissances, ils

sont susceptibles de bien évaluer les concepts pertinents, il existe toujours un risque que des concepts jugés comme non pertinents puissent conduire à une solution inattendue. La nature imprévisible d'une pertinence *a posteriori* rend une évaluation par les utilisateurs inévitablement incomplète. Pour cette raison, cette étude représente plutôt un complément à la comparaison déjà présente dans §5.4.1. Les conclusions que nous tirons de la performance de notre approche sont fondées sur l'effet synergique de ces deux expériences.

5.4.2.1. Contexte de l'étude

L'étude que nous avons menée vise principalement à évaluer la capacité de notre méthode à suggérer des concepts utiles dans un processus de diffusion d'un problème d'innovation à un public d'experts potentiels en ligne. Il est donc question de générer un rapprochement de concepts qui soit utile pour relier les problèmes et potentiels auteurs de solutions, d'une manière pertinente et inattendue, ainsi que nous l'avons déjà évoqué dans les besoins de notre cas d'utilisation décrit dans §5.1.2. Cette étude repose sur l'hypothèse que des utilisateurs ayant une expérience avec le processus de la résolution de problèmes en ligne soient capables d'évaluer la qualité des suggestions de concepts et leur utilité par rapport à ces besoins.

Les utilisateurs participant à nos expériences ont été choisis très attentivement. Nous avons pris soin de nous assurer que leur connaissance du processus de la résolution des problèmes en ligne, de la diffusion de problèmes et de la nature diversifiée des concepts sollicités soit suffisante. Ces participants appartenaient notamment aux groupes suivants :

- **Employés d'Hypios**. Nous avons notamment choisi des personnes travaillant directement dans le processus de diffusion de problèmes et dans la gestion des relations avec les auteurs de solutions. Des personnes participant aux relations avec les clients et partenaires, particulièrement expérimentées en formulation de problèmes, ont également pris part dans nos études.
- **Employés des entreprises partenaires d'Hypios spécialisés en innovation ouverte**. Nous avons eu la chance de travailler avec des employés de Bluenove[105], une entreprise de conseil en innovation ouverte qui dans le cadre d'un partenariat avec Hypios fournit un soutien à des clients en les aidant à formuler leurs problèmes et les soumettre chez Hypios. Leur capacité à juger l'utilité des concepts suggérés est susceptible d'être proche de celle des personnes en charge de la diffusion de problèmes chez Hypios, avec qui ils collaborent régulièrement.
- **Doctorants qui étudient le phénomène de l'innovation ouverte**. Des doctorants du laboratoire Dauphine Recherche en Management[106] de l'université Paris-Dauphine qui étudient, ayant pour perspective la science de gestion, le phénomène des plateformes d'innovation en ligne, ont également participé à nos études. Ayant étudié beaucoup de cas d'innovation en ligne et suivi les pratiques émergentes de l'innovation ouverte, leur jugement sur les concepts suggérés est peu contestable.

[105] http://www.bluenove.com/

[106] http://www.drm.dauphine.fr/

Dans chaque étude, nous avons fait appel à des participants appartenant à ces trois catégories.

Nous avons conduit deux études, dans les différentes phases de développement de notre approche. Dans une première étude nous avons comparé notre approche hiérarchique développée en premier avec Google Adwords Keyword Tool. Dans la deuxième étude nous avons comparé notre approche mixte avec ce même outil et aussi avec Random Indexing, dont nous avons pris connaissance entre-temps. Dans chaque étude les utilisateurs choisissaient au moins un problème d'innovation à traiter. Ils avaient le choix de travailler avec un problème d'Hypios dont le sujet concordait avec leurs connaissances ou bien de rédiger eux-mêmes l'énoncé d'un problème. Ensuite, nous avons présenté des concepts initiaux, extraits du problème par Zemanta pour vérification. À ce stade, l'utilisateur pouvait effacer certains concepts initiaux extraits par erreur où en ajouter certains qui manqueraient. Cinq à sept concepts initiaux étaient présents dans les concepts initiaux après vérification. Ensuite, nous présentions pour chaque problème 30 suggestions de concepts par système évalué, générés à partir des concepts initiaux et nous sollicitions les avis des utilisateurs. Il leur a spécifiquement été demandé d'évaluer deux aspects des suggestions :

- **Pertinence par rapport au problème** sur une échelle de 1 à 5, avec l'interprétation suivante

  1 : non pertinent (un expert en ce sujet ne pourrait certainement pas résoudre le problème)

  2 : assez peu pertinent (il est peu probable qu'un expert en ce sujet propose des solutions)

  3 : neutre (il existe une chance qu'un expert en ce sujet propose des solutions)

  4 : assez pertinent (un expert en ce sujet aurait probablement des idées de solutions)

  5 : pertinent (un expert en ce sujet pourrait certainement avoir des idées de solutions)

- **Aspect inattendu** des suggestions désignant leur caractère surprenant pour l'utilisateur dans le contexte du problème en question et des voies de résolution que l'utilisateur juge probables. Nous avons utilisé l'échelle suivante :

  1 : évident (le concept apparaît dans l'ensemble des concepts initiaux)

  2 : assez évident (le concept s'impose comme évident étant donné les concepts initiaux)

  3 : peu évident (le concept est peu évident par rapport aux concepts initiaux)

  4 : surprenant (l'utilisateur n'aurait pas pensé lui-même à ce concept dans le contexte du problème donné)

  5 : précédemment inconnu (l'utilisateur ne connaissait même pas ce concept).

Nous présentons par la suite les résultats des deux études.

5.4.2.2. Comparaison avec Google AdWords

Dix utilisateurs ont participé à notre première évaluation, en comparant des suggestions de hyProximity, plus précisément la fonction de proximité hiérarchique utilisant d2 comme fonction de distance et p2 comme mesure de pondération, avec des suggestions d'AdWords Keyword Tool. Les utilisateurs ont évalué des suggestions correspondant à 22 problèmes, faisant au total 2152 notes sur la pertinence et l'aspect inattendu des concepts suggérés.

Tableau 7 Note moyenne ± déviation standard dans l'étude 1

mesure	adWords	hyProximity
pertinence	3.682±0.16	3.362±0.17
inattendu	2.580±0.16	3.461±0.18
inattendu (pertinence>=4)	2.322±0.24	3.169±0.30
inattendu (pertinence=5)	1.939±0.25	2.502±0.38

En ce qui concerne la performance, les résultats montrent une différence relativement faible en faveur de système d'AdWords. Nous n'avons pas pu confirmer cette différence comme statistiquement signifiante par le test T ($p>0.05$). Par contre, la différence en terme du caractère inattendu des suggestions est beaucoup plus importante entre les deux systèmes ($p<0.0001$) en faveur du hyProximity. Néanmoins, le caractère inattendu, vu isolément, n'est pas un indicateur pertinent de l'utilité des suggestions, car une suggestion complètement inappropriée peut être jugé comme inattendue. Nous nous intéressons donc à évaluer la synergie de la pertinence et du caractère inattendu des suggestions. Pour ce faire, nous calculons le caractère inattendu des concepts jugés pertinents, et cela à deux niveaux de valeur de pertinence >=4 et pertinence =5. Pour les deux niveaux, la note moyenne des suggestions de hyProximity montre des résultats considérablement meilleurs que ceux d'AdWords. Une récapitulation de tous ces résultats est donnée dans le Tableau 7.

5.4.2.3. Comparaison avec Random Indexing

Une fois que notre approche a été enrichie par les fonctions transversales et mixtes comme prolongements naturels des travaux initiaux dédiés au développement d'une fonction de distance hiérarchique, il était question de répéter l'étude utilisateur pour comparer la nouvelle version de notre approche avec les approches existant dans l'état de l'art. L'approche Random Indexing ayant été publiée entre-temps, nous avons donc profité de notre nouvelle étude pour l'inclure dans la comparaison. Parmi nos fonctions alternatives, nous avons choisi de comparer la fonction mixte utilisant l'approche hiérarchique avec d2 et p2, et des fonctions d^t_{K5} et p^t_{K5} dans la partie transversale de la formule mixte.

Douze utilisateurs, dont seulement 3 participants de la première étude, appartenant aux 3 groupes désignés précédemment comme compétents, ont participé à l'étude en générant des évaluations de suggestions correspondant à 34 problèmes, faisant un total de 3060 concepts évalués.

Les résultats sont présentés dans le Tableau 8. Les deux systèmes utilisant le graphe de données liées (le nôtre et Random Indexing) ont une pertinence de concepts suggérés supérieure à celle de Google AdWords. La signifiance de la différence entre chaque

couple de systèmes a été confirmée par le test T ($p < 0.005$). Nous n'avons pas pu confirmer la même chose pour la différence

Tableau 8 Note moyenne ± déviation standard dans l'étude 2

mesure	AdWords	hyProximity (mixte)	Random Indexing
pertinence	2.930±0.22	3.693±0.23	3.330±0.25
inattendu	2.859±0.16	2.877±0.25	3.052±0.22
inattendu (pertinence>=4)	2.472±0.31	2.542±0.36	2.635±0.36
inattendu (pertinence=5)	1.760±0.22	1.842±0.31	1.767±0.36

en terme du caractère inattendu des suggestions par le test T ($p >0.05$).

En ce qui concerne les concepts à la fois pertinents et inattendus, hyProximity atteint la meilleure performance : si nous ne prenons en compte que les concepts les plus pertinents, alors que Random Indexing le surpasse légèrement si nous prenons en compte les concepts dont la pertinence est notée 4 et 5.

Dans une deuxième phase, nous avons demandé aux participants de se soumettre à un entretien afin de récolter leur opinion en leur permettant de s'exprimer plus librement sur les aspects positifs et négatifs des suggestions qu'ils ont reçues. Cette expérience nous a permis de parfaire nos connaissances sur la nature qualitative de leurs jugements et même, dans certains cas, expliquer qualitativement la différence entre les systèmes. Pour les utilisateurs, les systèmes ont été nommés « Système 1,2 et 3 », et leur ordre était aléatoire et différent pour chaque utilisateur. Grâce à un enregistrement de l'ordre de systèmes pour chaque étude, sauvegardé par notre méthode, nous avons pu déchiffrer à quel système se réfèrent les appréciations qualitatives des utilisateurs.

Lors des entretiens nous avons demandé aux utilisateurs de décrire avec leurs propres mots la nature des concepts suggérés. Presque tous les utilisateurs ont commencé par une déclaration de genre : « Il y a un système qui n'est pas du tout pertinent, et deux bons. Parmi ces deux-là, l'un est un peu mieux ». Les adjectifs les plus couramment utilisés pour caractériser les suggestions d'AdWords étaient « redondant » et « orienté recherche Web » pour souligner que les suggestions de ce système avaient un caractère peu enrichissant dans le contexte de la résolution de problème en ligne et de la recherche de concepts diversifiés. Une telle caractérisation ne surprend pas, compte tenu de la nature statistique du fonctionnement de ce système qui tire une notion de similarité par la cooccurrence et non pas par le sens similaire ou complémentaire de termes. Les suggestions de hyProximity furent le plus souvent jugées comme « très intéressantes » et « dans l'esprit de l'innovation ouverte » pour souligner leur caractère souvent très concret, inattendu et potentiellement pertinent. Les suggestions proposées par Random Indexing ont le plus souvent été caractérisées comme « très générales » pour évoquer la capacité de ce système à trouver des généralisations pertinentes sur des concepts initiaux.

5.5. Le logiciel hyProximity

Après nous être rassurés sur l'utilité de notre approche au travers de ces évaluations, nous avons construit un programme informatique visant à permettre aux utilisateurs finaux d'accéder à notre moteur de suggestion de concepts.

Le logiciel, réalisé en Java, est composé d'un cœur de moteur de suggestion de concepts accessible par la classe hyProximityFacade (Figure 50). Cette classe implémente les fonctionnalités clés du système, telles que :

- pour un texte donné, obtenir des concepts initiaux ;
- pour un texte donné, obtenir des suggestions de concepts ;
- pour un ensemble de concepts initiaux, obtenir des suggestions de concepts.

Il est possible d'accéder à ces fonctionnalités via deux modes :

- application Web, destinée à l'usage des utilisateurs finaux ;
- API qui permet aux autres applications Web d'accéder aux fonctionnalités du système.

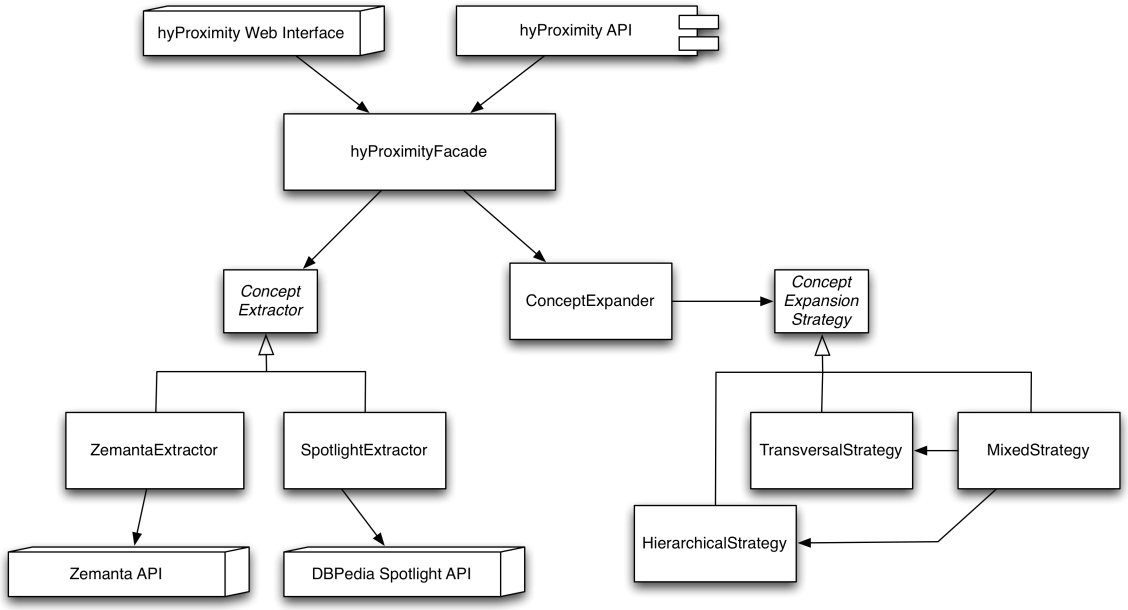


Figure 50 Vue d'haut niveau des composants principaux du logiciel HyProximity

Dans les deux modes, il est possible de configurer le nombre de suggestions désirées ainsi que de préciser la langue du texte fourni à l'input. Notre programme peut travailler avec plusieurs langues à condition qu'un extracteur de concepts soit disponible pour chaque langue[107]. Pour le moment, en l'absence d'un extracteur fiable de concepts de DBPedia qui soit capable de travailler avec d'autres langues que l'anglais, nous appliquons une astuce que nous avons testée par pure curiosité, et qui s'est avérée fonctionnelle. Celle-ci consiste à traduire le texte fourni en anglais, en passant par un système de traduction automatique tel que Google Translate[108]. Si la traduction automatique est d'une qualité assez pauvre et même parfois difficile à lire, ce texte traduit peut servir pour l'extraction de concepts, qui peut travailler avec les quelques phrases qui sont correctement traduites. Nous sommes conscients d'une possible perte d'information dans ce processus et pour cette raison nous déconseillons la mise en marche de notre logiciel avec les textes autres que l'anglais tant qu'un outil d'extraction

[107] Grace à sa conception modulaire, le système peut être configuré pour travailler avec des nouveaux extracteurs dès qu'ils deviennent disponibles.

[108] http://translate.google.fr/

de concepts multilingue, capable de travailler avec des concepts de DBPedia, n'est pas encore à notre disposition.

La fonctionnalité d'extraction de concepts initiaux est assurée par un module, utilisé également dans nos autres logiciels et décrit déjà dans la partie §6.3.2. Si la flexibilité de conception permet l'usage de plusieurs services d'extraction et même l'implémentation des composants permettant de s'appuyer sur d'éventuels nouveaux services, dans la configuration actuelle nous utilisons Zemanta.

La fonctionnalité de suggestion de concepts est assurée par la classe ConceptExpander. Cette classe implémente notre algorithme détaillé dans §5.3.1, d'une manière générale, sans fixer l'implémentation d'une fonction de proximité. Celle-ci est confiée à des classes implémentant l'interface ExpansionStrategy. Ces classes concrètes implémentent les fonctions hiérarchiques, transversales et mixtes. Lors de nos expériences nous avions implémenté une multitude de fonctions alternatives, mais nous n'avons gardé par la suite que la fonction hiérarchique utilisant d2 et p2, la fonction transversale utilisant d^t_{K5} et p^t_{K5}, et la fonction mixte combinant ces deux fonctions. Dans un fichier de configuration il est possible de spécifier la classe concrète à utiliser pour l'exécution de l'algorithme. La classe MixedExpansionStrategie implémentant la fonction mixte est utilisée le plus souvent.

Dans le calcul de proximités, le système utilise la librairie Jena pour effectuer des requêtes SPARQL sur le point d'accès en ligne du DBPedia.org. Il est également possible de configurer le système pour qu'il utilise une instance de DBPedia locale ou un autre jeu de données en spécifiant l'URI de son point d'accès SPARQL.

5.5.1. L'interface Web de hyProximity

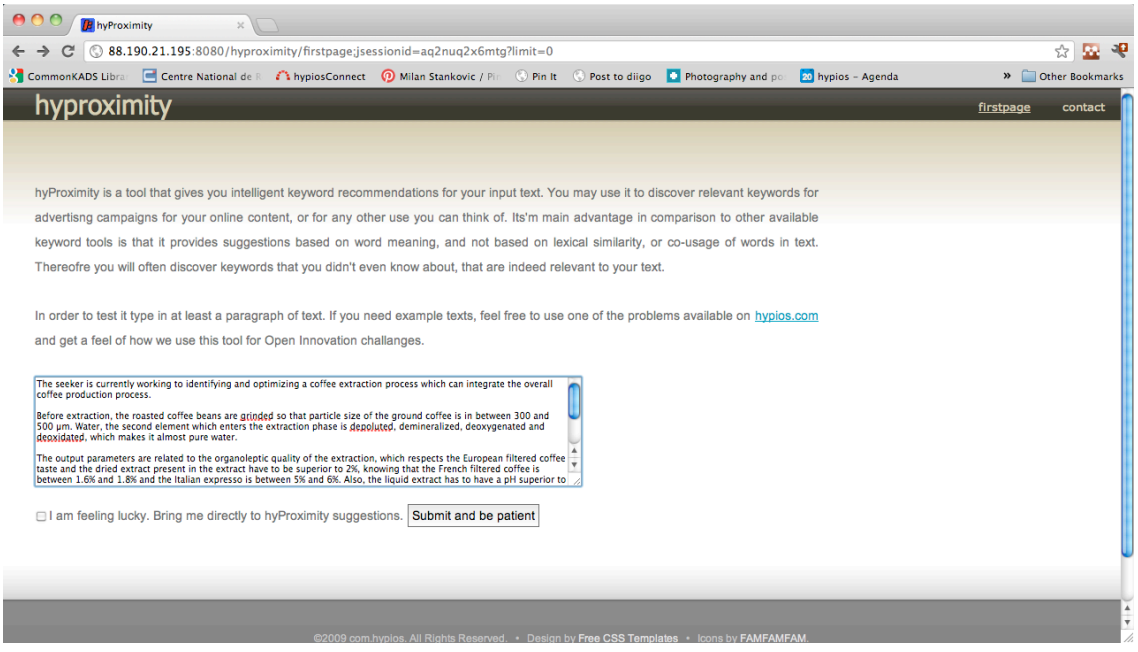


Figure 51 Interface hyProximity - page 1

L'interface Web de notre système est actuellement utilisée en interne par les employés d'Hypios. Une version, dont le nombre de suggestions fournies est limitée, est accessible de manière publique[109].

Le premier pas dans l'interaction avec le système est l'insertion d'un texte de problème d'innovation (ou d'un autre texte pour lequel les suggestions de concepts de ce type peuvent être utiles) dans le champ d'entrée textuelle (Figure 51). Ensuite l'utilisateur clique pour passer à une prochaine étape de vérification des concepts initiaux extraits directement du texte. Optionnellement il peut cocher la case « I'm feeling Lucky » pour signaler qu'il fait confiance au système d'extraction et qu'il souhaite être directement redirigé vers la page de suggestions hyProximity.

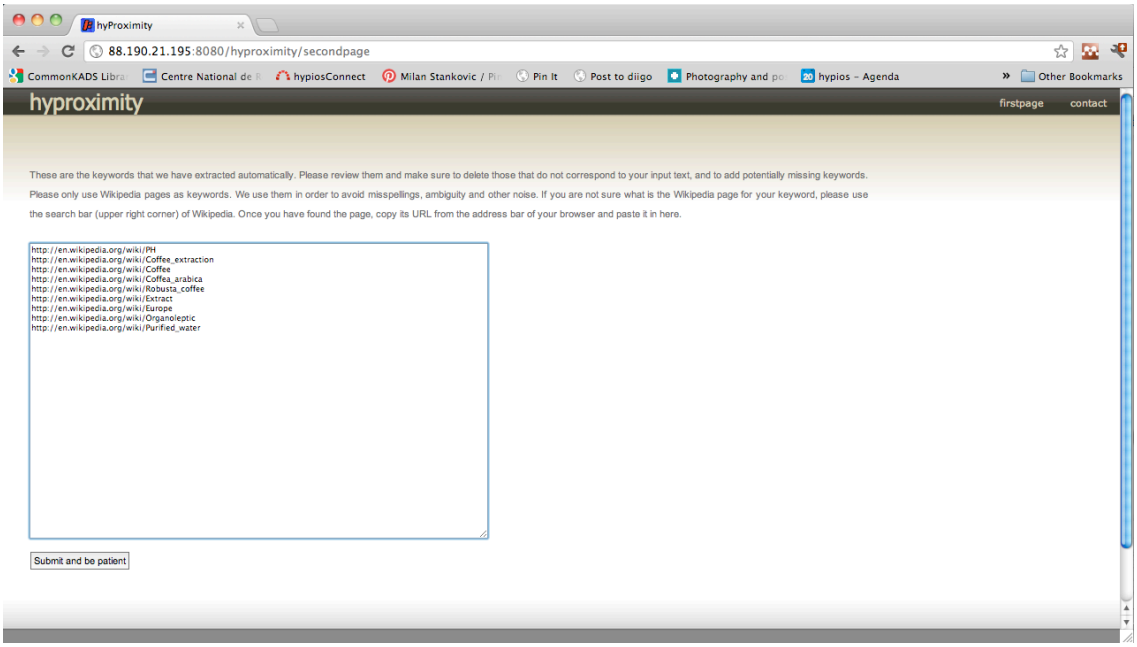


Figure 52 Interface hyProximity - page 2

Dans la deuxième étape (Figure 52), le logiciel fournit une liste de concepts initiaux. Ces concepts sont représentés par leur page Wikipédia. L'utilisateur peut modifier cette liste en retirant certains concepts ou bien en ajoutant d'autres liens Wikipédia. Nous avons choisi d'utiliser des liens Wikipédia pour supprimer toute ambiguïté conceptuelle. Il aurait également été possible d'implémenter un système d'auto-complétion d'input utilisateur, en s'assurant que la même représentation textuelle était utilisée chaque fois qu'un concept était rappelé – une approche par exemple adoptée par Faviki[110].

À l'étape suivante (Figure 53), le système fournit des concepts retrouvés par notre approche. Chaque concept est représenté par son titre en anglais, donné sous la forme d'un lien vers la page Wikipédia. Suite à la demande des utilisateurs, nous avons également implémenté l'affichage d'une partie de la description textuelle du concept (obtenu de DBPedia) au survol de la souris. Cette nouvelle fonctionnalité facilite le choix de concepts à utiliser par la suite dans différents scénarios.

[109] http://research.hypios.com/hyproximity , code:imsorrydave

[110] Faviki est une application d'organisation de marques-pages qui se sert des concepts DBPédia pour désambigüiser des tags http://www.faviki.com/

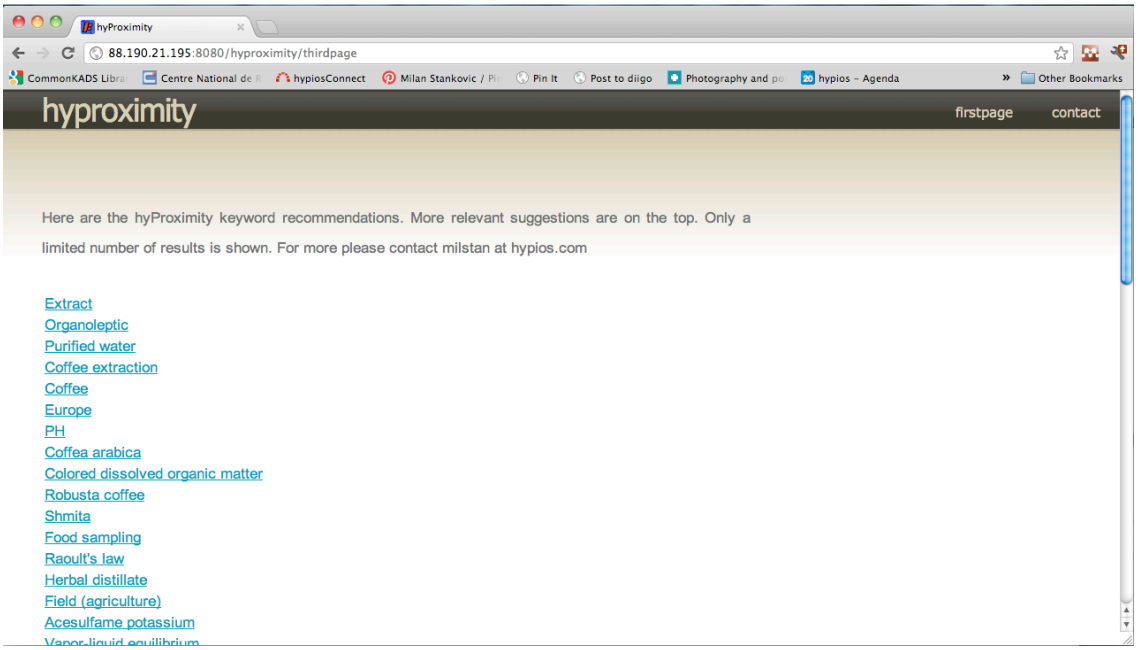


Figure 53 Interface hyProximity - page 3

5.6. Utilisation du système dans la plateforme Hypios

5.6.1. Cas d'utilisation 1

Nous avons observé l'usage d'hyProximity par Hypios dans le cadre d'une campagne de diffusion de problème aux experts. Cette campagne menée dans les débuts d'Hypios reposait toujours sur un système d'envoi automatique d'e-mails, sans intervention humaine. Le processus consistait à rechercher des experts correspondant à des concepts/mots-clés obtenus par un système (AdWords ou hyProximity) puis de leur envoyer un e-mail automatique contenant un lien vers la description du problème avec un texte d'invitation à prendre part dans sa résolution. La recherche d'experts fut effectuée sur une base d'experts que Hypios a constituée en parcourant le Web et en extrayant des informations pertinentes à partir des pages Web d'experts, de chercheurs et de leurs articles de recherche. Cette base comptait à l'époque de cette expérience plus de 800.000 personnes.

Grâce à un système de suivi de visites sur la page Web du problème, système mis en place par le département de développement d'Hypios, nous avons pu tracer le nombre de visiteurs distincts venant sur le site à travers le lien qui leur avait été envoyé dans cette campagne. Des liens différents ont été envoyés à chaque expert afin de pouvoir trier les résultats.

Nous avons choisi pour cet exercice un problème relatif à l'industrie alimentaire. Le problème concernait la recherche d'une bonne consistance pour une crème destinée à fourrer des barres chocolatées : elle devait être onctueuse, comparable à celle que l'on trouve dans les gâteaux pâtissiers, mais devait aussi pouvoir être conservée en magasin pendant au moins 6 mois.

Les résultats de la campagne de diffusion sont présentés dans le Figure 14. Le nombre d'experts trouvés dans la base en utilisant des suggestions d'AdWords était légèrement

supérieur au nombre d'experts trouvés en utilisant seulement des concepts de départ. Le nombre d'experts trouvés en utilisant des suggestions de hyProximity était considérablement supérieur.

En ce qui concerne le taux de clics sur les liens envoyés dans les e-mails (« nombre de réponses » dans le Figure 14), le résultat est assez similaire. 2.24% des experts identifiés par les concepts de hyProximity ont cliqué sur le lien, contre 1.98% des experts identifiés par les mots-clés d'AdWords et 1.83% des experts identifiés par les concepts initiaux. Il est donc raisonnable de conclure que l'usage des suggestions de hyProximity était utile dans le processus de diffusion de problème.

Tableau 9 Résultats de la campagne de diffusion du problème

Système	Nombre d'experts trouvés	Nombre de réponses	%
Concepts initiaux	1802	33	1.83
AdWords	2061	41	1.98
hyProximity	2849	64	2.24

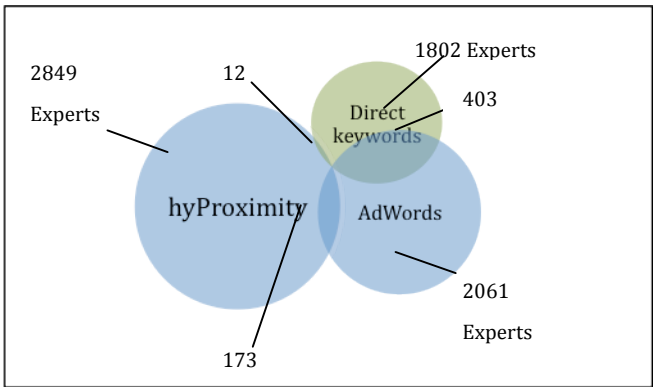


Figure 54 Intersections d'ensemble d'experts

Il est également intéressant d'observer les intersections des ensembles d'experts identifiés par les différents systèmes. Les experts identifiés par hyProximity ont une intersection notamment faible avec les experts identifiés par des concepts directs et par ceux suggérés par AdWords. Étant donné les taux de réponses présentés ci-dessus, il semble raisonnable de combiner des mots-clés provenant des différents systèmes pour atteindre une masse d'experts plus large.

Finalement, ce problème a reçu une solution inattendue qui a plu au client. Contrairement aux critères spécifiés dans l'appel par le client, cette solution proposait de changer non pas la structure de la crème mais plutôt celle de l'emballage en utilisant des nanotechnologies.

Afin de s'assurer que cette différence entre le nombre d'experts trouvés et intéressés n'est pas due au hasard, nous avons répété l'expérience avec 3 autres problèmes présents sur le site à l'époque (en février 2011). Les résultats, présentés dans Tableau 10, ont été très similaires.

Tableau 10 Résultats de la diffusion des 3 problèmes de février 2011

Système	Nombre d'experts trouvés	Nombre de réponses	%
Concepts initiaux	4562	67	1.47
AdWords	5120	83	1.62
hyProximity	6997	141	2.01

5.6.2. Cas d'utilisation 2

Lors du traitement d'un problème lié à la fabrication de café, Hypios a pris une démarche inhabituelle dans l'usage de hyProximity. Au lieu d'attendre une version finale de la description du problème, un premier brouillon de celui-ci a été soumis à hyProximity (Phase 1 sur la Figure 55). Un certain nombre de concepts a été extrait et sauvegardé. La description du problème fut ensuite changée encore 2 fois (phase 2 et 3 sur la figure Figure 55). À chaque phase, l'équipe d'Hypios fournissait la formulation du problème à hyProximity pour générer 50 concepts, en collaboration étroite avec un entreprise spécialisée dans le conseil en innovation et chargée de la rédaction du problèmes à soumettre aux chercheurs. Dans ce processus collaboratif, les personnes en charge faisaient le choix des concepts à retenir. Si les concepts extraits dans les premières phases ont aidé les rédacteurs à mieux élaborer certains points et développer davantage les aspects peu représentés et néanmoins importants, le client, une grande entreprise française, a également voulu exclure certaines parties de la formulation initiale pour des raisons de confidentialité. Ces changements ont contribué à un cumul de concepts assez important au fil du temps.

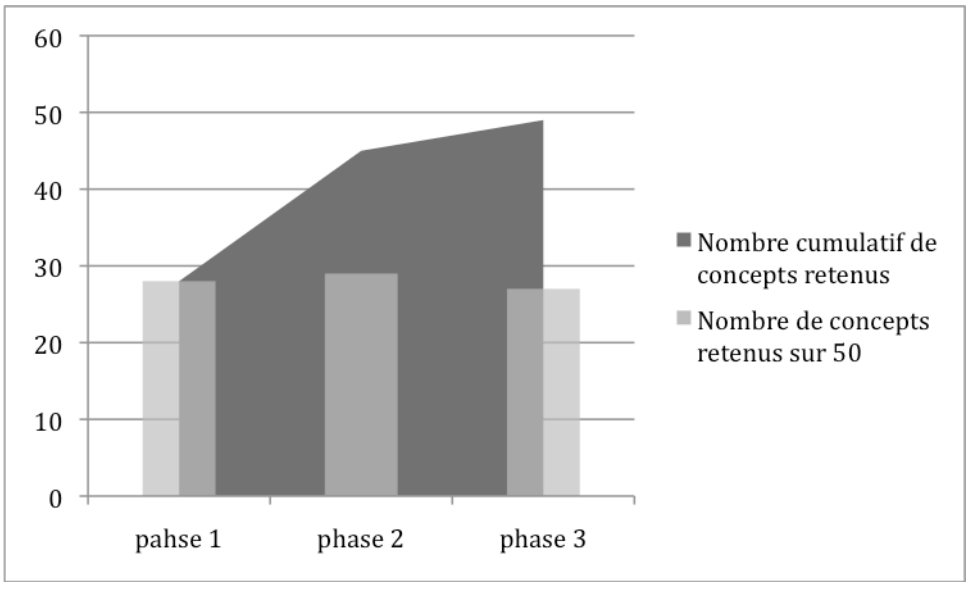


Figure 55 Rétention de concepts suggérés par hyProximity dans un processus itératif de rédaction de problème[111]

[111] Nous devons notre gratitude à Denis Wodrascka, Directeur de la relation clients chez Hypios, pour avoir institué une manière systématique de notation de tous les pas d'interaction avec hyProximity au

La Figure 55 montre une progression du nombre de concepts retenus lors du processus itératif. Il est clair que chaque nouvelle version permet d'ajouter un nombre considérable de concepts utiles. Les concepts retenus à partir de la première version et ne figurant pas après (12 concepts) seraient perdus dans un processus non itératif basé sur la version finale. L'équipe a donc décidé de garder tous les concepts (49 au total) retenus dans toutes les phases et de rajouter seulement 9 concepts ne figurant pas dans les suggestions pour effectuer leurs recherches d'auteurs potentiels de solutions et concevoir des campagnes de publicité en ligne pour diffuser le problème.

5.6.3. Exemples de suggestions

Afin d'illustrer l'utilité des concepts fournis par notre système, nous présentons ici quelques exemples de suggestions correspondant aux problèmes qui ont été postés sur hypios.com.

Tableau 11 Exemple de suggestions hyProximity 1

Concepts Found in the Problem	Concepts Suggested by hyProximity	
Refractive index	Nucleosol	Electromagnetic wave equation
Transparency and translucency	Quantum electrodynamics	Scattering
Wavelength	Phase (matter)	Dirac equation
Electromagnetic radiation	Equation of state	Heat capacity
	Emission spectrum	Superinsulator
	Polarization (waves)	Reflection (physics)
	Stimulated emission	Johari-Goldstein relaxation
	Viscosity	Fermion
	Electron	Superfluid

Le premier exemple concerne un problème lié à l'indice de réfraction du verre. Le défi était de trouver une manière innovante de changer l'indice de réfraction et de le maintenir changé sans qu'un recours permanent à l'électricité soit nécessaire. Dans le Tableau 11, nous présentons les concepts initiaux ainsi que les concepts trouvés par hyProximity. Parmi les suggestions, nous trouvons le concept Nucleosol. Si à première vue ce concept biologique ne parait pas très pertinent, une lecture attentive de sa description sur Wikipedia révèle que cette partie de cellule des organismes vivants a des propriétés électromagnétiques très intéressantes, qui pourront potentiellement inspirer une solution. Si la solution venait d'un biologiste, ce ne serait pas la première fois que la conception des systèmes technologiques s'inspirerait de la nature. D'autres concepts latéraux tels que : *Quantum electrodynamics*, *Fermion* et *Viscosity* figurent également dans la liste. Finalement, des concepts directement liés sont également d'une utilité non négligeable. Des concepts très spécifiques comme *Dirac equation* pourront servir pour la recherche d'experts utilisant un vocabulaire spécifique dans leurs publications, et des concepts plus généraux comme *Reflection* pourront servir dans la conception des

sein de l'équipe en charge des relations avec les futurs auteurs de solutions. Cette approche systématique nous a permis d'effectuer les observations présentées ici.

campagnes de publicité en ligne pour attirer un public large et potentiellement pertinent.

L'autre exemple de problème concernait la réduction de la consommation de l'électricité par les ménages à l'aide d'incitations innovantes. La description du problème était particulièrement mal formulée par rapport à la nature du problème. Elle concernait plutôt l'électricité et sa diffusion, alors que les parties relatives aux motivations pour la réduction de la consommation, au cœur du problème, étaient sous-représentées.

En dépit de cet obstacle, notre système, favorisant des concepts proches de plusieurs concepts de départ, a réussi à trouver des concepts à l'intersection des notions de l'électricité et de la motivation. Ces concepts concernaient notamment l'économie et les théories de la motivation des acteurs économiques. Ce problème a attiré 22 solutions de domaines divers. Satisfaite, l'entreprise cliente a continué à échanger avec un auteur de solution ayant un profil multidisciplinaire.

Tableau 12 Exemple de suggestions hyProximity 1

Concepts Found in the Problem	Concepts Suggested by hyProximity	
Electricity consumption	Virtuous circle and vicious circle	Convex preferences
Energy consumption	Household electricity approach	Lexicographic preferences
Incentive	Wealth	Opportunity cost
Electricity	World energy resources and consumption	Ac/dc (electricity)
Solution	Mesoeconomics	Indifference map
End-user	Foundations of economic analysis	Direct current
Finance	Epstein-zin preferences	Indifference curve
Consumption (economics)		Preference (economics)
Electrical power industry		

5.7. Conclusions

Les traces utilisateurs sont reliées aux concepts. Ce couple d'élément s'avère être très utile dans les scénarios liés à l'innovation. Cependant, les traces, et par conséquent les concepts qui les décrivent, ne représentent qu'une partie du spectre de connaissances et d'intérêts d'un utilisateur. Le Web n'a pas « conscience » de la totalité des intérêts d'un internaute, d'où le besoin de créer la notion de proximité de concepts pour faire face à cette incomplétude. Dans ce chapitre nous avons présenté nos travaux relatifs à la conception d'un rapprochement entre des concepts. Ils ont été testés lors de la diffusion de problèmes Hypios, de la recherche d'experts, etc. Nos travaux introduisent le principe de traitement différent de liens dans le graphe en fonction de leurs types. Ces types de liens sont définis à la fois par la propriété utilisée pour créer un lien dans le graphe, et par sa structure dans le graphe. Par ce dernier critère, nous avons établi une hiérarchie transversales de liens – i.e une arborescence de concepts, établie de manière directe. Nous avons proposé plusieurs méthodes alternatives pour quantifier l'importance de ces liens, dont la fonction de proximité, baptisée hyProximity. Grâce à un « étalon-or » construit à partir des données historiques d'Hypios, nous avons pu comparer nos différentes méthodes alternatives. Si l'utilisation de liens transversaux donne les

meilleurs résultats en termes de rappel, une fonction mixte utilisant les liens hiérarchiques et transversaux de manière différente apporte de meilleurs résultats en terme de précision.

Nous avons ensuite comparé notre approche aux approches issues de l'état de l'art dont : (1) Google Adwords Keyword Tool, un outil de suggestion de mots-clés dans le cadre de diffusion d'annonces sur le Web et (2) Random Indexing, une approche de proximité statistique, appliquée sur les graphes de données liés. Les études réalisées à l'aide de l' «étalon-or » ont montré une supériorité de notre approche. Les études utilisateurs nous ont permis d'effectuer des comparaisons plus détaillées en comparant à la fois la pertinence de concepts suggérés, et l'aspect inattendu, très sollicité dans les scénarios de l'innovation ouverte. Si en termes de performance notre approche s'est montrée très compétitive, c'est en termes de pertinence des suggestions inattendues qu'il a montré une supériorité importante. C'est donc grâce aux plusieurs évaluations de notre approche, et à l'effet synergique de celles-ci, que nous avons conduites que nous avons pu nous assurer d'une utilité réelle de nos recherches.

Nous avons implémenté notre approche sous forme d'une application Web. Elle a été utilisée dans le cadre de la prestation d'Hypios afin de trouver des nouveaux concepts pour orienter la recherche d'experts. Ce travail a également donné lieu au dépôt d'un brevet européen.

L'utilisation continue de cet outil au sein des équipes d'Hypios nous a permis d'effectuer plusieurs études de cas et d'analyser la vraie utilité de notre outil. Aujourd'hui encore, il continue à fasciner ses utilisateurs en leur proposant de concepts inattendus mais pertinents.

Bibliographie

[Brin & Page, 1998] Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In: *Seventh International World-Wide Web Conference (WWW 1998)*, April 14-18, 1998, Brisbane, Australia.

[Burton-Jones et al., 2003] Burton-Jones, A., Storey, V., Sugumaran, V., & Purao, S. (2003). A heuristic-based methodology for semantic augmentation of user queries on the web. *Conceptual Modeling-ER 2003*, 476–489. Springer. Retrieved January 19, 2011, from http://www.springerlink.com/index/TP1URDMGDM3F0WP3.pdf.

[Cilibrasi & Vitanyi, 2007] Cilibrasi, R. L., & Vitanyi, P. M. B. (2007). The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3), 370-383. doi: 10.1109/TKDE.2007.48.

[Cohen, 2008] Cohen, T.: Exploring medline space with random indexing and Pathfinder networks. Annual Symposium proceedings / AMIA Symposium. *AMIA Symposium* pp. 126-130 (2008)

[Damljanovic et al., 2011] Damljanovic, D., Petrak, J., Lupu, M. et al. (2011) Random Indexing for Finding Similar Nodes within Large RDF graphs. . In *The Semantic Web: ESWC 2011 Workshops, Workshops at the 8th Extended Semantic Web Conference*, ESWC 2011, Heraklion, Greece, May 29-30, 2011, Revised Selected Papers.

[Deerwester et al., 1990] Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.: Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41, pp.391-407 (1990)

[Fano, 1961] Fano, R M (1961), Transmission of Information: A Statistical Theory of Communications, MIT Press, Cambridge, MA (Chapter 2)

[Gracia & Mepa, 2008] Gracia, J., & Mena, E. (2008). Web-Based Measure of Semantic Relatedness. *In Proceedings of the 9th international conference on Web Information Systems Engineering* (WISE '08), pp.136-150.

[Jaccard, 1901] Jaccard, P. (1901). Etude Comparative de la Distribution Florale dans une Portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles* 37 547579, 37, 547-579.

[Jiang & Conrath, 1997] Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *In Proceedings of International Conference on Research in Computational Linguistics*, Taiwan.

[Karlgren & Sahlgren, 2001] Karlgren, J., Sahlgren, M.: From words to understanding. In: Uesaka, Y., Kanerva, P., Asoh, H. (eds.) *Foundations of Real-World Intelligence*, pp. 294{308. Stanford: CSLI Publications (2001)

[Kleinberg, 1999] J. Kleinberg. Authoritative sources in a hyperlinked environment. J. ACM, 46(5):604–632, 1999.

[Leacock & Chodorow, 1998] Leacock, C., Chodorow, M.: Combining local context and wordnet similarity for word sense identification. In Fellbaum, C., ed.: WordNet: *An electronic lexical database*. MIT Press (1998) 265–283

[Lehamnn et al., 2007] Lehmann, J., Schuppel, J., & Auer, S. (2007). Discovering Unknown Connections – the DBpedia Relationship Finder. *Proceedings of the 1st SABRE Conference on Social Semantic Web* (2007).

[Macdonald & Ounis, 2007] Macdonald, C., & Ounis, I. (2007). Expertise drift and query expansion in expert search. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management - CIKM '07*, 341. New York, New York, USA: ACM Press. doi: 10.1145/1321440.1321490.

[Matos et al., 2010] Matos, S., Arrais, J. P., Maia-Rodrigues, J., & Oliveira, J. L. (2010). Concept-based query expansion for retrieving gene related publications from MEDLINE. *BMC bioinformatics*, 11, 212. doi: 10.1186/1471-2105-11-212.

[Olivier & Senellart, 2007] Ollivier, Y., & Senellart, P. (2007). Finding related pages using Green measures: An illustration with Wikipedia. *Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI 2007)*. Vancouver, Canada. Retrieved from http://www.aaai.org/Papers/AAAI/2007/AAAI07-226.pdf

[Passant, 2010] Passant, A. (2010). dbrec — Music Recommendations Using DBpedia. In P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, et al. (Eds.), *Proceedings of the 9th International Semantic Web Conferene - ISWC 2010* (Vol. 1380, pp. 1-16). Shanghai, China: Springer Verlag.

[Rada et al., 1989] Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics 19 (1989)* 17–30

[Resnik, 1995] Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. *In: Proceedings of the 14th International Joint Conference on Artificial Intelligence*. Volume 1. (1995) 448–453

[Safar et al., 2004] Safar, B., Kefi, H. & Reynaud, C (2004). OntoRefiner, a user query refinement interface usable for Semantic Web Portals. *Proceedings of Application of Semantic Web technologies to Web Communities,* Workshop ECAI'04 (pp. 65-79).

[Sahlgren, 2005] Sahlgren, M.: An introduction to random indexing. In: Methods and Applications of Semantic Indexing Workshop at the *7th International Conference on Terminology and Knowledge Engineering*, TKE 2005.

[Salton, 1983] Salton, G.,McGill, M.J. (1983) *Introduction to modern information retrieval*,

[Sigurbjörnsson & Zwol, 2008] Sigurbjörnsson, B., & Zwol, R. van. (2008). Flickr tag recommendation based on collective knowledge. *Proceeding of the 17th international conference on World Wide Web - WWW '08*, 327. New York, New York, USA: ACM Press. doi: 10.1145/1367497.1367542.

[Strube & Ponzetto, 2006] Strube, M., & Ponzetto, S. P. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. *Proceedings of the National Conference on Artificial Intelligence (Vol. 21, p. 1419)*. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999. Retrieved from http://www.aaai.org/Papers/AAAI/2006/AAAI06-223.pdf

[Sussna, 1993] Sussna, M. (1993). Word sense disambigua- tion for free-text indexing using a massive semantic network. *In Proceedings of the Second International Conference on Information and Knowledge Management (CIKM-93),*Arlington, Virginia.

[Waitelonis & Sack, 2009] Waitelonis, J., & Sack, H. (2009). Towards Exploratory Video Search Using Linked Data. 2009 *11th IEEE International Symposium on Multimedia*, 540-545. Ieee. doi: 10.1109/ISM.2009.111.

[Wu & Palmer, 1994] Wu, Z., Palmer, M.(1994): Verb semantics and lexical selection. In: *Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics*. 133–138

[Zetch & Gurevych, 2007] Zesch, T., & Gurevych, I. (2007). Analysis of the Wikipedia Category Graph for NLP Applications. *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT 2007)*. Retrieved from http://acl.ldc.upenn.edu/W/W07/W07-02.pdf#page=11

6. Rapprochement de traces utilisateurs

Ce chapitre traite le rapprochement des traces utilisateurs entre elles. Plus précisément, il aborde la question de mise en correspondance de traces à faible contenu textuel (e.g. des tweets) avec des traces plus riches comme par exemple les articles de recherche relevés dans les tweets. Le scénario général est présenté dans la partie §5.1. La partie §5.2 présente des travaux liés tandis que la partie §5.3 présente une approche générale pour la mise en correspondance des tweets avec les articles. Pour finir, la partie §5.4 apporte un exemple d'application des alignements tweet-article qui démontre l'utilité des rapprochements pour des scénarios liés à l'innovation.

6.1. Introduction

6.1.1. Motivation générale

Le rapprochement de contenus Web est utile dans plusieurs situations. D'une manière générale, de nouveaux liens entre des éléments du Web permettent de mieux contextualiser un élément et d'avoir une base plus riche pour l'extraction et l'analyse de sens qui peut y être associée. Pour la tâche de l'annotation automatique, par exemple, l'appui sur ces nouveaux liens pourrait favoriser à la performance de l'approche. Un exemple de cet effet favorable est certainement l'étude de Abel et al. [Abel et al., 2011] qui montre la supériorité d'une approche d'annotation des tweets par des concepts lorsque celle-ci prend en compte les liens qui s'y trouvent. Ces liens pourront également permettre de mieux guider la navigation de l'utilisateur dans un espace informationnel pertinent pour une situation donnée. Par analogie, en opérant un lien entre un

utilisateur, expert potentiel, et tous les articles et projets qu'il a réalisés, il est possible d'envisager une optimisation de la recherche d'experts.

6.1.2. Le cas de tweets et d'articles scientifiques

Si la mise en correspondance de contenus Web est généralement utile, elle s'avère presque indispensable dans le cas du traitement de tweets, notamment pour le traitement automatique de leur sens. Ces billets de très petite taille, souvent inférieure à 140 caractères, sont particulièrement pauvres en contenu, ce qui rend très difficile leur annotation sémantique et l'extraction de concepts. L'importance croissante de ce type de ressources souligne l'ampleur du défi qu'impose leur traitement, tout d'abord en raison de la quantité croissante de données publiées sous cette forme. À lui seul, Twitter enregistre 200 millions de tweets par jour[112]. La seconde raison expliquant l'importance de cette source pour nos cas d'utilisation liés à l'innovation réside dans l'usage du microblog pour la communication scientifique et professionnelle, phénomène de plus en plus repéré dans la littérature. Reinhart et al. [Reinhart et al., 2009] étudient l'usage de Twitter lors des congrès scientifiques. Letierce et al. [Letierce et al., 2010] présentent une étude d'utilisation de Twitter pour la dissémination des résultats de recherche et la communication scientifique dans la communauté informatique. Veletsianos [Veletsianos, 2011] témoigne des pratiques d'échange sur Twitter employées par les enseignants. L'usage de Twitter lors de conférences est également un phénomène qui attire de plus en plus d'attention dans la communauté de recherche. En 2010, nous avons mis en évidence l'utilité des tweets créés lors de conférences pour plusieurs scénarios, tels que la création de profils utilisateurs, la recherche d'experts, et le suivi de la popularité de certains sujets, relayés par des présentateurs et articles lors d'évènements scientifiques [Stankovic et al., 2010]. D'autres auteurs [Weller et al., 2011] ont également souligné les possibilités d'exploiter la richesse des références présentes dans les tweets liés aux conférences. C'est cette richesse de références aux entités du monde physique ou virtuel qui rend ce type de microposts particulièrement apte à être mise en correspondance avec d'autres entités sur le Web.

Motivés par ces deux raisons nous nous focaliserons sur Twitter et notamment sur la mise en correspondance de tweets et d'articles académiques, tout en espérant que celle-ci aiderait à mieux traiter les traces d'une nouvelle facette d'échanges scientifiques ayant lieu dans le Web Social, pour l'instant assez peu abordée dans la littérature. Nous espérons également que les techniques et méthodes développées constitueront un bon exemple pour inspirer des rapprochements d'autres types de ressources sur le Web.

Lors des congrès, les tweets des utilisateurs peuvent concerner une multitude d'entités du monde réel ou du Web. Parfois un tweet contient une appréciation d'une présentation, du présentateur, de ces transparents et parfois il concerne le contenu de la présentation souvent concrétisée sous la forme d'un article. On peut aisément imaginer une mise en correspondance de tous ces éléments, à condition qu'ils aient une représentation dans le graphe Web. Il est pourtant difficile, dans le graphe Web actuel, de trouver des descriptions de toutes ces entités, et les articles représentent généralement la trace la plus communément trouvable. Pour cette raison nous visons à lier un tweet à un article. Il est important de noter que la description de l'article ici

[112] D'après le blog officiel de Twitter du 30 Juin 2011. http://blog.twitter.com/2011/06/200-million-tweets-per-day.html

représente une référence à la fois à l'article lui-même et à l'évènement associé avec sa présentation.

L'utilité de la mise en correspondance proposée serait multiple en permettant de :

- **Recueillir de l'information sur la réception des articles par le public du congrès**. En récupérant tous les tweets qui concernent un article et sa présentation dans le congrès, il est possible d'analyser la popularité de l'article et l'attention qu'il a attiré. Si la mesure de cette popularité peut être utile pour analyser la capacité d'un auteur à influencer sa communauté de recherche, elle peut également servir pour calculer la popularité de certains sujets au sein de la communauté et pour étudier des tendances. Une analyse supplémentaire des opinions exprimées dans des tweets pourrait donner lieu à des analyses encore plus profondes de la popularité des articles.
- **Recueillir de l'information sur les intérêts des utilisateurs de Twitter, experts potentiels**. Connaitre les présentations d'articles auxquelles un utilisateur a assisté permet de construire une image plus détaillée de ses centres d'intérêt. Ces articles sont beaucoup plus riches en contenu que les tweets et permettent aux extracteurs de concepts d'atteindre un fonctionnement supérieur de celui que l'on obtient avec le texte d'un tweet. Une telle mise en correspondance pourrait également, au fil du temps, permettre de surveiller la dynamique des intérêts des utilisateurs et cerner leurs intérêts les plus récents, et donc de faire émerger les thématiques sur lesquelles ils sont susceptibles de travailler à l'avenir. Elle pourrait également servir d'aide-mémoire pour permettre aux chercheurs de facilement retrouver des articles qui étaient leur cible lors des congrès.

Ces deux types d'informations qui émergent en conséquence de la mise en correspondance proposée permettront ensuite de tirer des conclusions sur l'expertise des auteurs des articles et potentiellement de celle des participants du congrès.

Nos travaux s'inscrivent dans une tendance de recherche qui essaye de tirer une nouvelle utilité de ce media. Dans cette même branche d'approches nous trouvons des travaux exploitant Twitter pour analyser des opinions politiques [Šķilters et al., 2011] et pour personnaliser les recommandations d'articles de presse [Abel et al., 2011a], pour ne citer que quelques exemples.

6.2. Travaux liés

Dans cette partie nous nous intéresserons notamment à deux catégories de travaux liés. D'une part, nous étudierons des travaux d'enrichissement de tweets par annotation de concepts sur des tweets, ce qui est en quelque sorte un objectif très proche de celui que nous envisageons. D'autre part, nous nous intéresserons aux travaux de mise en correspondance de tweets avec d'autres types d'objet Web. Ces travaux appliquent une idée similaire, potentiellement utile dans d'autres cas d'utilisation que celui de l'innovation sur le Web.

6.2.1. Annotation de tweets par des concepts

L'annotation de tweets a souvent pour objectif de compenser la pauvreté de leur contenu informationnel. Pour rendre un tel billet plus utile pour une multitude d'usages

possibles, il est notamment intéressant de fournir des éléments liés au contexte de la création du billet (utilisateur, localisation, temps, appareil utilisé, etc.) ainsi que de fournir des concepts étant en relation avec les tweets.

Certaines approches proposent la création de ces métadonnées au moment de la rédaction du billet. C'est le cas de SMOB [Passant et al., 2010], un système de microblogging sémantique qui publie le contenu des tweets avec les métadonnées associées directement dans le format RDF. Ce système à source ouverte crée de manière automatique les métadonnées conformes à l'ontologie SIOC et propose à l'utilisateur d'associer des tags présents dans les tweets avec les concepts DBPedia qu'ils sont susceptibles de représenter. Laisser le choix aux utilisateurs permet notamment de désambigüiser le sens de certains tags, dont la signification pourrait varier selon le contexte ; une tâche pas évidente pour des systèmes automatiques.

Malgré ces nombreux avantages, SMOB compte relativement peu d'utilisateurs, surtout en comparaison avec Twitter[113] qui est devenu la centrale globale d'échange des tweets. L'usage des tweets exige donc une annotation *a posteriori* de leur publication. Du fait de la taille du contenu des tweets et du jargon spécifique qui émerge en conséquence naturelle de la taille limitée des phrases, l'annotation par concepts s'avère un défi majeur que de nombreuses approches tâchent d'affronter.

C'est dans ce sens que Gentile et al. [Gentile et al., 2011] ont présenté une étude visant à évaluer la capacité des différents services existant à reconnaître les entités nommées dans des textes de petite taille. Cette étude montre que la taille du billet est critique quant aux performances de la plupart des systèmes d'analyse. En effet, un billet limité à 140 caractères donne des résultats médiocres alors que des textes d'une taille supérieure à 280 caractères et plus révèlent de meilleurs résultats. Ce même défi d'extraction à partir des tweets a fait l'objectif d'une compétition TREC Microblog Track[114]. La tâche de construire un système de recherche de tweets a attiré des propositions de plusieurs équipes de chercheurs. Des mesures de pertinence des tweets par rapport à un mot ou un thème s'imposaient bien évidemment comme indispensables pour un tel système. Nous présenterons par la suite quelques approches proposées à la compétition.

Cao et al. [Cao et al., 2011] ont proposé une mesure de pertinence de tweets par rapport à une requête (ensemble de mots) basée sur la mesure BM25[Robertson et al., 1994], une mesure couramment utilisée dans le classement des résultats de recherche Web qui s'inspire fortement de la mesure TF-IDF [Salton, 1983]. Cette mesure tient compte de l'importance d'un mot par rapport au tweet dans lequel il se trouve ainsi que de la probabilité de son occurrence dans un ensemble de tweets. Une approche relativement similaire est utilisée dans [Ferguson et al., 2011] dans le cadre de la même compétition. Mccreadie et al. [Mccreadie et al., 2011] construisent une approche d'apprentissage automatique pour apprendre une fonction de pertinence de tweets par rapport à un sujet. Leur modèle se sert de la pertinence calculée par la mesure BM25 ainsi que la qualité du tweet i.e. la probabilité qu'il soit un spam. Ces approches, développées pour la

[113] Twitter compte 100 millions d'utilisateurs actifs en date du 9/9/2011. http://www.washingtonpost.com/blogs/faster-forward/post/twitter-hits-100-million-active-users/2011/09/09/gIQAs9QjEK_blog.html

[114] https://sites.google.com/site/trecmicroblogtrack/

recherche de tweets, pourront bien évidemment être utilisées pour annoter des tweets par des sujets pour lesquels ils sont pertinents.

6.2.2. Différents travaux de mise en correspondance de tweets et autres ressources

Plusieurs approches de mise en correspondance de tweets avec d'autres entités ont été proposées dans la littérature, certaines même inspirées de nos travaux présentés en 2010 [sdow2010].

Choudhury et Breslin [Choudhury & Breslin, 2011] proposent un système de mise en correspondance de tweets avec des micro-événements, notamment lors d'une compétition sportive. Cette mise en correspondance peut ensuite servir pour effectuer différentes analyses ou bien pour construire de nouvelles manières de suivre les matches à l'aide de tweets associés. Une approche de classification par apprentissage automatique est ainsi mise en place, en s'appuyant sur des attributs tels que la mention des joueurs, la présence d'un tag, et le contexte du tweet. De la même manière, une approche d'apprentissage automatique proposée par Abel et al [Abel et al., 2011] permet de mettre en correspondance des tweets avec des incidents rapportés par des services de sûreté publique aux Pays-Bas et d'en construire un système de surveillance de la réaction publique face à différents incidents.

Si ces approches mettent en évidence une méthode de mise en correspondance qui semble assez généralisable, ils ne traitent pas le cas de mise en correspondance des tweets produits lors de congrès et des articles scientifiques, qui est notamment utile pour la détection des sujets d'intérêt des experts potentiels utilisant Twitter lors des congrès.

6.3. L'approche RATED

Dans cette partie nous présenterons notre approche générale de mise en correspondance de tweets et d'articles de conférence appelée RATED (*Rich Activity Twitter Event Data*). Notre démarche d'enrichissement de tweets commence par l'identification d'un évènement, souvent un congrès, qui fait l'objet du traitement. Ensuite il est nécessaire de récupérer des tweets qui le concernent. Ceci est largement facilité par la pratique d'usage des #hashtags (labels précédés par le signe # inclus dans les tweets) identifiant les évènements. Par exemple les auteurs des tweets associés avec la conférence International Semantic Web Conference 2010 incluront souvent le hashtag « #iswc2010 » dans leurs billets. Notre approche se décline ensuite en trois phases, représentées à la Figure 56.

Figure 56 Production de données liées à partir de tweets et articles

Dans la première phase, des tweets sont transformés en format RDF et mis à disposition en tant que données liées. Ces données sont ensuite enrichies (phase 2) via un extracteur de concepts. Finalement, (phase 3) les tweets enrichis sont mis en

correspondance avec des articles. Une condition préalable au fonctionnement de notre approche est la disponibilité des résumés d'articles ou d'autres descriptions du contenu des présentations qui peuvent être utilisés pour la mise en correspondance. Nous détaillons chacune de ces étapes ci-après.

6.3.1. Transformation de tweets en RDF

Pour effectuer la transformation des tweets en format RDF nous avons créé un programme Java chargé de cette transformation et de leur mise à disposition dans un entrepôt de triplet (nous avons choisi une base de triplets en ligne – Talis[115]). Dans les phases suivantes nous allons faire appel à ce même entrepôt pour récupérer des tweets. Il est important de souligner que ce choix architectural–permet une séparation de fonctionnalités en phases qui crée une modularité du système. Grâce à cette modularité, les phases ultérieures pourront fonctionner même si la manière de transformer des tweets en RDF change.

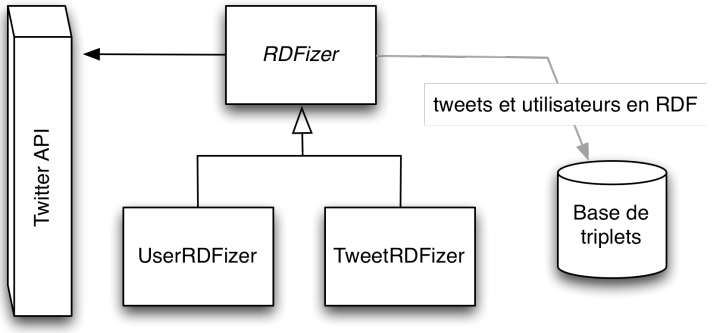


Figure 57 Structure du système pour la transformation des tweets en RDF

Dans notre pratique, nous avions en effet été amenés à modifier le fonctionnement du module pour la transformation des tweets suite à un changement majeur dans la politique d'usage officielle de Twitter. L'accès aux tweets par l'API de Twitter étant limité à des tweets très récents (uniquement sur les quelques derniers jours), nous nous étions initialement appuyés sur des services d'archivage de tweets, tels que TwapperKeeper[116] qui gardaient des archives de tweets correspondant aux hashtags fournis par des utilisateurs. Notre système à l'époque utilisait ces archives comme point de départ dans la transformation de tweets. Entre temps, suite au changement de politique de Twitter, la diffusion des archives des tweets était devenue illégale. Pour faire face à ces nouvelles circonstances, nous avons changé le module de récolte de tweets en implémentant une communication directe avec l'API Twitter. La nouvelle version de notre système permet donc de récupérer des tweets correspondant à un congrès uniquement dans les jours qui suivent le congrès. L'architecture d'haut niveau de notre module est présentée sur la Figure 57. Une classe (de type RDFizer) est créée pour chaque type d'objet i.e. pour des profils d'utilisateurs (UserRDFizer) et pour des tweets (TweetRDFizer). Lors du traitement des tweets, un TweetRDFizer est appelé pour chaque tweet. Ce module utilise notamment des ontologies SIOC et OPO[117]. Un exemple de tweet dans le format RDF est présenté dans la Figure 58. L'UserRDFizer n'est appelé que si l'utilisateur, auteur du tweet en cours de traitement, n'est pas déjà

[115] http://docs.api.talis.com/about

[116] http://twapperkeeper.com/index.html

[117] http://online-presence.net/ontology.php

traité est sauvegardé dans la base de triplets. L'ontologie FOAF est utilisée pour représenter en RDF les différents aspects de profil utilisateur.

Grâce au fait que les phases successives de notre approche utilisent des tweets décrits par des ontologies standards et couramment utilisées, il est possible de substituer notre module de transformation de tweets avec d'autres convertisseurs, comme par exemple Twarql [Mendes et al., 2010] ou KASAS [Wagner, 2010] ou bien d'utiliser directement des tweets créés dans le format RDF par le client SMOB.

```
<rdf:Description rdf:about="http://data.Hypios.com/tweets/tweet-15162225891">
  <dc:description>A tweet :Noshir Contractor at #eswc2010 speaking of data-
  driven social network analysis of MMORPG such as WoW and EQ2</dc:description>
  <rdf:type rdf:resource="http://online-presence.net/opo/ns#OnlinePresence"/>
    <opo:customMessage>
      <sioct:MicroblogPost
        rdf:about="http://data.Hypios.com/tweets/tweet-15162225891-cm">
        <sioc:content>Noshir Contractor at #eswc2010 speaking of data-driven social network
analysis of MMORPG such as WoW and EQ2</sioc:content>
        <sioc:id>15162225891</sioc:id>
        <dcterms:language>en</dcterms:language>
        <foaf:maker rdf:resource="http://data.Hypios.com/tweets/user-ciro"/>
        <dcterms:date>2010-06-01T09:16:46+0200</dcterms:date>
        <dcterms:subject
            rdf:resource="http://dbpedia.org/resource/Social_network"/>
      </sioct:MicroblogPost>
```

Figure 58 Un exemple de tweet dans le format RDF

6.3.2. Enrichissement par concepts

L'extraction de concepts à partir des éléments textuels des tweets est confiée à des services externes. L'architecture de notre module Java responsable de cette extraction repose sur les mêmes principes de flexibilité implémentés dans la séparation de système en modules. Une interface ConceptExtractor est mise en place en tant qu'abstraction de la fonctionnalité effectuée par ce module, quel que soit le service externe appelé pour la mettre en œuvre. Des classes de cette interface prennent des tweets déjà présents dans l'entrepôt de triplets, effectuent une extraction de concepts à partir des parties textuelles et réintroduisent dans l'entrepôt des nouveaux triplets ayant pour objectif de relier des tweets avec des concepts. Pour effectuer cette liaison nous avons utilisé la propriété dct :subject, souvent utilisée pour désigner des concepts.

Nous avons implémenté deux versions concrètes de ConceptExtractor : ZemantaExtractor et SpotlightExtractor. Le premier fait appel à l'API de Zemanta[118], un service d'annotation des textes par des liens pertinents, pour traiter le texte et en extraire des concepts. Zemanta propose, pour un texte donné, des liens vers différentes ressources multimodales, articles de journaux et blogs, pertinentes pour le texte ainsi que des pages Wikipedia correspondant aux notions directement mentionnées dans le texte. Nous n'en gardons que les liens Wikipedia, pour les transformer en URIs des concepts DBPedia correspondants. Le deuxième extracteur repose sur DBPediaSpotlight, un service proposé par DBPedia qui permet d'annoter le texte avec

[118] http://www.zemanta.com/api/

des concepts DBPedia. Il est possible de changer à tout moment l'extracteur utilisé par une simple modification du fichier de configuration.

Dans notre expérience avec les deux APIs nous avons remarqué que Zemanta donne des résultats beaucoup plus satisfaisants en termes de quantité de tweets dans lesquelles il peut identifier des concepts. Lors d'une expérience, nous avons fait traiter les mêmes échantillons de tweets par les deux services. Le nombre de tweets où Zemanta a trouvé des concepts était 2 à 3 fois supérieur à celui de tweets où DBPedia Spotlight a trouvé des concepts. Si cette expérience n'avait pas la valeur d'une comparaison statistiquement signifiante, puisqu'elle avait surtout l'objectif d'entamer une première interrogation sur les performances respectives de ces deux systèmes, c'est l'étude menée par Rizo et Troncy [Rizzo & Troncy, 2011] comparant les performances d'une multitude de services d'extraction de concepts à partir du texte qui nous a confortés dans notre choix de Zemanta comme extracteur principal.

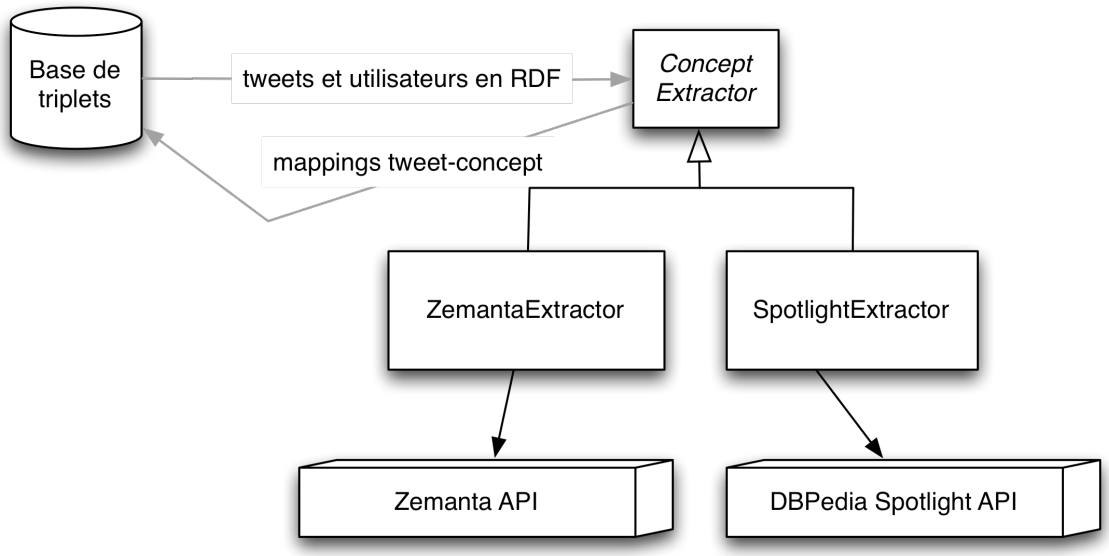


Figure 59 Structure du système pour l'enrichissement du contenu par concepts

6.3.3. Alignements tweet-article

La dernière phase de notre approche est l'alignement des tweets avec des articles. Afin de montrer qu'une telle mise en correspondance est une possibilité réelle et utile, nous avons conduit une expérience visant à produire, de manière manuelle, des liens entre des tweets associés avec la conférence ESWC 2010 (*European Semantic Web Conference*) et ses articles. Par la suite nous décrivons cette expérience et des travaux que nous avons menés en coopération avec Matthew Rowe d'Open University, UK pour mettre en place une approche d'alignement automatique par une classification à base d'apprentissage automatique. Enfin nous aborderons quelques scénarios d'usages potentiels rendus possibles par les nouveaux liens.

6.3.3.1. Alignement manuel

Pour effectuer des alignements manuels, nous avons d'abord récupéré des tweets correspondant à ESWC 2010 (hashtag « #eswc2010 ») qui étaient à l'époque disponibles sur TwapperKeeper. Ces tweets ont ensuite été traités par notre système d'enrichissement pour produire leurs descriptions sémantiques et les lier aux concepts

DBpedia. Un total de 1082 tweets a été traité, dont seulement 213 ont pu être liés à des concepts, ce qui a produit 252 liens tweet-concept. Les concepts les plus fréquemment trouvés dans des tweets sont présentés dans le Tableau 13.

Nous avons ensuite fait appel à trois participants de la conférence ESWC, utilisateurs de Twitter pour leur demander de mettre en correspondance manuellement un échantillon de 200 tweets avec des articles. L'expérience a été effectuée dans la semaine suivant la conférence et les participants de notre étude ont pu toujours se souvenir du contexte de publication des tweets particuliers, des discussions et autres éléments permettant d'effectuer l'alignement. Afin d'obtenir un alignement crédible il était nécessaire de s'assurer que les trois participants étaient d'accord sur leurs alignements. Le taux d'accord calculé par la métrique kappa, déjà présentée dans la partie §4.3.3.3, n'était pas suffisant dans le premier tour (*k* moyenne était 0.328). Nous avons donc dû procéder à un deuxième tour, permettant aux participants de voir de manière anonyme les alignements des autres et de changer leur opinion[119]. Après le deuxième tour le taux d'accord (k moyenne 0.820) montrait une convergence d'opinions qui nous a permis de prendre 89 paires tweet-concept en tant qu'alignements valides, sur lesquelles au moins 2 sur 3 participants de l'étude étaient d'accord. Les autres 111 tweets étaient soit liés à un évènement qui n'avait pas d'article correspondant (e.g. conférences invitées) soit n'étaient pas du tout liés à un évènement concret. Le Tableau 14 démontre une récapitulation des concepts les plus fréquemment trouvés dans les articles alignés avec des tweets.

Tableau 13. Les concepts les plus populaires dans les tweets correspondant à ESWC2010

Concept	*Nombre d'occurrences*
http://dbpedia.org/resource/Semantic_Web	30
http://dbpedia.org/resource/Linked_Data	24
http://dbpedia.org/resource/SPARQL	20
http://dbpedia.org/resource/Resource_Description_Framework	11
http://dbpedia.org/resource/Social_network	9
http://dbpedia.org/resource/Uniform_Resource_Identifier	8
http://dbpedia.org/resource/Radio-frequency_identification	7
http://dbpedia.org/resource/Hypertext_Transfer_Protocol	6

Tableau 14. Les concepts les plus populaires ESWC2010 dans les articles liés aux tweets

Concept	*Nombre d'occurrences*
http://dbpedia.org/resource/Data	15
http://dbpedia.org/resource/Semantics	13
http://dbpedia.org/resource/Semantic_Web	12

[119] Nous rappelons qu'une description complète de la méthodologie est donnée sur le site http://milstan.net/stankovic_rowe_methodology/

http://dbpedia.org/resource/Resource_Description_Framework	6
http://dbpedia.org/resource/Graph_%28mathematics%29	5
http://dbpedia.org/resource/DBpedia	5

En observant les deux tableaux il devient clair que des concepts trouvés dans les tweets sont légèrement différents de ceux observés dans les articles référencés par des tweets. Les concepts comme DBpedia ou Graph, apparaissent par exemple dans les articles, alors que dans les tweets ils semblent être peu fréquents. Une explication possible est que ces concepts sont peut-être sous-entendus dans les tweets.

Certains concepts sont fréquents dans les tweets mais rares dans les articles concernés par les tweets. C'est le cas du concept de « Social Network ». Cela peut être dû au fait que ce concept est considéré comme trop spécifique pour les textes scientifiques, dans lesquels « Social Web » est souvent préféré pour distinguer les différents sites sociaux, y compris les réseaux sociaux. Dans le langage parlé, ce concept reste néanmoins très utilisé, ayant même un sens plus général.

Ces variations de vocabulaire dans les différents types de traces (tweets et articles), non seulement concernant la terminologie utilisée mais aussi le choix de concepts et leur spécificité, impliquent que des mises en correspondance entre des tweets et des articles pourront effectivement être utiles pour la compréhension du sens des deux types d'entités.

6.3.3.2. Automatisation des alignements

En coopération avec Matthew Rowe d'Open University, UK nous avons construit une approche automatique de mise en correspondance des tweets et des articles basés sur une méthode de classification par apprentissage automatique. Une fois la mise en correspondance effectuée, nous utilisons la propriété lode :illustrate, définie dans l'ontologie *Linking Open Description of Events (LODE)*[120] pour introduire le lien entre le tweet et l'événement correspondant à la présentation de l'article.

Nous considérons le problème d'alignement comme un problème d'apprentissage automatique. Plus particulièrement, nous disposons d'un ensemble de tweets X et d'un ensemble d'articles identifiés par leur URIs Y. Cet ensemble Y peut en effet être considéré comme un ensemble de labels de catégories des éléments de X. Notre tâche devient donc de trouver une fonction qui assigne le label le plus approprié aux tweets f : X→Y. Pour construire notre approche, nous avons considéré deux techniques de construction de clusters: (1) clustérisation à base de mesures de proximité : (similaire à K-means [ref]), en utilisant des distances : Manhattan et euclidienne et (2) classification Naive Bayes. Nous avons également varié les différents ensembles d'attributs (en. *feature sets*) afin de comparer leur performance :

1. les éléments informationnels appartenant à la description de tweets (e.g., auteur, date, etc.)
2. les éléments informationnels trouvables dans le graphe Web et étant associés au tweet avec un lien à distance 2. Par exemple, si un tweet est associé à l'utilisateur

[120] http://linkedevents.org/ontology/

A qui en est l'auteur, et que celui-ci connait l'utilisateur B, nous prendrons l'utilisateur B comme un attribut de tweet.

3. Les concepts DBPedia associés avec le tweet après nos transformations.

Nous avons comparé la performance de toutes ces variations de l'approche de classification avec celle des alignements manuels créés préalablement (dont la création est décrite dans §6.3.3.1). La classification Naive Bayes a montré de meilleurs résultats.

Cette approche étant le résultat d'un travail collaboratif dans lequel nous n'avons participé que partiellement, nous ne la présentons pas en détails dans ce mémoire. Un lecteur intéressé pourra en trouver une description détaillée dans l'article [Rowe & Stankovic, 2011].

6.4. Exemple d'usage des alignements

Selon l'hypothèse disant que la création d'un tweet lié à un article est une déclaration implicite de l'intérêt porté à celui-ci, et indirectement de l'intérêt porté aux concepts associés à l'article, il est possible de se servir des alignements tweet-article pour créer des profils d'intérêt d'utilisateurs. Nous montrons ici un exemple d'un tel profil en le comparant avec un profil que l'on aurait créé en s'appuyant seulement sur des tweets.

Prenons l'exemple de l'utilisateur de Twitter @clauwa (Claudia Wagner) qui a assisté à ESWC 2010. Les tweets créés par Claudia lors du congrès ont permis l'extraction des concepts suivants :

http://dbpedia.org/resource/Semantic_Web
http://dbpedia.org/resource/Resource_Description_Framework

Après avoir effectué la mise en correspondance, nous avons pu identifier un nombre plus grand de concepts associés à des articles faisant l'objet de ses tweets :

http://dbpedia.org/resource/Ontology_(information_science)
http://dbpedia.org/resource/Semantic_Web
http://dbpedia.org/resource/Data
http://dbpedia.org/resource/Unsupervised_learning
http://dbpedia.org/resource/Tag_cloud
http://dbpedia.org/resource/Semantics
http://dbpedia.org/resource/Rule_of_inference
http://dbpedia.org/resource/Resource_Description_Framework
http://dbpedia.org/resource/Research
http://dbpedia.org/resource/Proprietary_format
http://dbpedia.org/resource/Logical_schema
http://dbpedia.org/resource/Gold_standard
http://dbpedia.org/resource/Folksonomy
http://dbpedia.org/resource/Flickr
http://dbpedia.org/resource/Biology

Nous avons ensuite vérifié avec Claudia que les concepts de cette deuxième liste représentaient en effet ses centres d'intérêt. Elle nous l'a affirmé pour tous les concepts à l'exception de la biologie. Bien que cette comparaison n'ait pas le poids d'une étude statistique rigoureuse, il apparait néanmoins clairement que la mise en correspondance ouvre de nombreuses possibilités pour la construction de profils. Celle-ci pourrait

notamment être affinée par une approche de prise en compte de l'importance des articles (en fonction de leur popularité, rareté d'occurrence dans les tweets ou autres paramètres).

6.5. Conclusions

Dans ce chapitre, nous avons présenté la notion de rapprochement de traces utilisateur. En raison de leur contenu pauvre en texte, certains types de traces, tels que des tweets, sont difficiles à exploiter. Cela rend ainsi l'identification des centres d'intérêt et sujets d'expertise très délicate. Il paraît donc opportun de les mettre en relation avec d'autres traces, plus riches en information. Prenons l'exemple des congrès scientifiques. L'utilisation de Twitter y est fréquente. Par conséquent, il arrive souvent que des tweets mentionnent des éléments, comme les articles présentés lors des congrès, qui font référence à d'autres traces sur le Web. Nous avons donc proposé une approche générale de mise en correspondance entre tweets et articles. Une implémentation concrète de cette approche a été effectuée en collaboration avec Matthew Rowe de KMi, Open University, Royaume Uni. Le système intègre des techniques d'apprentissage automatique. Nous avons pu montrer l'utilité de ce système à l'aide d'une étude de cas. Une preuve supplémentaire de sa pérennité fut donnée par l'analyse de travaux indépendants publiés après notre article [Stankovic et al., 2010]. De plus, les utilisations diverses de l'idée d'alignement des tweets et autres traces (suivi des événements sportives [Choudhury & Breslin, 2011], suivi des incidents [Abel et al., 2011], visualisations de blogs et tweets [Cameron et al., 2011], etc) fournit une preuve de sa généralité.

Bibliographie

[Abel et al., 2011] Fabian Abel, Claudia Hauff, Geert-Jan Houben, Richard Stronkman, Ke Tao. Twitcident: Fighting Fire with Information from Social Web Stream. *In Proceedings of International Conference on World Wide Web (WWW)*, Lyon, France, March 2011

[Abel et al., 2011a] Abel, F., Gao, Q., Houben, G. J., & Tao, K. (2011). Analyzing user modeling on twitter for personalized news recommendations. *User Modeling, Adaption and Personalization*, 1–12. Springer. Retrieved from http://www.springerlink.com/index/02815H2269QG3MV6.pdf

[Cameron et al., 2011] Cameron, D., Finlayson, A., & Wotzko, R. (2011). Visualising Social Computing Output: Mapping Student Blogs and Tweets. *Social Media Tools and Platforms in Learning Environments* (pp. 337-350). Springer. Retrieved from http://books.google.com/books?hl=en&lr=&id=qt5myODFp2sC&oi=fnd&pg=PA337&dq=Visualising+social+computing+output+:+Mapping+student+blogs+and+tweets&ots=BK0BXyNnre&sig=60DOB-zPsfNKd1WBkPZRjY8mMfE

[Cao et al., 2011] Cao, P., Gao, J., Yu, Y., Liu, S., Liu, Y., & Cheng, X. (2011). ICTNET at Microblog Track TREC 2011. *TREC 2011 Microblog Track* (pp. 1-4).

[Choudhury & Breslin, 2011] Choudhury, S., & Breslin, J. (2011). Extracting Semantic Entities and Events from Sports Tweets. *Proceedings of Making Sense of Microposts #MSM2011 (pp. 22-32)*. Retrieved from http://www.johnbreslin.org/files/publications/20110530_msm_2011.pdf

[Ferguson et al., 2011] Ferguson, P., Hare, N. O., Lanagan, J., Smeaton, A. F., Phelan, O., Mccarthy, K., & Smyth, B. (2011). *CLARITY at the TREC 2011 Microblog Track Temporal Re-weighting*. TREC 2011 Microblog Track (pp. 1-6).

[Gentile et al., 2011] Gentile, A. L., Cano, A. E., Dadzie, A.-S., Lanfranchi, V., & Ireson, N. (2011). Does Size Matter? When Small is Good Enough. *Proceedings of Making Sense of Microposts #MSM2011*. Retrieved from http://trec.nist.gov/pubs/trec20/papers/ICTNET.microblog.pdf

[Letierce et al., 2010] Letierce, J., Passant, A., Decker, S., & Breslin, J. (2010). Understanding how Twitter is used to spread scientific messages. In: *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*. Raleigh, NC: US. Retrieved from http://journal.webscience.org/314/2/websci10_submission_79.pdf

[Mccreadie et al., 2011] Mccreadie, R., Macdonald, C., Santos, R. L. T., & Ounis, I. (2011). University of Glasgow at TREC 2011 : Experiments with Terrier in Crowdsourcing , Microblog , and Web Tracks. *TREC 2011 Microblog Track*.

[Mendes et al., 2010] Mendes P.N., Kapanipathi, P., Passant A." Twarql: Tapping into the Wisdom of the Crowd,". Triplification Challenge 2010 at *6th International Conference on Semantic Systems (I-SEMANTICS),* Graz, Austria, 1-3 September 2010.

[Passant et al., 2010] Passant, A., Bojars, U., Breslin, J. G., Hastrup, T., Stankovic, M., & Laublet, P. (2010). An Overview of SMOB 2: Open, Semantic and Distributed Microblogging. *4th International Conference on Weblogs and Social Media, ICWSM* (pp. 303-306). Retrieved from http://swl.slis.indiana.edu/files/smob-icwsm-poster.pdf

[Reinhart et al., 2009] Reinhardt, W., Ebner, M., Beham, G., & C. (2009). How people are using Twitter during conferences. *5th EduMedia conference*, (March). Retrieved from http://www.edna.edu.au/edna/referral/browse/http://lamp.tu-graz.ac.at/~i203/ebner/publication/09_edumedia.pdf

[Rizzo & Troncy, 2011] Rizzo, G., & Troncy, R. (2011). NERD : Evaluating Named Entity Recognition Tools in the Web of Data. *Proceedings of the 11th Interational Semantic Web Conference ISWC2011*. Bonn, Germany.

[Robertson et al., 1994]Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, S. and Gatford, M. Okapi at TREC-3. *In Proceedings of the Third Text REtrieval Conference (TREC 1994)*. Gaithersburg, USA, November 1994

[Rowe & Stankovic, 2011] Rowe, M. and Stankovic, M. Aligning Tweets with Events: Automation via Semantics. *Semantic Web Journal, Special Issue on Real-time and Ubiquitous Social Semantics*. 2011. IOS Press

[Salton, 1983] Salton, G.,McGill, M.J. (1983) *Introduction to modern information retrieval*,

[Šķilters et al., 2011] Šķilters, J., Kreile, M., Bojārs, U., Brikše, I., Pencis, J., & Uzule, L. (2011). The Pragmatics of Political Messages in Twitter Communication. *Proceedings of Making Sense of Microposts #MSM2011* (pp. 69-68). Heraklion, Crete, Greece.

[Stankovic et al., 2010] Stankovic, M., Rowe, M., & Laublet, P. (2010). Mapping Tweets to Conference Talks: A Goldmine for Semantics. in *Proceedings of the 3rd Social Data on the Web Conferece, SDOW2010,* collcoated with International Semantic Web Conference ISWC2010. Shanghai, China.

[Veletsianos, 2011] Veletsianos, G. (2011). Higher education scholars' participation and practices on Twitter. *Journal of Computer Assisted Learning*,. doi:10.1111/j.1365-2729.2011.00449.x

[Wagner, 2010] Wagner, C. (2010). Exploring the Wisdom of the Tweets: Towards Knowledge Acquisition from Social Awareness Streams. *In Proceedings of Extended Semantic Web* Conference 2010, Heraklion, Crete.

[Weller et al., 2011] Weller, K., Droge, E., & Pushcmann, C. (2011). Citation Analysis in Twitter. Approaches for Defining and Measuring Information Flows within Tweets during Scientific Conferences. *Proceedings of Making Sense of Microposts #MSM2011* (pp. 1-12). Retrieved from http://files.ynada.com/papers/msm2011.pdf

Lösch, U., & Müller, D. (2011). Mapping Microblog Posts to Encyclopedia Articles. *Tagungsband Informatik* 2011. Retrieved from http://www.user.tu-berlin.de/komm/CD/paper/040121.pdf

7. Rapprochement de collaborateurs dans les scénarios d'innovation sur le Web

Dans ce chapitre, nous nous intéressons à un rapprochement entre des collaborateurs potentiels pouvant travailler ensemble sur la résolution d'un problème. Le succès d'une collaboration repose sur la complémentarité des individus en termes de compétences, mais également sur leur compatibilité et leur capacité à travailler ensemble. Pour affiner la recherche de collaborateurs, nous appliquons, en plus des critères précédents, un rapprochement de concepts. Ce travail représente donc une occasion d'expérimenter des rapprochements multiples. Dans la partie §7.1 nous présentons le contexte et les besoins particulièrs de notre cas d'utilisation. La partie §7.2 est dédiée à la présentation des approches issues de l'état de l'art. Dans la partie §7.3, nous étudierons en détail la méthode de rapprochement de collaborateurs. Elle est ensuite évaluée dans la partie §7.4, où nous nous intéressons plus particulièrement à l'impact du rapprochement de concepts sur la performance de notre méthode. Pour finir, nous illustrons dans la partie §7.5 ce concept de rapprochement à l'aide d'un exemple concret.

7.1. Rapprochement de collaborateurs

Le besoin de l'autre, de l'interaction, de la complicité dans le succès et dans l'échec s'entremêle avec de nombreux aspects de la vie humaine. C'est la communication qui permet aux idées de se développer, se propager, de devenir virales. C'est aussi la communication qui les fait s'affiner et évoluer. Dans le cadre de l'innovation et de la recherche, l'aspect socio-collaboratif de la création est d'une importance vitale et peut-être même plus important que dans la plupart des autres situations.

Deux tendances soulignent cette importance de l'aspect social i.e. de la collaboration dans l'innovation :

- **La nature pluridisciplinaire des défis de l'innovation et de la recherche**. L'aspiration à l'expertise, d'une importance capitale dans les processus industriels, exige une certaine spécialisation des compétences. De plus, la nature complexe des problèmes que l'industrie rencontre aujourd'hui nécessite d'une pluralité de perspectives et de savoirs dans leur résolution. La collaboration entre experts de plusieurs disciplines s'impose alors comme la solution optimale pour relever ce défi. Des études menées en sciences de gestion confirment l'impact positif de la multidisciplinarité d'équipes R&D sur les résultats du processus de l'innovation [Roure & Keely, 1990] [Johne & Snelson, 1988].
- **Les collaborations internationales dans la recherche**. Le monde de la recherche est de plus en plus ouvert à la collaboration. Le financement de la recherche publique par projet, imposé par les gouvernements et par la Commission Européenne, comme par exemple à travers le programme FP7[121], crée une ambiance très favorable à la formation d'équipes internationales ainsi qu'aux collaborations entre universités. Des analyses de tendances dans la publication scientifique [Jones, 2008] démontrent une domination parmi les articles acceptés dans des conférences et journaux, des articles écrits par des équipes d'auteurs provenant d'universités différentes.

Il est donc clairement utile d'envisager un rapprochement entre des personnes participant dans le processus d'innovation afin de mieux les soutenir dans leur mission, celle d'innovateurs. Dans ce chapitre nous proposons un rapprochement de potentiels (co)auteurs de solutions dans l'innovation sur le Web. Ce rapprochement est censé répondre au besoin d'un auteur de solutions dont les compétences répondent seulement en partie à celles requises par le problème d'innovation et qui souhaite découvrir dans un potentiel collaborateur les compétences complémentaires aux siennes.

Contrairement aux rapprochements qui ont fait l'objet des chapitres précédents, où il était toujours question de rapprocher deux entités de manière utile, le rapprochement présenté dans ce chapitre a un caractère trilatéral. Il est nécessaire de rapprocher un utilisateur initial d'utilisateurs candidats collaborateurs dans le contexte d'un problème d'innovation. La Figure 60 illustre la complexité des liens entre ces entités dans le graphe Web. Les utilisateurs, potentiels collaborateurs, sont liés entre eux et avec l'utilisateur initial par différents liens d'amitié, de travail et de connaissance. Grâce à des traces qu'ils laissent sur le Web, il est possible de les considérer comme étant également

[121] http://cordis.europa.eu/fp7/home_en.html

liés aux concepts différents. Le problème d'innovation est à son tour également lié à certains concepts, et les concepts sont bien évidemment aussi interconnectés.

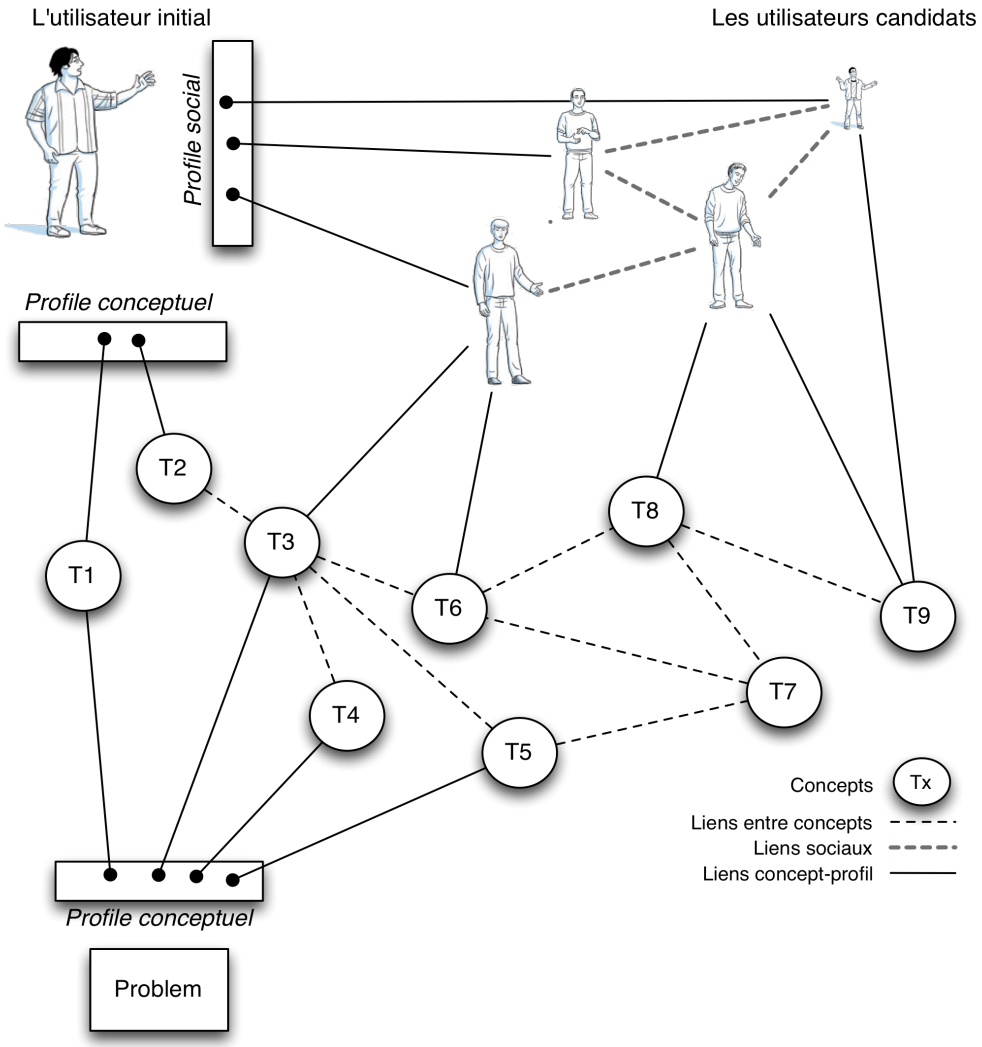


Figure 60 Exemple de sous-graphe Web pertinent pour le rapprochement de collaborateurs potentiels

Le défi que nous affrontons donc dans ce chapitre consiste à utiliser tous ces liens d'une manière intelligente, pertinente et utile pour effectuer de possibles rapprochements de collaborateurs dans le processus de l'innovation sur le Web.

7.1.1. Qualités désirées d'une approche de découverte de collaborateurs

Inspirés par le cadre général de la thèse aspirant à certaines qualités (e.g., la sérendipité), aussi bien que par la littérature existante révélant partiellement la nature de la collaboration humaine, nous déduisons quelques caractéristiques que notre approche devrait avoir afin d'être vraiment utile. Pour un problème donné l'utilisateur initial devrait, grâce à notre approche, pouvoir découvrir des collaborateurs, inconnus auparavant, qui :

- Sont complémentaires avec lui en terme de compétences pour affronter un problème ;
- Seront compatibles avec lui pour un travail d'équipe.

Il existe certaines études sociales permettant de mieux comprendre la préférence des membres d'une équipe dans un contexte de projet industriel pour une diversification des collaborateurs. Une telle étude [Hinds et al., 2000] démontre que la similarité (e.g., la similarité de centres d'intérêt) joue un rôle important dans le choix des collaborateurs aussi bien que la compétence et la complémentarité des candidats collaborateurs. Une autre étude [Huckman & Staats, 2010] confirme l'effet positif de la familiarité des membres d'équipe. Elle évoque des effets positifs de la différence de niveaux d'expérience des membres qui peut améliorer la robustesse d'une équipe et son adaptabilité au changement de tâches. Nous fondons notre approche sur ces résultats et nous nous en inspirons pour explorer d'autres caractéristiques potentiellement utiles pour le choix de collaborateurs.

7.2. Etat de l'art

S'il est très difficile de trouver une étude proposant le même type de rapprochement, la nature complexe de celui-ci rend nécessaire la prise en compte de nombreuses approches développées pour faire face aux différentes parties du puzzle.

Le rapprochement de concepts par exemple fait partie du rapprochement complexe dans le graphe présenté à la Figure 60. L'état de l'art général sur ce sujet est déjà présenté dans la partie §5.2.

Les approches de recherche d'experts qui s'appuient sur les hypothèses d'expertise H19 et H20 (basées sur des tweets), présentées dans la partie §4.2.1.1, seront notamment pertinentes par rapport à l'implémentation concrète et l'évaluation de notre approche, effectuée à l'appui de tweets.

Le rapprochement entre des personnes et leurs sujets de compétence à travers des traces a déjà fait l'objet du Chapitre 4, comprenant également un état de l'art.

Il reste donc dans ce chapitre à faire la revue de la littérature proposant des méthodes de recommandation de personnes dans différents scénarios, plus ou moins similaires au nôtre.

7.2.1. Recommandation de personnes

La recommandation de membres d'équipe n'est pas une idée neuve dans le monde de la recherche. Dans [Barreto et al., 2008] un système de sélection de membres d'équipe pour des projets de développement de logiciel est proposé. Ce système utilise une approche de satisfaction de contraintes et peut optimiser ses propositions par rapport aux différents critères e.g., le coût de travail, la compétence des membres etc. La compatibilité de membres n'a pas été considérée. Celle-ci prend une place plus importante dans des travaux liés à l'innovation ouverte, tels que l'approche théorique de suggestion d'équipes pour les projets d'innovation ouverte proposée dans [Brocco & Groh, 2009]. Le modèle théorique proposé identifie des paramètres à prendre en compte lors du choix des membres d'une équipe. Ces paramètres incluent : les caractéristiques des membres (compétence, paramètres démographiques,

psychologiques, sociaux), les caractéristiques intrinsèques de l'équipe (synergie, longévité, conflits), les caractéristiques externes à l'équipe (organisation, système de récompenses, etc.). Aucune réalisation concrète de ce modèle n'a été proposée par les auteurs.

Différentes mesures de similarité de personnes ont été proposées, même dans les travaux ne cherchant pas forcement à former des équipes. Ces mesures se distinguent notamment par (1) les sources d'information utilisées pour construire des profils utilisateurs dont la similarité sera mesurée, et (2) les mesures de similarité utilisées pour comparer des profils. Ces mesures font le plus souvent partie des mesures vectorielles déjà présentées en §5.2.2 Par exemple, une étude [Spertus et al., 2005] présente une comparaison de performances de 6 mesures de similarité de profils utilisateurs dans la recommandation des communautés en ligne intéressantes pour les membres du réseau social Orkut. Une mesure simple à base de normes de vecteurs et de leur intersection est montrée comme étant plus performante que des mesures probabilistes, telles que Pointwise Mutual Information, TF-IDF et LogOdds [Cornfield, 1951]. L'intuition derrière une telle approche de recommandation est que les utilisateurs aimeraient rejoindre des communautés dans lesquelles participent des personnes qui leur sont similaires. Une même logique est appliquée à d'autres systèmes essayant de recommander différents services ou produits en se basant sur la similarité de goût des utilisateurs. Le système de recommandation de films proposé dans [Luo et al., 2008], qui utilise une notion transitive de la similarité pour compenser le manque d'information et la pauvreté de profils utilisateurs, en est un bon exemple. Une étude présentée dans [Guy et al., 2010] compare la performance de différentes sources d'informations utilisées pour construire des profils utilisateurs (telles que : les intérêts, les contacts sociaux, la localisation des activités des utilisateurs) dans le cadre de calcul de similarités à l'aide de la mesure de Jaccard. Différentes sources montrent des performances variées selon les scénarios (recherche d'experts, découverte d'information, etc.). Une autre étude [Chen et al., 2009] montre que les mesures de similarités basées sur les liens sociaux des utilisateurs ont de meilleures performances dans des scénarios où l'utilisateur cherche des contacts connus, alors que la similarité du contenu généré par les utilisateurs conduit à de meilleures performances dans la recommandation de personnes inconnues.

La similarité n'est pas la seule relation entre deux utilisateurs employée dans des systèmes Web. Facebook, par exemple, utilise la notion d'affinité[122] d'un utilisateur envers l'autre en fonction de leurs différentes interactions. Cette mesure reflète l'intérêt qu'un utilisateur porte aux informations qui concernent un autre utilisateur, ainsi que la fréquence de leurs interactions.

Certaines approches utilisent une combinaison (linéaire) de plusieurs fonctions de similarité pour établir une mesure complexe. Par exemple, une approche d'apprentissage automatique est utilisée [Lakiotaki et al., 2011] pour apprendre des facteurs de pondération de différentes mesures de similarité dans un système de recommandation de personnes similaires basé sur un site de filtrage collaboratif. Similairement, dans [Kazienko et al., 2010] une multitude de critères basés sur des interactions des utilisateurs avec des objets Web est utilisée pour concevoir une mesure de similarité complexe. Dans la recommandation de divers éléments Web (contenus,

[122] http://techcrunch.com/2010/04/22/facebook-edgerank/

utilisateurs, et groupes) dans le réseau social Flickr [Siersdoref & Sizov, 2009], une approche d'apprentissage automatique combinant des fonctions de similarité basées sur les liens sociaux et sur le contenu créé par des utilisateurs s'est montrée utile.

7.3. Vers une approche de recommandation des collaborateurs dans le cadre de l'innovation à travers le Web

Afin de répondre aux besoins identifiés dans la partie §7.1.1 en s'inspirant d'études sociales qui ont révélé certains facteurs influençant le choix de collaborateurs dans de vraies équipes, nous avons conçu une approche pour la recommandation de collaborateurs potentiels pour un utilisateur (utilisateur initial par la suite) dans le processus de résolution de problèmes en ligne qui essaye de maximiser :

- **La complémentarité** de concepts associés à des collaborateurs recommandés (profil conceptuel) avec ceux associés à l'utilisateur initial par rapport aux concepts qui sont associés à un problème. Ce paramètre sert à assurer que des collaborateurs recommandés ajouteraient des compétences et intérêts pertinents pour le problème et pas assez présents chez l'utilisateur initial ;
- **La similarité des profils conceptuels** des collaborateurs recommandés avec le profil de l'utilisateur. Indépendamment du problème en question, ce paramètre permet de s'assurer que des collaborateurs recommandés auraient des centres d'intérêt partagés avec l'utilisateur initial ;
- **L'appartenance** des collaborateurs recommandés aux mêmes cercles sociaux que l'utilisateur initial. Ce paramètre garantit une certaine proximité sociale. Connaître les mêmes personnes est souvent un facteur facilitant la communication et la création de nouveaux liens sociaux. Il est donc raisonnable de s'attendre à des effets positifs de ce critère dans la recommandation de collaborateurs.

Si le premier critère est motivé par la capacité d'une future équipe d'auteurs de solutions à affronter le problème avec une synergie de leurs capacités intellectuelles et de leurs connaissances, les deux derniers critères sont motivés par le besoin de former des équipes cohérentes et compatibles, qui partageraient des points communs (des intérêts communs, des amis communs, etc.) leur permettant de développer une affinité personnelle plus facilement et de s'accepter les uns les autres.

Nous recherchons notamment, dans l'esprit de l'innovation ouverte sur le Web, à recommander des personnes susceptibles d'apporter de nouvelles perspectives à l'équipe. La conception de notre approche s'inspire donc de nos précédents travaux sur la découverte de concepts latéralement pertinents et inattendus. L'évaluation que nous avons conduite se concentre également sur le scénario de résolution de problèmes en ligne.

Notre travail part de l'hypothèse selon laquelle il existe un réseau social, une partie du graphe Web, qui met en relation, d'une manière ou d'une autre, les utilisateurs entre eux ainsi que les utilisateurs et les concepts. Autrement dit, nous supposons qu'il est possible de générer des profils sociaux (PS) et des profils conceptuels (PC) tels qu'ils

sont décrits par les formules 2 et 4 dans la partie §3.3.1.1 (page 56), à partir d'une source de données disponible. La plupart des sites sociaux permettant la génération du contenu par les utilisateurs représenteront donc une source de données éligible pour l'application de notre approche. L'évaluation et l'implémentation d'un prototype sont effectuées à partir de réseau social Twitter, en raison de la disponibilité publique de ses données.

7.3.1. L'approche générale

Notre approche générale consiste en plusieurs phases, présentées graphiquement sur la Figure 61, qui peuvent être effectuées de différentes manières. Ces différentes alternatives feront l'objet d'une comparaison dans notre évaluation présentée dans la partie §7.4.

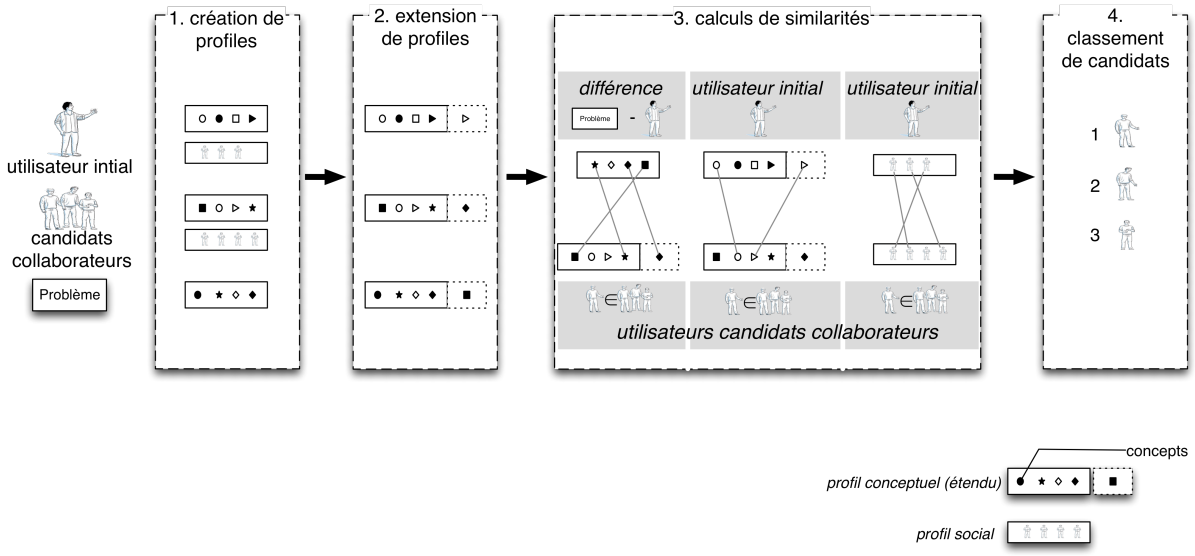


Figure 61 Quatre phases de la découverte de collaborateurs possibles

La première phase consiste à construire des profils conceptuels et sociaux des utilisateurs, l'utilisateur initial aussi bien que les utilisateurs candidats collaborateurs, ainsi que le profil conceptuel du problème.

Dans une deuxième phase, optionnelle, une extension des profils conceptuels est effectuée pour y ajouter d'autres concepts pertinents. Si un enrichissement de ce type est souvent pratiqué afin de compléter les profils éventuellement partiels, dans notre scénario l'extension est également motivée par :

- le fait qu'un concept (souvent représentant une unité de sens très spécifique et restreinte) trouvé en relation avec l'utilisateur révèle souvent des connaissances et des intérêts plus vastes qui restent non représentés dans le profil utilisateur. Par exemple un utilisateur intéressé à la transformation de Fourier est susceptible d'être intéressé par d'autres concepts, notions mathématiques nécessaires pour la compréhension de celle-ci.
- Le besoin de latéralité dans la recommandation i.e. la capacité de prendre en compte une pertinence latérale de deux concepts dans le calcul de la similarité de profils afin de s'assurer de la prise en compte des différentes perspectives. Il semble donc opportun d'enrichir les profils par des concepts pertinents afin de mieux trouver des profils similaires.

La troisième phase consiste à calculer des similarités entre différents profils, et plus précisément entre :

- des profils conceptuels de candidats collaborateurs et la différence de profils conceptuels de problème et de l'utilisateur initial. Cette différence de profils permet d'identifier des concepts non représentés dans le profil utilisateur mais néanmoins pertinents pour le problème. La similarité du profil d'un candidat collaborateur avec cette différence indiquerait une certaine complémentarité de celui-ci avec l'utilisateur initial, par rapport au problème en question. La différence conceptuelle peut être calculée en partant du profil initial ou bien du profil étendu dans la phase précédente ;
- des profils conceptuels de l'utilisateur initial et des candidats utilisateurs. Cette similarité peur être calculée à partir de profils initiaux ou étendus dans la phase précédente ;
- des profils sociaux de l'utilisateur initial et des candidats utilisateurs.

Ces similarités correspondent aux propriétés des candidats collaborateurs que nous avons identifiées en tant que désirables dans l'introduction de la partie §7.1.1.

Une fois les similarités élémentaires calculées, dans la quatrième phase il est nécessaire de classer les candidats collaborateurs en fonction des similarités calculées dans la phase précédente.

Nous présentons ci-après des méthodes alternatives utilisables dans chaque phase.

7.3.2. Création des profils

D'une manière générale, la création de profils conceptuels est discutée dans le chapitre 4 lors de la présentation des différentes hypothèses d'expertise et des travaux existants qui les appliquent pour former des profils d'experts et approvisionner la fonctionnalité de la recherche d'experts. Des approches appliquant les hypothèses d'expertise H19[123] et H20[124] sont notamment pertinentes pour nos travaux présentés dans ce chapitre, car ils se focalisent sur la construction de profils à partir de tweets, une approche entreprise dans nos réalisations. Le chapitre 6 traite également la question de la création de profils conceptuels à partir de tweets ainsi qu'à partir de correspondances entre tweets et articles de recherche.

Dans la phase de création de profils, il est nécessaire de créer des profils conceptuels de problèmes et des utilisateurs (l'utilisateur initial aussi bien que les utilisateurs candidats collaborateurs) et des profils sociaux des utilisateurs, en s'appuyant sur les données disponibles au sein d'un réseau social, Twitter dans notre cas. Nous rappelons qu'une définition formelle des profils est donnée par les formules 2 et 4 dans la partie §3.3.1.1. Les profils conceptuels que nous utilisons dans ce chapitre consistent à la fois en des concepts et des valeurs d'importance de chaque concept par rapport à l'objet de profil (un utilisateur ou un problème).

[123] H19 : Si l'utilisateur à créé un certain nombre de tweets sur le sujet X, alors il peut être considéré comme expert sur un sujet X.

[124] H20 : Si un sujet X apparaît de manière récurrente dans les tweets créés par l'utilisateur, alors celui-ci peut être considéré comme expert sur le sujet X.

Dans notre implémentation concrète, nous construisons des profils conceptuels à partir des éléments textuels en effectuant une extraction de concepts DBPedia à partir du texte. Nous rappelons que l'extraction de concepts n'est pas un sujet au cœur de nos travaux et que nous avons donc fait appel aux services externes pour effectuer celle-ci. Dans le cas de profils de problèmes, nous extrayons des concepts depuis le texte des problèmes en utilisant Zemanta. Nous considérons tous les concepts retrouvés par ce service comme étant de la même importance. En ce qui concerne les profils conceptuels des utilisateurs, nous effectuons une extraction de concepts à partir des 300 derniers tweets, et nous considérons le nombre d'occurrences de concepts dans ces tweets pour déterminer leur importance pour l'utilisateur. Le choix de cette technique particulière est fait en tenant compte de deux nécessités relatives à notre cas d'utilisation :

- la nécessité de favoriser la prise en compte de concepts qui représentent des intérêts récents ;
- la nécessité d'engranger une connaissance la plus complète possible de l'ampleur des intérêts d'une personne. Une telle focalisation va dans le sens de l'innovation ouverte dont l'inclusion des personnes marginalement compétentes pour un sujet est un des principes fondamentaux.

Il est néanmoins possible d'employer d'autres stratégies de construction de profils qui soient plus restrictives et focaliseront les profils sur les concepts maîtrisés par l'utilisateur. Dans des cas d'utilisation exigeant une restriction plus élevée, il serait notamment intéressant d'appliquer des stratégies de prise en compte de la dynamique de concepts et leur persistance dans les tweets d'un utilisateur à travers le temps ([Ziamatin, 2011], [Abel et al., 2011] et [Stan et al., 2011]), qui ont déjà montré des effets positifs sur la précision de profils de compétence.

Les profils sociaux des utilisateurs ont été construits à partir de tous leurs contacts sur Twitter.

7.3.3. Extension des profils

D'une manière générale, l'extension de profils consiste à faire appel à une mesure de proximité de concepts pour obtenir, à partir de l'ensemble de concepts présents dans le profil initial, un ensemble de concepts suggérés. Ces concepts sont ajoutés au profil initial pour créer le profil étendu. Les valeurs de proximité données par la fonction de proximité sont prises comme valeurs d'importance de ces nouveaux concepts dans les profils étendus. Nous avons expérimenté trois méthodes d'extension de profils : hyProximity, une mesure distributionnelle et le pseudo retour de pertinence.

Si notre mesure hyProximity, présentée dans le chapitre 5 a montré de bons résultats dans la recommandation de concepts, il semble pertinent de l'expérimenter plus avant et de déterminer son utilisabilité dans l'extension de profils pour la recommandation de collaborateurs. Dans ces expériences nous utilisons la mesure de proximité hyProximity - mixte.

Dans l'état le l'art décrivant des approches existantes pour le calcul de similarités de concepts, nous avons mentionné quelques approches distributionnelles, y compris Normalized Google Distance proposé dans [Cilibrasi & Vitanyi, 2007]. Cette mesure simple a inspiré beaucoup d'autres approches et reste souvent utilisée par des systèmes réels. Nous l'adaptons ici pour concevoir une fonction de proximité de concepts à partir de leur distribution dans les profils des utilisateurs.

$$Dist_{\tau}(c_1,c_2) = \frac{cooccurrence(c_1,c_2)}{occurrence(c_1) + occurrence(c_2)} \quad (18)$$

Notre mesure distributionnelle de deux concepts c_1 et c_2, dont la représentation formelle est donnée par la formule 18, représente un rapport de nombre de profils où ces deux concepts co-occurrent et la somme de leurs occurrences individuelles dans des profils des utilisateurs. Elle est calculée à partir d'un ensemble de profils utilisateurs τ. Cet ensemble représente, dans une version standard de l'approche, la totalité des profils connus par le système. Nous partons ici de l'idée que des concepts qui co-occurrent souvent dans les profils utilisateurs sont probablement assez proche (similaires ou complémentaires/dépendants).

Notre troisième fonction est inspirée par le pseudo retour de pertinence (**PRP**). Dans certaines approches de recherche Web [Macdonald & Ounis, 2007] ou de recherche d'experts [Serdyukov et al., 2007], il est possible de trouver une méthode d'enrichissement de requêtes de recherche par des termes qui occurrent souvent dans des résultats obtenus par la requête initiale. L'intuition derrière cette approche est que les concepts ou termes découverts par cette notion de proximité représenteront potentiellement des précisions ou enrichissements utiles par rapport à la requête initialle.

7.3.4. Mesures de similarité et calcul

Dans cette troisième phase, nous utilisons des mesures de similarité vectorielles pour calculer les similarités de différents profils. Au préalable, une normalisation de profils s'impose pour créer des vecteurs qui peuvent être comparés par les mesures vectorielles standards.

Par exemple, pour calculer la similarité de deux profils conceptuels P_1 et P2, nous constituerons d'abord un ensemble de tous les concepts pertinents C, qui représentent l'union des concepts présents dans ces deux profils. Ensuite nous construisons un vecteur pour chaque profil de manière suivante : pour chaque concept $c_j \in C$, si $c_j \notin P$, nous ajoutons la valeur 0 dans le vecteur ; et si $c_j \in P$ la valeur de l'importance que ce concept possède dans ce profil, i.e. nous ajoutons la valeur i, telle que $(c_j,i_j) \in P$. Pour des profils sociaux qui ne contiennent pas de valeur d'importance, nous prenons la valeur 1 comme étant l'importance de toutes les connexions sociales.

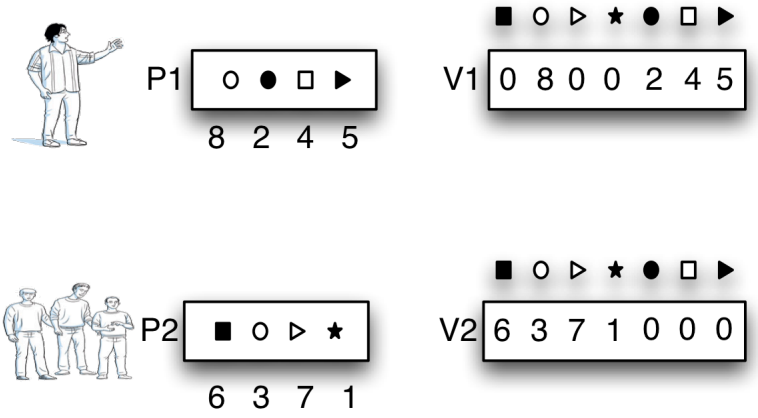


Figure 62 Vectorisation de deux profils conceptuels

La Figure 62 montre un exemple de deux profils transformés en forme vectorielle. Les symboles graphiques (similairement à la Figure 61) représentent des concepts différents et les chiffres inscrits au-dessous, leur valeur d'importance.

D'après ces vecteurs normalisés nous calculons deux mesures : la **mesure intersection** et la **similarité cosinus**.

La mesure intersection, calculé entre deux profils, P1 et P2, représente une somme de valeurs d'importance, pour tous les concepts qui apparaissent dans les deux profils[125]. En utilisant notre représentation vectorielle des profils (V1 et le vecteur dérivé de P1 et V2 dérivé de P2) cette mesure peut être représentée par la formule 19. Il est important de noter que cette mesure n'est pas symétrique. Sur l'exemple présenté à la Figure 62, INTERSECTION(P1,P2) serait 8, alors que INTERSECTION(P1,P2) serait 3. Cette simple mesure est souvent utilisée dans la littérature, comme par exemple dans le cas de l'approche pour le calcul de similarité de profils d'experts présentée dans [Balog & Rijke, 2007].

$$INTERSECTION(V_1,V_2) = \sum_{v1_i \in V_1, v2_i \in V_2, v2_i \neq 0} v1_i \qquad (19)$$

Une autre mesure, souvent utilisée dans la clustérisation et dans l'analyse de similarité de textes est la similarité cosinus [Salton 1983]. Elle est par exemple utilisée dans [Viswanathan & Finnin, 2010] dans le cadre du calcul de similarités textuelles. Cette mesure repose sur le calcul de cosinus de l'angle entre deux vecteurs représentant deux profils. C'est en fait en raison de cette mesure que nous normalisons nos profils sous forme vectorielle. Sa représentation formelle est donnée par la formule 20. Le calcul est effectué en divisant le produit scalaire des deux vecteurs par le produit des normes de ces mêmes vecteurs.

$$\cos(V_1,V_2) = \frac{V_1 \bullet V_2}{\| V_1 \| \cdot \| V_2 \|} \qquad (20)$$

Chacune de ces deux mesures peut être utilisée pour calculer les similarités entre des profils différents et notamment :

- **Similarité de centres d'intérêt (SCI)** i.e. la similarité entre les profils conceptuels de l'utilisateur initial et un utilisateur candidat collaborateur.
- **Similarité sociale (SS)** i.e. la similarité entre les profils sociaux de l'utilisateur final et un utilisateur candidat collaborateur.
- **Similarité avec la différence conceptuelle (SDC)**. Pour cette similarité nous calculons la différence de profil de problème et le profil conceptuel de l'utilisateur, et ensuite nous calculons la similarité de celle-ci avec le profil conceptuel d'un utilisateur candidat collaborateur. La Figure 63 montre un exemple graphique du calcul d'une telle différence.

[125] avec une valeur d'importance différente à 0

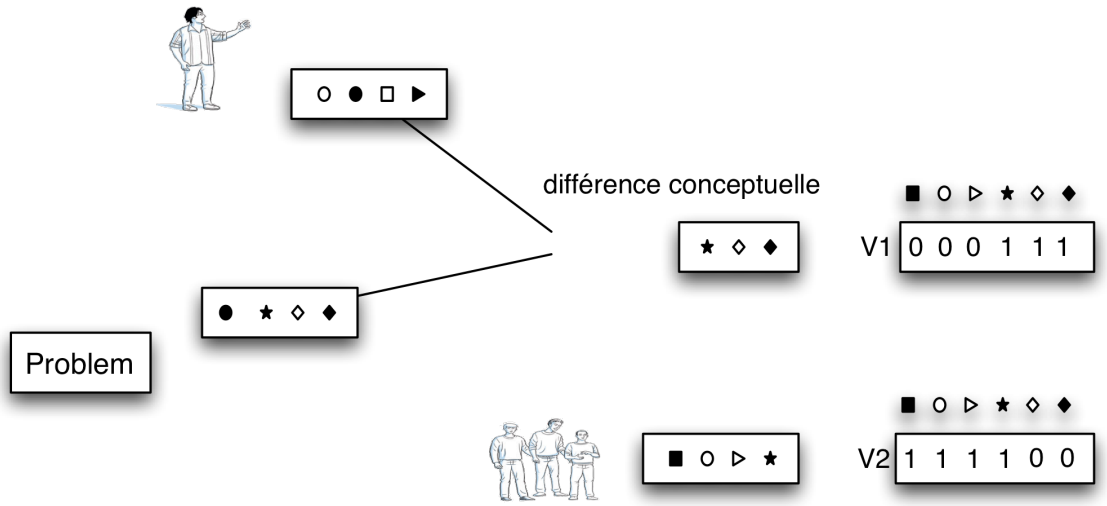


Figure 63 Calcul de la différence conceptuelle et vectorisation des profils. Sous l'hypothèse que toutes les valeurs d'importance soient = 1 pour des concepts présents dans le profil

7.3.5. Classement de candidats collaborateurs

Afin de classer les candidats collaborateurs et de proposer à l'utilisateur les meilleurs candidats dans cette 4ème phase, il est possible de se servir des similarités élémentaires (SCI, SS et SDC) ou bien de calculer une mesure composée afin de classer les candidats ayant un bon niveau de similarité par chaque mesure. A cette fin, il est possible de se servir d'une combinaison linéaire de ces mesures, aussi bien que d'un produit. Dans notre implémentation nous utilisons le produit Prod(SCI,SS,SDC) afin de pénaliser les candidats qui ont des bonnes valeurs pour un type de similarité mais des valeurs très mauvaises, voire nulles, pour les autres (x*0=0). Dans les cas où la combinaison linéaire est utilisée, une approche d'apprentissage automatique pourrait être mise en place pour adapter l'approche aux préférences de l'utilisateur en ajoutant des pondérations des fonctions élémentaires dans leur combinaison. Une telle approche est assez commune dans des systèmes de création de profils multicritères, comme celui présenté dans [Lakiotaki et al., 2011].

Quelle que soit la mesure de similarité choisie, dans la phase de classement, les candidats sont classés par ordre décroissant de similarité et un certain nombre de candidats les mieux classés est présenté à l'utilisateur.

7.4. Evaluation

Afin de comparer les différentes alternatives de notre approche, nous avons effectué deux évaluations. Nous avons sollicité des avis d'utilisateurs sur un nombre de potentiels collaborateurs. Ces avis nous ont servi d' « étalon-or » pour comparer des approches alternatives. Les deux évaluations se focalisent sur deux scénarios différents. Dans la première (Evaluation 1 par la suite) un groupe de collaborateurs cherche un membre complémentaire à leur équipe pour résoudre un problème d'innovation. Dans la deuxième (Evaluation 2 par la suite), c'est un auteur de solutions individuel qui cherche un collaborateur pour travailler avec lui sur un problème d'innovation. Ces deux scénarios, par leur complémentarité, assurent une capacité de conclusion évaluative plus élevée. Si dans le cadre de l'évaluation basée sur les opinions d'un groupe de

solveurs, la validité des conclusions est assurée par le taux d'accord entre les membres de l'équipe, dans le cas des auteurs de solutions individuelles, c'est un nombre de participants, légèrement plus grand, qui assure notre capacité de tirer des conclusions.

Les deux évaluations sont effectuées sur 3 problèmes multidisciplinaires, dont les descriptions sont disponibles en ligne[126]. L'un des problèmes repose sur des sujets liés au Web Sémantique et au Marketing en ligne ; l'autre sur le Web Sémantique et les Capteurs et le troisième sur le Web Sémantique et l'analyse des entreprises (start-up). Ces sujets de problèmes nous ont permis de choisir des évaluateurs parmi des experts reconnus en Web Sémantique, qui pour résoudre ces problèmes auraient besoin d'un collaborateur complémentaire. Les problèmes d'innovation, quant à eux, étaient formulés en s'inspirant des sujets des différents projets.

Le premier pas dans l'évaluation fut la création des profils des utilisateurs qui ont accepté d'y participer en tant qu'évaluateurs. Par des méthodes déjà décrites dans la partie §7.3.2 et en utilisant les comptes Twitter des utilisateurs, nous avons construit des profils conceptuels et sociaux pour chacun d'eux. Dans le cas de l'évaluation 1, nous avons construit les profils du groupe en unissant les profils individuels des membres du groupe à la recherche d'un collaborateur. Nous avons également créé un profil pour chacun des problèmes d'innovation.

Dans un second temps, nous avons obtenu pour chaque paire <utilisateur/groupe, problème> un nombre de candidats collaborateurs. Ces candidats ont été obtenus en récoltant de Twitter :

- Tous les utilisateurs trouvés dans une proximité sociale de l'utilisateur/groupe en question. Nous avons notamment cherché le deuxième cercle d'amis (des amis des amis) de l'utilisateur, en excluant ses contacts directs.
- Tous les utilisateurs trouvables par l'outil de recherche de personnes sur Twitter en effectuant des requêtes correspondant aux concepts faisant partie du profil conceptuel du problème et de l'utilisateur initial. Cet outil effectue la recherche de personnes à partir de l'occurrence des mots-clés de requête dans des tweets et des résumés des utilisateurs.

Ces deux moyens de récolte de candidats collaborateurs correspondent aux intentions de la recherche de collaborateurs décrits dans la partie §7.1.1 et devraient assurer une bonne couverture des bons candidats.

Après avoir récolté des candidats collaborateurs nous appliquons notre méthode de découverte de collaborateurs potentiels comprenant l'extension des profils puis l'application de différentes mesures de proximité, comme décrit dans la partie §7.3.1. Nous générons des recommandations de collaborateurs pour chaque alternative de l'approche. À l'aide d'un retour des évaluateurs, nous évaluons la performance des différentes alternatives de l'approche. Ces comparaisons seront détaillées par la suite.

Pour comparer des alternatives de l'approche nous utilisons deux mesures : DCG (en. *Discounted Cumulative Gain*) et AMP (Apport Moyenne en Pertinence).

La première mesure, DCG, est définie par la formule 21, et représente la valeur cumulative des appréciations des évaluateurs des candidats proposés par le système dans les premières *n* (10 dans notre cas) propositions. Ces appréciations sont

[126] http://research.hypios.com/?page_id=184

représentées dans la formule par la fonction $rating_i$ qui prend la valeur 1 si l'utilisateur proposé par le système à la position de classement *i* est pertinent (figure dans les appréciations des évaluateurs avec une note positive), et 0 dans le cas contraire. Cette valeur est ensuite décomptée par $\log_2 i$ pour favoriser des classements où des bons candidats sont placés en premier. Prenons l'exemple de 10 résultats de suggestion dont 4 sont bons. Si dans un classement A nous trouvons tous ces 4 bons candidats classés en premier, et dans un classement B ces 4 candidats se trouvent dans des positions 1,3,5, et 9, alors le classement A aurait une meilleure valeur de DCG que le classement B.

$$DCG = rating_1 + \sum_{i=2}^{10} \frac{rating_i}{\log_2 i} \quad (21)$$

L'autre mesure, AMP, représente la surface sous la courbe de précision par rappel. Elle permet de mesurer la valeur moyenne de la précision en fonction de la croissance de rappel. Le calcul de cette fonction consiste à calculer, à chaque position de classement de suggestions *i* la valeur de précision sur les premières *i* suggestions et de multiplier cette valeur par l'incrément de rappel entre la position *i* et *i -1*. Le cumul de ces valeurs jusqu'à un niveau *n* donne la valeur finale de la mesure. Un choix qui aurait de meilleures valeurs de cette fonction avec un nombre plus petit de suggestions prises en compte, serait meilleure par rapport à la valeur de ces suggestions qui sont les mieux classées.

$$AMP_n = \sum_{i=1}^{n} precision(i) \bullet \Delta recall(i) \quad (22)$$

7.4.1. Evaluation 1

7.4.1.1. Contexte de l'étude

Nous avons sollicité la participation de 3 utilisateurs, experts en Web Sémantique, qui travaillent souvent ensemble, en tant qu'évaluateurs dans cette expérience. Les trois problèmes ont été montrés aux utilisateurs afin qu'ils se familiarisent avec eux.

Nous avons ensuite généré des candidats collaborateurs et fait marcher nos mesures de similarité afin de recommander des collaborateurs. Ensuite, nous avons pris les 10 collaborateurs les mieux classés de chaque alternative en les combinant dans une liste à ordre aléatoire. Dans une interface Web, nous avons demandé à nos évaluateurs de juger, pour chaque collaborateur proposé, s'il travaillera avec lui/elle sur la résolution du problème en question et tenant en compte sa complémentarité à leur équipe existante. L'interface Web montrait pour chaque candidat collaborateur son nom d'utilisateur chez Twitter et une biographie courte. Si ces informations ne suffisaient pas à l'évaluateur pour prendre sa décision, il pouvait aussi visiter le profil de l'utilisateur sur Twitter et voir ses tweets pour se faire une idée plus précise des centres d'intérêt du collaborateur proposé.

Dans cette évaluation, nous nous focalisons sur la similarité avec la différence conceptuelle, et la similarité des centres d'intérêt parce que la similarité sociale avec un groupe est difficile à évaluer.

Le taux d'accord des évaluateurs sur la complémentarité des collaborateurs suggérés n'était pas suffisant pour toutes les paires d'évaluateurs dans la première phase. Nous avons donc appliqué notre méthodologie itérative, détaillée sur notre page Web

dédiée[127] afin d'arriver à un accord. Après la deuxième itération, le coefficient d'accord k [Fleiss, 1981] était supérieur à 0.6 (en moyenne 0.74) pour toutes les paires d'évaluateurs pour les 3 problèmes. Nous avons pris tous les candidats jugés favorablement par au moins 2 évaluateurs.

7.4.1.2. Résultats

Les résultats de cette étude permettent notamment d'évaluer l'impact des différentes techniques d'enrichissement des profils. Sur la Figure 64, nous présentons une comparaison de nos différentes méthodes d'enrichissement, par la métrique DCG. Dans ce cas, nous avons appliqué nos 3 méthodes d'enrichissement sur les profils de problèmes, et ensuite calculé la similarité (similarité cosinus aussi bien que la similarité Intersection) avec la différence conceptuelle, en gardant des profils utilisateurs initiaux (non étendus). La mesure hyProximity montre de meilleurs résultats. La différence entre cette mesure et la mesure Dist a été confirmée comme signifiante par le test T ($p < 0.05$).

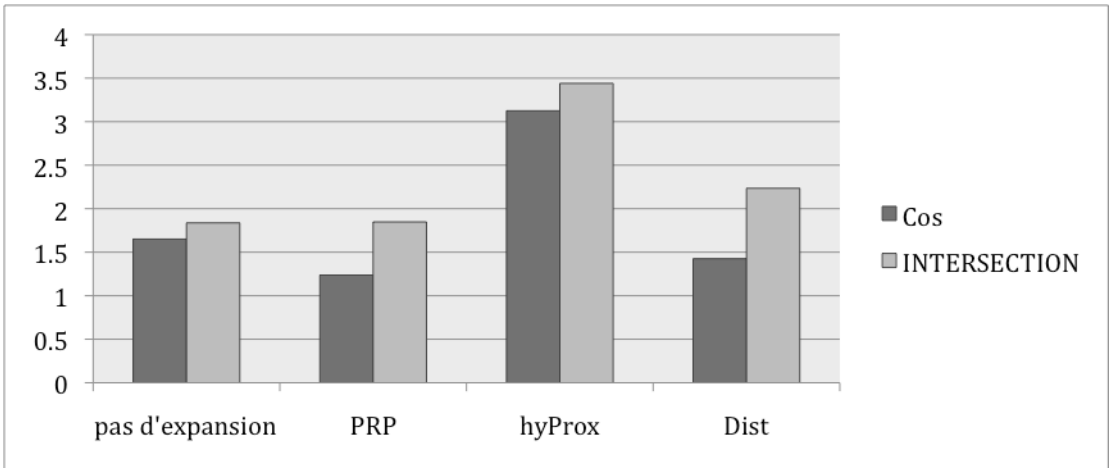


Figure 64 Résultats de la métrique DCG appliquée sur le calcul de la similarité SDC, avec l'extension de profils de problème par différentes méthodes

Dans la Figure 65, nous présentons une autre comparaison, cette fois-ci dans le cadre de l'extension des profils de l'utilsiateur initial (le groupe des utilisateurs). Danc ce cas, la mesure hyProximity montre des résultats inferieurs à la mesure distributionelle. Contrairement au cas précédent, nous n'avons pas pu confirmer la signification statistique de cette différence par le test T ($p > 0.05$).

Dans les deux cas, l'usage de pseudo retour de pertinence n'ajoute pas beaucoup de valeur par rapport au calcul de similarités de base.

D'une manière générale, les résultats obtenus ne sont guère surprenants. Il est raisonnable de s'attendre à ce qu'une mesure basée sur la distribution des sujets dans des profils utilisateurs se montre le meilleur choix dans l'expansion des profils pour le calcul des similarités entre des utilisateurs. D'un autre côté, il n'est pas surprenant qu'une mesure basée sur le sens des concepts soit la meilleure dans l'extension du champ conceptuel du problème.

[127] http://milstan.net/stankovic_rowe_methodology/

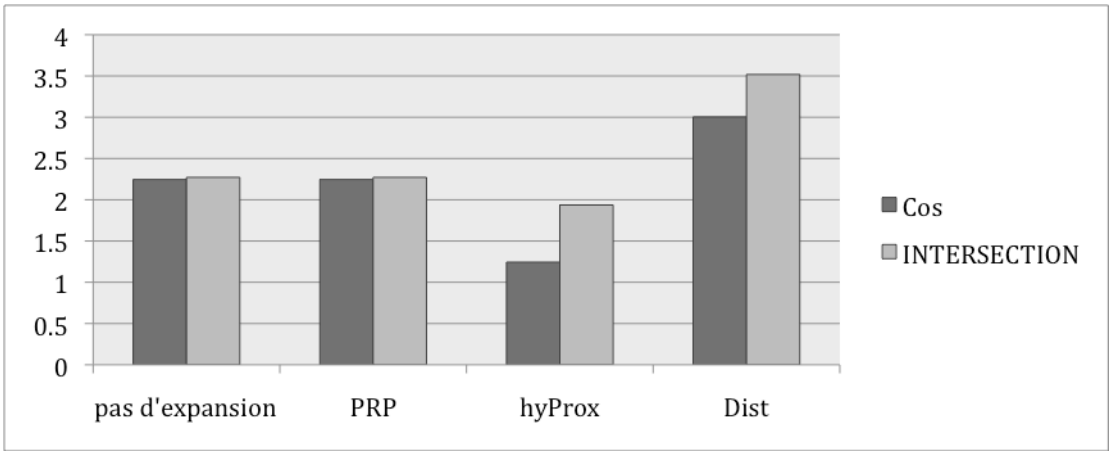


Figure 65 Résultats de la métrique DCG appliquée sur le calcul de la similarité SCI, avec l'extension de profils de problème par des méthodes différentes

7.4.2. Evaluation 2

7.4.2.1. Création de « l'étalon-or »

Dans cette étude nous avons sollicité la participation de 12 utilisateurs Twitter indépendants, experts en Web Sémantique. Pour chacun d'entre eux et pour chacun des 3 problèmes, nous avons généré des candidats collaborateurs (36 ensembles de candidats). La taille de ces ensembles de candidats (80-240) était inférieure à celle de l'évaluation précédente, en raison de la nature individuelle des profils, ce qui a rendu possible une évaluation de tous les candidats collaborateurs récoltés. Pour obtenir leur jugement sur les candidats, nous avons posé la même question que dans la précédente étude, à savoir s'ils travailleraient avec l'utilisateur en question sur le problème d'innovation donné. Nous avons également explicité que leur choix de bons collaborateurs devrait prendre en compte la capacité perçue de cet utilisateur à les aider à résoudre le problème ainsi qu'un jugement de compatibilité personnelle.

Cette évaluation à priori de la recommandation des candidats nous a permis de constituer un étalon-or qui serait utile dans la comparaison de plusieurs méthodes, et pas seulement celles pré-choisies avant l'évaluation. C'est en obtenant à posteriori des candidats suggérés par différentes méthodes et en vérifiant s'ils figurent dans l'étalon-or que nous pourrons calculer nos métriques de performance DCG et AMP.

7.4.2.2. Résultats

Dans cette étude nous avons notamment été intéressés par les comparaisons des différentes alternatives du produit des différentes mesures de similarité i.e. par les résultats d'une notion de similarité multicritères. Nous comparons les alternatives suivantes :

- SCI-cos*SS-cos*SDC-cos : le produit des mesures de similarité SCI, SS et SDC en utilisant la fonction de similarité cosinus.
- PRP(SCI-cos*SS-cos*SDC-cos) : extension de profil de problème par le pseudo retour de pertinence et l'application de la mesure précédente.
- SCI-cos*SS-cos*hyProx(SDC-cos) : extension du profil de problème par hyProximity avant le calcul de la mesure de similarité avec la différence

conceptuelle par la similarité cosinus ; puis le produit de cette mesure avec SCI et SS également obtenus en appliquant la similarité cosinus sur des profils initiaux.

- Dist(SCI-cos)*SS-cos*SDC-cos : extension du profil de l'utilisateur initial par la mesure distributionnelle Dist avant le calcul de la mesure de similarité des centres d'intérêt par la similarité cosinus ; puis le produit de cette mesure avec SS et SDC également obtenus en appliquant la similarité cosinus sur des profils initiaux.
- Dist(SCI-cos)*SS-cos*hyProx(SDC-cos) : extension du profil de problème par hyProximity avant le calcul de la mesure de similarité avec la différence conceptuelle par la similarité cosinus ; extension du profil de l'utilisateur initial par la mesure distributionnelle Dist avant le calcul de la mesure de similarité des centres d'intérêt par la similarité cosinus ; puis le produit de ces deux mesures avec SS également obtenu en appliquant la similarité cosinus sur des profils initiaux.

Ce choix particulier des alternatives est fait en s'inspirant par des résultats de l'évaluation 1. Nous avons choisi d'étendre les profils par hyProximity dans le contexte du calcul de la similarité SDC parce que cette méthode d'extension donnait des meilleurs résultats pour ce type de similarité dans l'évaluation 1. Pour des mêmes raisons l'extension à l'aide de la mesure distributionnelle a été choisie pour la similarité SCI.

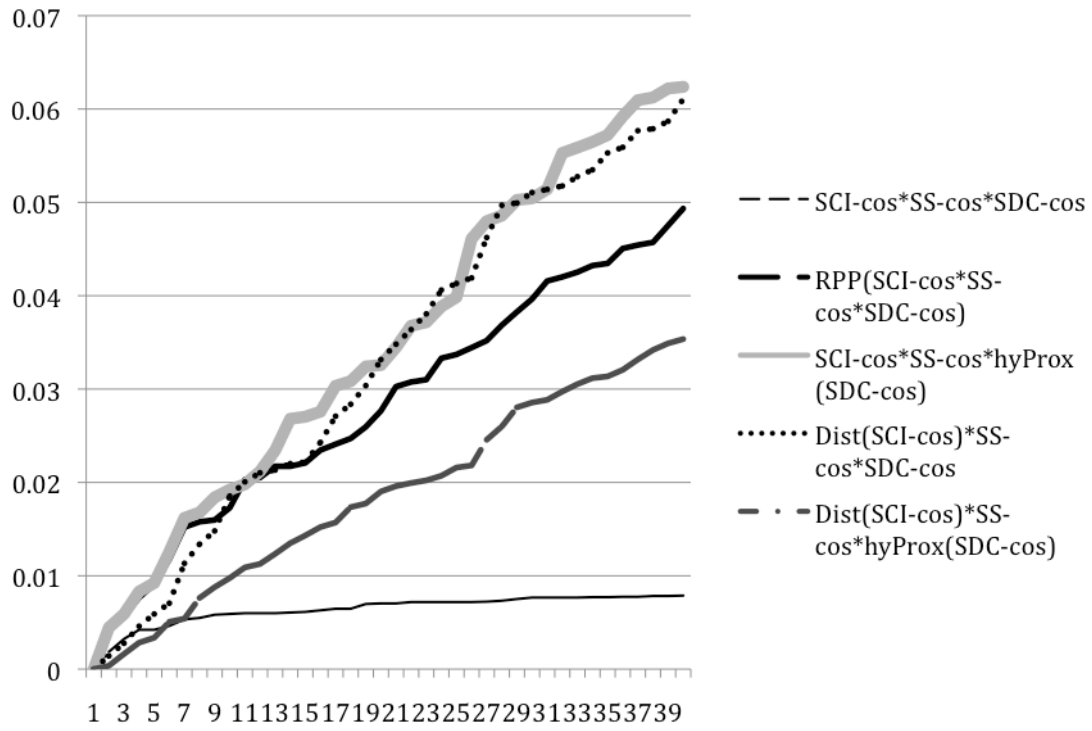


Figure 66 mesure AMP calculée pour différentes positions dans le classement (axe x)

La Figure 66, montre une comparaison de ces alternatives par rapport à la mesure AMP. Il est clair que toutes les alternatives d'enrichissement de profils permettent d'atteindre de meilleures valeurs d'AMP plus rapidement que l'approche de base. L'enrichissement des profils de problèmes par hyProximity ressort comme le meilleur choix, suivie dans la plupart du spectre de positions de classement par l'enrichissement des profils des utilisateurs initiaux par la mesure distributionnelle. L'enrichissement par le pseudo

retour de pertinence atteint des performances inférieures à celles-ci pour des positions de classement supérieures à 10. L'enrichissement mixte ne se montre pas très prometteur, avec des performances même inférieures à l'approche de base pour certaines positions de classement. Il est possible que cela soit dû à une divergence très importante par rapport aux sujets de départ, causée par le double enrichissement.

La comparaison des mêmes alternatives par la métrique DCG confirme les conclusions que nous avons tirées en analysant la mesure AMP. Toutes les mesures d'enrichissement montrent de meilleurs résultats que l'approche de base sur les 3 problèmes. Dans 2 problèmes sur 3 l'enrichissement des profils de problèmes par hyProximity apporte le plus de bénéfice à la performance de la suggestion. L'approche par deux enrichissements, encore une fois, a des effets d'une utilité redoutable.

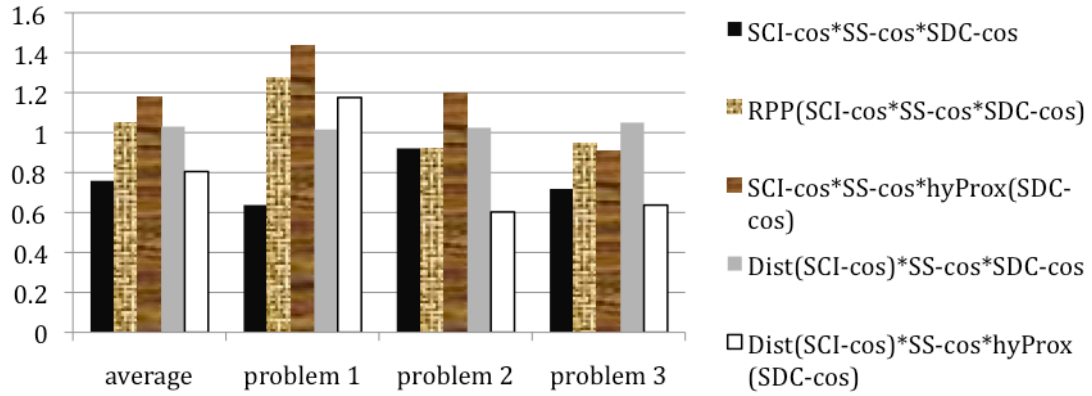


Figure 67 DCG calculé pour les 10 premières suggestions de chaque approche

Si nous nous sommes appuyés sur la mesure cosinus pour effectuer nos suggestions de collaborateurs, c'était pour des raisons d'uniformité. Nous avons également fait de semblables comparaisons avec la mesure INTERSECTION, mais les résultats montraient des différences négligeables par rapport à ceux obtenus par la mesure cosinus. La mesure F, similairement aux deux autres mesures, confirmait les mêmes performances et n'apportait pas de valeur analytique ajoutée. Leur présentation était donc d'un intérêt limité pour ce mémoire.

7.5. Un exemple de collaborateurs suggérés

Si nos évaluations nous ont permis de comparer la performance de différentes alternatives, nous avons jusqu'ici peu parlé de la vraie utilité des suggestions obtenues. C'est pourquoi nous donnons un exemple de collaborateurs suggérés, en espérant que cela aidera les lecteurs à mieux comprendre la nature des suggestions et leur pertinence.

Pour l'utilisateur Twitter avec l'identifiant « milstan », les collaborateurs suggérés pour le problème 3, en utilisant la mesure SCI-cos*SS-cos*hyProx(SDC-cos) étaient :

- davidsrose
- fundingpost
- ECVentureCapita
- BVCA

- vc20
- AndySack
- CVCACanada
- Austin_Startups
- tgmtgm
- davidblerner

Cette liste correspond en effet à l'intention de trouver des collaborateurs complémentaires, qui compenseraient le manque de compétence en matière de start-ups et de capital requises chez milstan et requises pour le problème en question.

L'utilisateur David Rose (davidrose) est un investisseur (en. angel investor) spécialisé en technologies de l'information, un intérêt qu'il partage avec milstan. Thomas McInerney (tgmtgm) est un interpréteur notamment expérimenté dans des start-ups, travaillant dans le domaine des sites sociaux. Il a participé à l'équipe de la start-up KLOUT qui fournit des métriques de popularité au sein des réseaux sociaux. David Blerner (davidblerner) et Andy Sack (AndySack) sont des entrepreneurs et investisseurs intéressés par les technologies de l'information. Les autres utilisateurs proposés représentent les comptes Twitter d'institutions[128], telles que des incubateurs de start-ups, qui ont certainement au sein de leurs équipes des experts en matière de l'analyse des start-ups, mais seront probablement plus difficiles à approcher que des particuliers.

7.6. Conclusions

Dans ce chapitre nous abordons le scénario de la résolution de problème en équipe. Nous proposons ensuite un rapprochement entre des collaborateurs potentiels. Ce rapprochement vise à répondre au besoin d'une complémentarité en termes de compétences, et d'une compatibilité personnelle (partage d'intérêts et appartenance aux mêmes cercles sociaux). Evidemment, ce rapprochement est plus complexe que ceux traités précédemment car il comprend à la fois une proximité entre collaborateurs, et une proximité entre les collaborateurs et le problème. Nous avons proposé une approche générale ainsi que plusieurs alternatives que nous avons ensuite comparées. Notre approche consiste d'abord à élargir les profils conceptuels des utilisateurs afin de mieux orienter les rapprochements et de compenser le manque de d'information de profils, inhérente à tout profil construit à partir de traces trouvées sur le Web. Ce n'est qu'une fois que cette étape est achevée que nous pouvons faire le calcul des proximités de profils utilisateurs. Nous avons essayé plusieurs méthodes de rapprochements de concepts dans le but d'élargir les profils. En plus de hyProximity, nous avons expérimenté avec une méthode basée sur la cooccurrence des concepts dans des profils, et une méthode de pseudo retour de pertinence.

Nos études ont permis de tirer des conclusions suivantes :

- L'extension de profils a des effets bénéfiques sur la performance de la recommandation de collaborateurs par rapport à l'usage des profils non élargis.

[128] La pratique de communiquer via Twitter, Facebook et d'autres média sociaux est devenu de plus en plus courante pour des entreprises, surtout dans l'industrie informatique.

- L'usage de hyProximity pour élargir des profils montre de meilleurs résultats lorsqu'il s'agit de calculer la complémentarité des candidats collaborateurs. Par contre, la mesure basée sur la cooccurrence se montre meilleure pour le calcul de compatibilité (similarité de profils conceptuels).
- Cependant, la combinaison de ces deux méthodes d'élargissement de profils ne semble pas donner des résultats optimaux.

Nous avons illustré l'utilité de notre approche sur un exemple de suggestion de collaborateurs basé sur des données issues de Twitter. Celui-ci montre les caractéristiques souhaitées pour ces suggestions, notamment pour des problèmes pluridisciplinaires où une complémentarité doit être couplée avec la compatibilité des collaborateurs.

Le travail présent dans ce chapitre, et notamment l'idée d'extension de profils, représente une base pour les travaux d'avenir. Certains ont déjà été entamés, comme la prédiction de liens sociaux sur Twitter où le même principe d'expansion pourrait s'appliquer. Similairement, il serait possible d'imaginer d'autres cas d'application de notre approche.

Bibliographie

[Abel et al., 2011] Abel, F., Gao, Q., Houben, G. J., & Tao, K. (2011). Analyzing Temporal Dynamics in Twitter Profils for Personalized Recommendations in the Social Web. *Web Science Conference*. Koblenz.

[Barreto et al, 2008] Barreto, a, Barros, M., & Werner, C. (2008). Staffing a software project: A constraint satisfaction and optimization-based approach. *Computers & Operations Research*, 35(10), 3073-3089. doi:10.1016/j.cor.2007.01.010

[Brocco & Groh, 2009] Brocco, M., & Groh, G. (2009). Team recommendation in open innovation networks. *Proceedings of the third ACM conference on Recommender systems - RecSys '09*, 365. New York, New York, USA: ACM Press. doi:10.1145/1639714.1639789

[Chen et al., 2009] Chen, J., Geyer, W., Dugan, C., Muller, M., & Guy, I. (2009). Make new friends, but keep the old: recommending people on social networking sites. *Proceedings of the 27th international conference on Human factors in computing systems*. ACM.

[Cilibrasi & Vitanyi, 2007] Cilibrasi, R. L., & Vitanyi, P. M. B. (2007). The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3), 370-383. doi: 10.1109/TKDE.2007.48.

[Cornfield,1951] Cornfield, J (1951). "A Method for Estimating Comparative Rates from Clinical Data. Applications to Cancer of the Lung, Breast, and Cervix". Journal of the National Cancer Institute 11: 1269–1275.*Proceedings of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining*. New York. USA

[Fleiss, 1981] Fleiss, J. L. (1981) *Statistical methods for rates and proportions*. 2nd ed. (New York: John Wiley)

[Guy et al., 2010] Guy, I., Jacovi, M., Perer, A., Ronen, I., & Uziel, E. (2010). Same places, same things, same people?: mining user similarity on social media. *Proceedings of the 2010 ACM conference on Computer supported cooperative work (pp. 41-50)*. ACM.

[Hinds et al., 2000] Hinds, P. J., Carley, K. M., Krackhardt, D., & Wholey, D. (2000). Choosing Work Group Members: Balancing Similarity, Competence, and Familiarity• 1. *Organizational Behavior and Human Decision Processes*, 81(2), 226–251. Elsevier.

[Huckman & Staats, 2010] Huckman, R. S., & Staats, B. R. (2010). Fluid Tasks and Fluid Teams : The Impact of Diversity in Experience and Team Familiarity on Team Performance Fluid Tasks and Fluid Teams : The Impact of Diversity in Experience and Team Familiarity on Team Performance. *Knowledge Creation Diffusion Utilization*.

[Johne & Snelson, 1988] Johne, A.F. & Snelson, P.A. (1988) Success factors in product innovation: A selective review of the literature. *Product Innovation Management*, 5, 114–128

[Jones, 2008] Jones, B. F., Wuchty, S., & Uzzi, B. (2008). Multi-university research teams: shifting impact, geography, and stratification in science. Science, 322(5905), 1259-62.

[Kazienko et al., 2010] Kazienko, P., Musial, K., & Kajdanowicz, T. (2011). Multidimensional Social Network in the Social Recommender System. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans,* 41(4), 746-759.

[Lakiotaki et al., 2011] Lakiotaki, K., Matsatsinis, N., & Tsoukias, A. (2011). Multicriteria User Modeling in Recommender Systems. *IEEE Intelligent Systems*, 26(2), 64-76.

[Luo et al., 2008] Luo, H., Niu, C., Shen, R., & Ullrich, C. (2008). A collaborative filtering framework based on both local user similarity and global user similarity. *Machine Learning*, 72(3), 231-245.

[Macdonald & Ounis, 2007] Macdonald, C., & Ounis, I. (2007). Expertise drift and query expansion in expert search. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management - CIKM '07*, 341. New York, New York, USA: ACM Press.

[Roure & Keely, 1990] Roure, J.B. & Keeley, R.H. (1990) Predictors of success in new technology based ventures. *Journal of Business Venturing,* 5, 221–239

[Salton, 1983] Salton, G.,McGill, M.J. (1983*) Introduction to modern information retrieval*

[Serdyukov et al., 2007] Serdyukov, P., Chernov, S., & Nejdl, W. (2007). Enhancing expert search through query modeling. Advances in Information Retrieval, 29th European Conference on IR Research, ECIR [Balog & Rijke, 2007] Balog, K., & de Rijke, M. (2007). Finding similar experts. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07,* 821. New York, New York, USA: ACM Press. doi:10.1145/1277741.1277926

[Siersdoref & Sizov, 2009] Siersdorfer, S., & Sizov, S. (2009). Social recommender systems for web 2.0 folksonomies. *Proceedings of the 20th ACM conference on Hypertext and hypermedia – HT'09*, 261. New

[Spertus et al., 2005] Spertus, E., Sahami, M., & Buyukkokten, O. (2005). Evaluating similarity measures.

[Stan et al., 2011] Stan, J.; Viet-Hung Do; Maret, P., "Semantic User Interaction Profils for Better People Recommendation," *Advances in Social Networks Analysis and Mining* (ASONAM), 2011, M Rowe, S

[Viswanathan & Finnin, 2010] Viswanathan, K. K., & Finin, T. (2010). Text Based Similarity Metrics and Deltas for Semantic Web Graphs. In P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, et al. (Eds.), *Proceedings of the 9th International Semantic Web Conferene* - ISWC 2010. Shanghai, China

[Ziamatin, 2011] Ziaimatin, H. (2011). DC Proposal : Capturing Knowledge Evolution and Expertise in Community-driven Knowledge Curation Platforms. *International Semantic Web Conference*. Bonn, Germany.

8. Ontologies pour faciliter les rapprochements

Ce chapitre aborde la question de la représentation d'éléments sous forme structurée dans le contexte des rapprochements dans les scénarios liés à l'innovation. Le besoin d'une représentation riche émerge à la fois d'une volonté d'assurer la richesse informationnelle des éléments du graphe Web, et de faciliter de futurs rapprochements. La partie §8.2 est consacrée à la modélisation des quatre types d'éléments Web qui participent aux rapprochements. Cette modélisation consiste à utiliser les ontologies existantes, les combiner et créer de nouvelles ontologies pour compléter la gamme existante. La partie §8.3 traite notamment de la modélisation des systèmes de rapprochement et de leur fonctionnement. Dans cette partie, nous présentons les ontologies qui servent à représenter des hypothèses de pertinence sur lesquelles reposent les rapprochements. Ces hypothèses peuvent être des hypothèses d'expertise, utilisées pour construire des profils utilisateur, ou bien les hypothèses de proximité sémantique, utilisées pour rapprocher des concepts et des utilisateurs. Dans la partie §8.4, nous présentons les ontologies permettant de représenter les résultats de chaque rapprochement. Enfin, nous présentons dans la partie §8.5 un scénario de réutilisation de rapprochements pour de nouveaux rapprochements.

8.1. Cadre général des ontologies

8.1.1. Introduction

Dans les chapitres précédents, nous avons montré comment il était possible de mettre à profit la richesse de la structure du graphe de données liées dans différent cas d'utilisation. Dans ce chapitre nous abordons les questions liées à l'émergence de cette même richesse, à la fois structurelle et informationnelle. Nous nous intéressons notamment à la description des ressources informationnelles pertinentes pour les scénarios de l'innovation sur le Web. Ce sont bien évidemment des ressources décrivant des utilisateurs, potentiels auteurs de solutions, des traces utilisateurs, des concepts, des problèmes d'innovation. Si ces éléments sont utiles pour alimenter nos différents outils de rapprochement, il ne faut pas négliger le besoin de décrire des résultats des rapprochements. Comme nous l'avons vu dans le Chapitre 7, certains rapprochements complexes peuvent s'appuyer sur des résultats d'autres rapprochements, dits élémentaires. Il est donc utile de publier les résultats de rapprochements, et cela de manière assez riche pour permettre aux applications consommatrices de ces données d'estimer l'applicabilité des rapprochements éventuellement déjà faits et potentiellement utiles dans une situation donnée. Dans cette estimation il est également utile de mettre à la disposition des consommateurs une description riche du rapprochement fait i.e. de l'hypothèse de proximité qui est utilisée pour rapprocher des éléments.

8.1.2. Les éléments modélisés

Tableau 15 Informations à modéliser dans le cas de rapprochements différents

<table>
<tr><th>Rapprochement</th><th>Eléments Web utilisés</th><th colspan="2">Fonctionnement du rapprochement (hypothèse de pertinence)</th><th>Résultats du rapprochement</th></tr>
<tr><td>utilisateur – concept</td><td rowspan="4">Utilisateurs, Traces (différents types), Concepts, Problèmes</td><td rowspan="4">Système de rapprochement</td><td>Hypothèse d'expertise</td><td>Profil d'utilisateurs</td></tr>
<tr><td>concept – concept</td><td rowspan="2">Mesure (hypothèse) de proximité sémantique</td><td>Proximités sémantiques de concepts</td></tr>
<tr><td>utilisateur – utilisateur</td><td rowspan="2">Proximité (compatibilité) des utilisateurs</td></tr>
<tr><td>trace – trace</td><td>Mesure (hypothèse) de proximité sociale</td></tr>
<tr><td></td><td></td><td></td><td>Hypothèse de proximité de traces</td><td>Correspondances des traces</td></tr>
</table>

Pour chaque rapprochement que nous avons conçu dans cette thèse, il peut être utile de modéliser trois types d'éléments : les éléments Web qui participent aux rapprochement, le fonctionnement du rapprochement et ses résultats. Dans le Tableau 15 nous présentons ces éléments dans le contexte de nos rapprochements.

Pour tous les scénarios il est d'abord nécessaire de bien décrire les éléments Web fondamentaux du scénario de l'innovation, i.e. des éléments suivants :

- **Les utilisateurs**. La description des utilisateurs doit permettre à la fois d'identifier les utilisateurs concernés par des traces, et cela parfois à travers différents sites Web où ils maintiennent leur présence, et éventuellement de les contacter. Les liens entre utilisateurs sont également très importants pour certains scénarios et nécessitent donc d'être modélisés.
- **Les concepts, sujets d'intérêt et de compétence.** Il est nécessaire de pouvoir représenter et identifier de manière non ambigüe des concepts de compétence et d'intérêt. Si la manière dont les concepts sont décrits dans le LOD actuel suffit largement à cet objectif, nous discutons par la suite certaines implications pratiques de l'usage de cette notion de concept dans le LOD .
- **Les traces utilisateurs**. Les traces utilisateurs sont à la source de beaucoup de rapprochements. La représentation de la notion des traces à un haut niveau d'abstraction, surtout pour distinguer de types de traces, est également importante pour l'échange de données sur les rapprochements parmi des systèmes hétérogènes.
- **Les problèmes d'innovation**. Si la description des problèmes d'innovation est importante pour les rapprochements, elle trouve sa place également dans l'échange des problèmes d'innovation entre des plates-formes. Nous avons donc abordé la modélisation des problèmes en considérant un contexte d'usage plus vaste.

Ensuite, il est nécessaire de modéliser le processus de rapprochement i.e. de décrire la nature de son fonctionnement. Celle-ci est donnée notamment par les hypothèses de pertinence utilisées pour effectuer le rapprochement et par le système qui l'a effectué. La Figure 68 présente une vue d'ensemble (à un haut niveau d'abstraction) de tous ces éléments modélisés et de leurs relations.

La richesse de cette description de rapprochements permettrait ensuite de décrire les résultats de rapprochements dans leur contexte, et de rendre possible leur réutilisation. Nous abordons la question de ces représentations plus en détail dans la suite de ce chapitre.

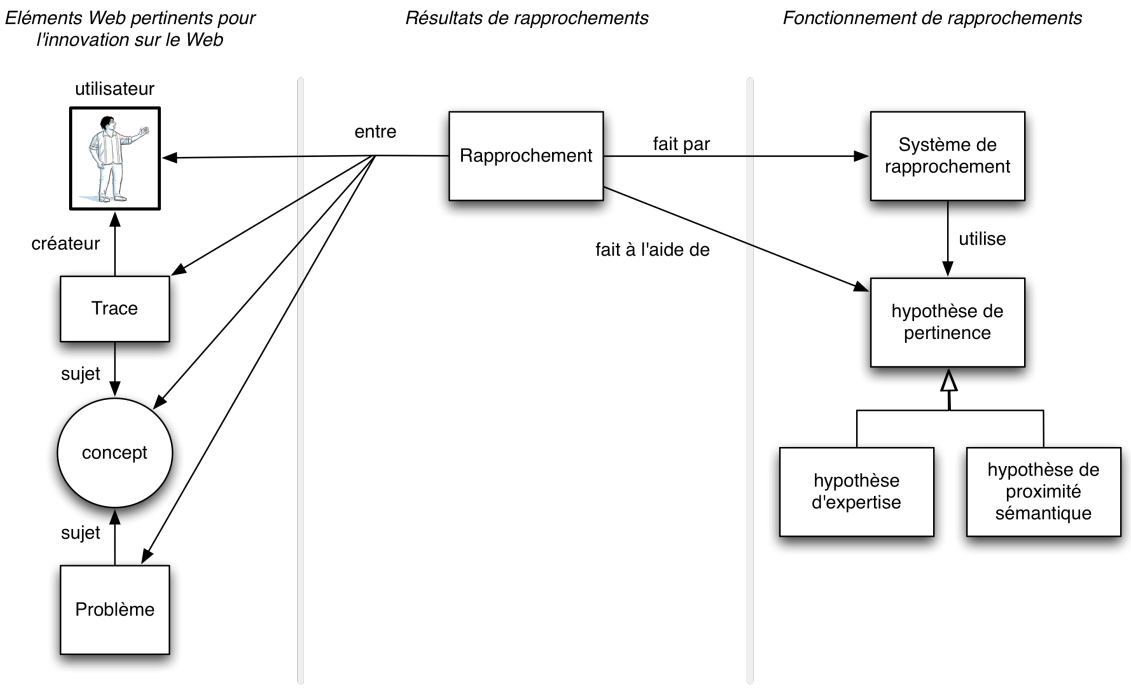


Figure 68 Eléments modélisés et leurs relations

8.1.3. Principes de modélisation

Avant de rentrer dans les détails de nos différents choix de modélisation, il est important de présenter quelques principes qui nous ont guidés dans nos travaux. Ces principes se sont établis au sein de la communauté du Web Sémantique en tant que règles de bonne pratique. Le premier principe est la création d'**ontologies de petite taille**. Cette petite taille impose notamment l'utilisation de structures simples et facilement compréhensibles pour la description de connaissances. Une autre conséquence de ce choix de taille est sans doute la spécialisation de ces ontologies, qui couvrent souvent un domaine très spécifique et restreint, bien maitrisé par leurs créateurs.

Un autre principe important, très compatible avec la pratique de création des ontologies de petite taille, est la **réutilisation**. Ce principe suggère la réutilisation des éléments d'une ou plusieurs ontologies existantes, dont l'utilisation est répandue, dans les nouvelles ontologies, afin de créer des liens et augmenter la compatibilité des données publiées en utilisant des ontologies différentes. Ce principe est utile non seulement pour éviter la redondance mais aussi pour rendre l'espace conceptuel des ontologies du Web Sémantique plus cohérente et plus connexe.

8.2. Modélisation des éléments fondamentaux

Dans cette partie nous présentons des façons de décrire nos éléments Web fondamentaux en RDF à l'aide d'ontologies différentes. Si bon nombre de ces ontologies existaient déjà, certaines sont le fruit de nos travaux. C'est le cas de l'ontologie d'appel aux solutions (PCO).

8.2.1. Modélisation d'utilisateurs

Une des premières ontologies à émerger pour le Web Social fut l'ontologie FOAF. Au fil des années FOAF a gagné en couverture grâce à son processus de développement fortement collaboratif, permettant la participation de toutes les parties intéressées via une liste de diffusion. Cette couverture concerne notamment la capacité de représenter différentes informations apparaissant dans des pages qui représentent des utilisateurs sur divers sites sociaux. FOAF permet de représenter principalement :

- des données démographiques de base (âge, sexe, origine,…) ;
- les liens sociaux (qui connaît qui) ;
- les données de contact et les données permettant d'identifier l'utilisateur dans des sources hétérogènes (adresse e-mail, site Web personnel, etc.) ;
- les intérêts de l'utilisateur ;

Grâce à sa simplicité ainsi qu'à son processus d'évolution très démocratique, cette ontologie est largement utilisée dans le Web de données. Nous allons l'utiliser à la fois pour décrire nos utilisateurs dans les situations qui l'exigent et pour consommer des informations sur les utilisateurs déjà décrites dans le graphe Web.

```
<foaf:Person rdf:about="#milstan" xmlns:foaf="http://xmlns.com/foaf/0.1/">
  <foaf:name>Milan Stankovic</foaf:name>
  <foaf:age>27</foaf:age>
  <foaf:homepage rdf:resource="http://milstan.net/" />
  <foaf:schoolHomepage rdf:resource="http://www.paris-sorbonne.fr" />
  <foaf:mbox rdf:resource="mailto:milstan@gmail.com" />
  <foaf:img rdf:resource="http://milstan.net/images/me.jpg" />
  <foaf:topic_interest
rdf:resource="http://dbpedia.org/resource/Semantic_Web" />
  <foaf:knows rdf:resource="http://ns.inria.fr/fabien.gandon#me" />
</foaf:Person>
```

Figure 69 Description d'un utilisateur à l'aide de l'ontologie FOAF, en RDF (syntaxe XML)

Nous présentons un exemple de description d'utilisateur, fait à l'aide de FOAF dans la Figure 69. Dans cet exemple, le nom et prénom de l'utilisateur sont représentés par la propriété foaf :name, l'âge par foaf :age. La propriété foaf :homepage permet d'indiquer la ou les page(s) Web de l'utilisateur. Cette propriété, ainsi que foaf :mbox utilisée pour l'adresse e-mail, permet d'identifier l'utilisateur de manière unique à travers différentes descriptions qui peuvent le concerner partout dans le Web. Ce type de propriété est dite « inversement fonctionnelle » (en. *inverse functional property*). Grâce à une telle caractéristique, un moteur de raisonnement peut considérer tous les utilisateurs ayant la même valeur pour ces propriétés comme étant en fait le même utilisateur.

Par la propriété foaf :topic_interest, nous avons indiqué que l'utilisateur s'intéresse au Web Sémantique. Pour distinguer ce dernier, nous nous sommes servis d'une ressource désignant ce concept, définie dans DBPedia. Finalement, à l'aide de la propriété foaf :knows nous avons indiqué que l'utilisateur connaissait un autre utilisateur, Fabien Gandon. Ce dernier est identifié par son URI dont le déréférencement pourrait donner accès à plus d'informations structurées.

Si FOAF est largement utilisé, cette ontologie n'est pas sans alternative. Par exemple, une initiative récente, Schema.org[129], essaye de proposer des représentations alternatives de tout élément pouvant être décrit sur le Web, y compris les utilisateurs en faisant donc concurrence à FOAF. Ce projet ambitieux, soutenu notamment par Google et Yahoo, risque d'entrer dans l'usage de masse car ces deux géants de la recherche Web promettent d'indexer toute donnée structurée marquée en utilisant le vocabulaire de Schema.org. Confrontés à l'incertitude quant à l'avenir de ce projet, nous avons préféré utiliser FOAF en tant qu'ontologie confirmée, à laquelle se conforment des masses de données existantes. Si jamais Schema.org devenait prédominant, il serait relativement facile de créer des transformations de données d'un format à l'autre, au moins pour les éléments dont le sens est similaire dans les deux ontologies.

8.2.2. Modélisation de concepts

Si dans ce mémoire nous utilisons très souvent des concepts définis chez DBPedia, il ne faut pas négliger les autres sources de concepts qui peuvent également être utilisées à sa place, telles que Freebase et Yago[130], ni écarter l'hypothèse d'émergence de nouvelles sources, éventuellement plus riches. Une telle source pourrait par exemple émerger du projet récent WikiData[131], visant à construire directement, par l'effort des utilisateurs, un Wikipedia sémantique, éditable de manière collaborative. Une approche similaire est appliquée dans le projet SemanticPedia[132] dont l'objectif est de construire une source francophone, similaire à DBpedia.

Il paraît donc utile d'étudier les structures de connaissances utilisées pour décrire des concepts dans toutes ces sources de concepts afin de pouvoir au mieux les exploiter. Si les concepts sont principalement identifiés par leur URI, il est souvent nécessaire de leur attribuer un titre lisible par des humains. La propriété rdfs :label, définie au niveau du langage RDFS, est souvent utilisée dans ce but. Le fait que sa définition soit fournie par le langage et non pas par une ontologie indépendante fait de cette propriété un choix favorable qui garantit, de manière incontestable, que sons sens puisse être compris par une large gamme de consommateurs de données. Les mêmes raisons motivent l'usage de rdf :comment, souvent utilisée pour fournir une description textuelle courte du concept. Cependant, un certain nombre de propriétés utilisées pour représenter d'autres informations restent spécifique aux sources, qui ont souvent leurs propres ontologies, e.g. DBPedia Ontology[133], Freebase Metaschema[134]. Si une hétérogénéité à ce niveau est inévitable, il est souvent possible de trouver des liens de correspondance entre des concepts représentant les mêmes choses dans des sources différentes. La propriété owl :sameAs, dont le sens fait polémique [Halpin & Hayes, 2010], est souvent utilisée à cette fin. Ceux qui contestnt l'applicabilité d'owl :sameAs, militent pour l'utilisation de rdfs :seeAlso qui n'a pas vocation d'impliquer une égalité de concepts, mais plutôt d'avertir le consommateur de l'existence d'autres concepts ayant un sens potentiellement pertinent ou similaire.

[129] http://schema.org/

[130] http://www.mpi-inf.mpg.de/yago-naga/yago/

[131] http://meta.wikimedia.org/wiki/Wikidata

[132] http://lab.wikimedia.fr/

[133] http://dbpedia.org/Ontology

[134] http://metatopic.freebaseapps.com/about

Pour décrire des relations hiérarchiques entre des concepts, souvent organisés en arborescences, un vocabulaire, SKOS [Miles & Brickley, 2005] (en. Simple Knowledge Organization Shceme) s'est imposé à la fois par sa simplicité et par le soutien institutionnel de W3C qui est derrière ce projet. Les propriétés telles que skos :broader et skos :narrower permettent d'indiquer des concepts ayant une extension plus large ou plus restreinte qu'un concept donné.

Dans la figure 3 nous présentons un exemple de quelques concepts liés par des relations décrites dans SKOS. Dans cet exemple, le concept de « Social Network » recouvre une extension plus large que le concept « Facebook », qui lui-même est pertinent par rapport au concept de « Social Media ». Cette dernière relation skos :related est moins contraignante que skos :broader, mais comporte moins d'informations sur la relation des deux concepts.

```
<skos:Concept rdf:about="http:/example.com/Concept/Facebook">
 <skos:prefLabel>Facebook</skos:prefLabel>
 <skos:broader rdf:resource="http:/example.com/Concept/Social_Network"/>
 <skos:related rdf:resource="http:/example.com/Concept/Social_Media"/>
</skos:Concept>
```

Figure 70 Exemple d'usage du SKOS pour décrire les relations entre des concepts.

8.2.3. Modélisation de traces

8.2.3.1. Ontologies existantes pour des traces

Si la notion de traces utilisateurs représente une abstraction que nous avons introduite dans ce mémoire pour faciliter la manipulation de l'information concernant les utilisateurs et les évidences de leurs intérêts et compétences, les données qui font l'objet de cette abstraction existent indépendamment. Il est possible, comme nous l'avons évoqué dans la partie § 4.2.2, de trouver des sources de données sur le LOD qui contiennent des types de traces différents sous différents formats et avec une richesse informationnelle plus ou moins élevée. Nous présentons ici quelques ontologies qui sont utilisées pour représenter des traces actuellement disponibles dans le graphe Web.

Si les ontologies présentées offrent des points de vue spécifiques en fonction de la nature des différentes traces, il est important de mentionner deux propriétés générales, souvent utilisées, en combinaison avec les propriétés des ontologies spécialisées. Une de ces propriétés est dct :subjet, définie dans l'ontologie Dublic Core Terms[135] qui sert souvent à relier une trace utilisateur à un concept, son sujet. Le lien entre l'utilisateur, créateur de la trace, et la trace elle-même est souvent assuré par la propriété foaf :maker, définie dans FOAF et très couramment utilisée.

8.2.3.1.1. Contenu généré par les utilisateurs

Dans le cadre des premiers efforts pour combiner le pouvoir du Web Social et du Web Sémantique, l'ontologie SIOC (en. *Semantically Interlinked Online Communities*), a émergé en tant que complément naturel de FOAF, qui elle représente uniquement les

[135] http://dublincore.org/documents/dcmi-terms/

données de l'utilisateur et de son réseau social. SIOC a pour objectif de représenter les différents types de contenu que les utilisateurs créent et de capter la richesse des interactions (e.g. commentaires, etc.) autour de ces contenus. Ce projet ambitieux et collaboratif fournit des mécanismes pour représenter (entre autres) :

- Blogs et microblogs,
- Forums et communautés de questions et réponses,
- Canaux IRC,
- Pages wiki.

Dans la Figure 71, nous donnons un exemple de billet représenté par SIOC. Cet exemple, repris du site officiel de l'ontologie, représente déjà dans sa forme originale une trace dans le sens de nos travaux, car il contient à la fois une référence à un concept (Web Sémantique) et une référence à un utilisateur (John Breslin), créateur du billet.

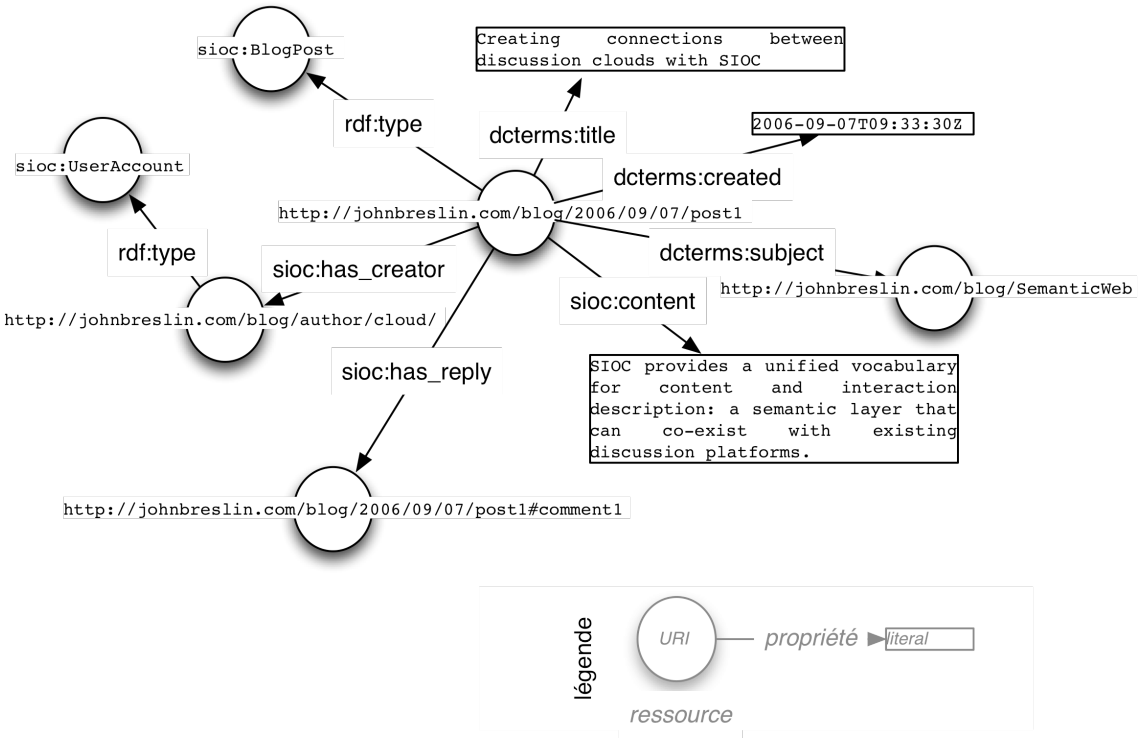


Figure 71 Exemple d'un billet représenté par SIOC, repris du site officiel[136]*.*

Les premières années du Web Social ont également vu naître un engouement pour les sites de sauvegarde de marques-pages. Ces sites permettaient à leurs utilisateurs d'associer les URLs des pages Web avec des mots, dits « tags », pour les classifier. S'il est de toute évidence utile de proposer un service dédié à l'organisation des pages visitées, la nature publique de ces marques-pages les a également rendus utiles pour la recherche de pages pertinentes pour une requête. Si aujourd'hui l'usage de ces sites en en baisse, la pratique du *tagging* reste fondamentalement implantée dans d'autres sites de Web Social, tels que les sites de partage de photos, les blogs etc. Un certain nombre d'ontologies a émergé pour permettre la représentation des tags et notamment de leur

[136] http://rdfs.org/sioc/spec/

sens. Ces représentations visent d'un coté à préserver le contenu textuel du tag et de l'autre coté à capter un sens riche de ce tag, de préférence sous la forme d'un concept.

MOAT[137] est un bon exemple d'une telle ontologie, très bien alignée avec les autres ontologies du spectre du Web Social Sémantique telles que SIOC et FOAF. Cette ontologie enrichit une autre, initialement proposée par Richard Newman, TagOntology[138], dont l'objectif est de représenter le triplet essentiel d'un tag : <utilisateur, ressource, tag> en permettant d'associer un sens précis à ce dernier élément du triplet. Cela permet par exemple de désambigüiser le sens du label « apple », soit comme pomme (fruit), soit comme l'entreprise Apple Inc. Nous donnons un exemple de tag représenté à l'aide de MOAT sur la Figure 72.

```
<tag:RestrictedTagging>   <tag:taggedResource
rdf:resource="http://example.org/post/1"/>   <foaf:maker
rdf:resource="http://example.org/alex"/>   <tag:associatedTag
rdf:resource="http://tags.moat-project.org/tag/apple"/>
<moat:tagMeaning
rdf:resource="http://dbpedia.org/resource/Apple_Inc."/>
</tag:RestrictedTagging>
```

Figure 72 Exemple du tag représenté par MOAT, repris du site officiel[139].

D'autres ontologies de tags ont également été proposées. CommonTag[140] par exemple résulte d'une initiative de plusieurs acteurs industriels (gestionnaires de sites de sauvegarde de marques-pages). L'ontologie proposée s'inspire fortement de MOAT et vise principalement la simplicité d'expression.

L'ontologie NiceTag[141] propose une vision très riche et très ambitieuse du sens des tags. Cette ontologie ne se limite pas à la désambigüisation du sens du tag, mais vise également à désambigüiser le sens de la relation entre le tag et la ressource taguée par l'utilisateur, créateur du tag [Limpens, 2009]. Grace à cette ontologie il est possible de distinguer le cas où un tag se réfère à un concept, e.g. Blog (nicetag :isAbout) de cas où un tag représente le concept lui-même (nicetag :is).

En ce qui concerne les autres types de traces, une ontologie[142] a été proposée pour décrire les avis de consommateurs/utilisateurs sur le Web, notamment dans le cadre du site Revyu qui collectionne ces différents avis. IRC Ontology[143] fut proposée pour représenter des messages instantanés échangés dans des canaux IRC. Les articles de recherche sont le type de trace qui a inspiré peut être la plus grande hétérogénéité de modèles. L'ontologie AKT[144] modélise des articles, leurs auteurs et les projets de recherche. The Bibliographic Ontology[145] se focalise sur les citations qui figurent dans

[137] http://moat-project.org/ontology

[138] http://www.holygoat.co.uk/projects/tags/

[139] http://moat-project.org/ontology

[140] http://commontag.org/Home

[141] http://ns.inria.fr/nicetag/2009/09/25/voc.html

[142] http://purl.org/stuff/rev#

[143] http://fabien-gandon.blogspot.fr/2008/09/irc-ontology-rdf-vocabulary-to.html

[144] http://www.aktors.org/ontology/

[145] http://purl.org/ontology/bibo/

les articles. L'ontologie UDO[146] (en. *Unified Discours Ontology*) modélise différents types de citations, ainsi que différentes données pouvant être présentées dans un article pour soutenir un discours, avec l'objectif final de faciliter la traçabilité du discours à travers les articles. L'ontologie SWCO[147] (en. Semantic Web Conference Ontology) permet de décrire des articles de recherche dans le contexte de leur présentation lors d'un congrès.

Parmi ses nombreux types, Schema.org fournit également des moyens de représentation des contenus générés par les utilisateurs (articles, blogs, etc.).

8.2.3.1.2. Activités et accomplissement des utilisateurs

L'ontologie DOAP (en. *Description of a Project*) permet de décrire des projets (notamment des projets de développement de logiciel) et la participation des utilisateurs dans ceux-ci. La notion d'expérience professionnelle est également prise en compte dans FOAF avec la propriété foaf :workspaceHomepage, mais cette propriété, si elle renseigne l'employeur, offre néanmoins peu d'informations sur la nature de l'emploi et de la compétence engagée.

SWCO permet également de distinguer des membres reconnus de la communauté scientifique par leur rôle dans les évènements académiques (participation à des comités de programme, etc.).

L'ontologie AKT est également utilisée dans les descriptions de projets de recherche financés par les fonds publics d'EU dans le cadre du jeu de données CORDIS[148]. Les financements étant accordés par l'intermédiaire d'un concours, le fait de gagner une telle compétition est un témoignage *per se* de la reconnaissance portée aux membres du projet.

Le vocabulaire Open Graph[149] de Facebook permet de représenter différentes activités (e.g. d'avoir aimé quelque chose) des utilisateurs.

8.2.3.1.3. Réputation des utilisateurs

Bien qu'un certain nombre d'ontologies visant à représenter la réputation des utilisateurs en ligne ait été proposé (ex. [Chang et al., 2006]), les systèmes s'appuyant sur une notion de réputation dans la recherche d'experts utilisent souvent leur propres définitions de la réputation et reviennent donc à l'usage directe des traces peu structurées, telles que du contenu généré par les utilisateurs et leurs interactions avec ces contenus, disponibles dans des formats qui pré-datent le Web Sémantique.

8.2.4. Modélisation de problèmes

Au début de nos travaux, la notion de plates-formes de résolution de problèmes sur le Web était relativement nouvelle, et il n'y avait pas à notre connaissance de vocabulaire permettant de décrire des problèmes. Nous avons donc développé une ontologie appropriée, PCO (en. *Problem Challenge Ontology*) que nous décrivons par la suite.

Si notre conception du modèle visait principalement à répondre aux besoins de nos rapprochements, il était également nécessaire de soutenir d'autres scénarios liés au

[146] http://vocamp.org/wiki/UDO_(Unified_Discourse_Ontology)

[147] http://data.semanticweb.org/ns/swc/swc_2009-05-09.html

[148] http://cordis.rkbexplorer.com/

[149] http://ogp.me/

fonctionnement des plates-formes et à l'intégration des problèmes issus de plusieurs plates-formes hétérogènes. Nous présentons ces scénarios par la suite.

Il est important de noter que l'ontologie que nous avons conçue représente principalement les différents aspects de la résolution d'un problème et ne concerne pas la description détaillée des différents défis de recherche et développement du dit problème. Nous avons été amenés à faire ce choix en raison de la nature diverse et imprévisible de ces défis : une plate-forme de résolution de problèmes peut en effet héberger des problèmes de toutes disciplines. Leur modélisation précise serait difficile à effectuer en raison principalement du nombre relativement faible de problèmes disponibles en ligne dans toutes les plates-formes.

8.2.4.1. Scénarios d'usage de données

Dans notre analyse d'usages, nous avons identifié quatre scénarios principaux pour lesquels une ontologie d'appel aux solutions peut servir.

Scénario 1 : Une application A1 consomme des informations sur l'appel aux solutions de la plate-forme de résolution de problèmes.

Des renseignements concernant la description du problème, les modalités de solutions acceptées, le prix, les conditions d'acceptation des solutions et de leur classement etc. peuvent être pertinents pour ce scénario.

Scénario 2 : Devant un référentiel de descriptions d'appels aux solutions, provenant potentiellement de plusieurs sources, un utilisateur cherche des problèmes auxquels il peut répondre.

Les modalités de transfert de la propriété intellectuelle peuvent être notamment importantes dans ce scénario. Les experts d'aujourd'hui, souvent déjà contraints par des contrats d'exclusivité et de non concurrence, pourraient s'intéresser uniquement aux appels aux solutions compatibles avec leur situation juridique.

Scénario 3 : Une application, en possession de certaines informations identifiant les compétences potentielles d'un utilisateur, cherche quels appels à solutions recommander à l'utilisateur.

Les concepts d'un problème sont notamment importants pour ce scénario, ainsi qu'un éventuel historique des solutions de l'utilisateur pouvant faciliter la découverte de ses préférences.

Scénario 4 : Un utilisateur travaille sur un problème d'innovation. Une application cherche à le soutenir en lui suggérant des ressources et des collaborateurs pertinents.

Plusieurs alternatives, plus précises, de ces scénarios généraux sont disponibles sur notre site dédié à la conception de l'ontologie PCO[150].

8.2.4.2. Ontologie PCO

Nous avons conçu l'ontologie PCO en tenant compte des besoins informationnels de nos scénarios, mais aussi en analysant le contenu informationnel des différents problèmes publiés sur les 7 plates-formes actives à l'époque de nos travaux : Hypios,

[150] http://research.Hypios.com/neologism/node/55

Innocentive[151], NineSigma[152], Presans[153], InnovationExchange[154], InnoGet[155] et Ideaken[156]. L'ontologie résultante est décrite dans une spécification Web[157], d'où une version légère est téléchargeable. La version complète est disponible en annexe de ce mémoire. Un document supplémentaire détaillant nos différents choix de modélisation est également disponible sur le site de l'ontologie[158].

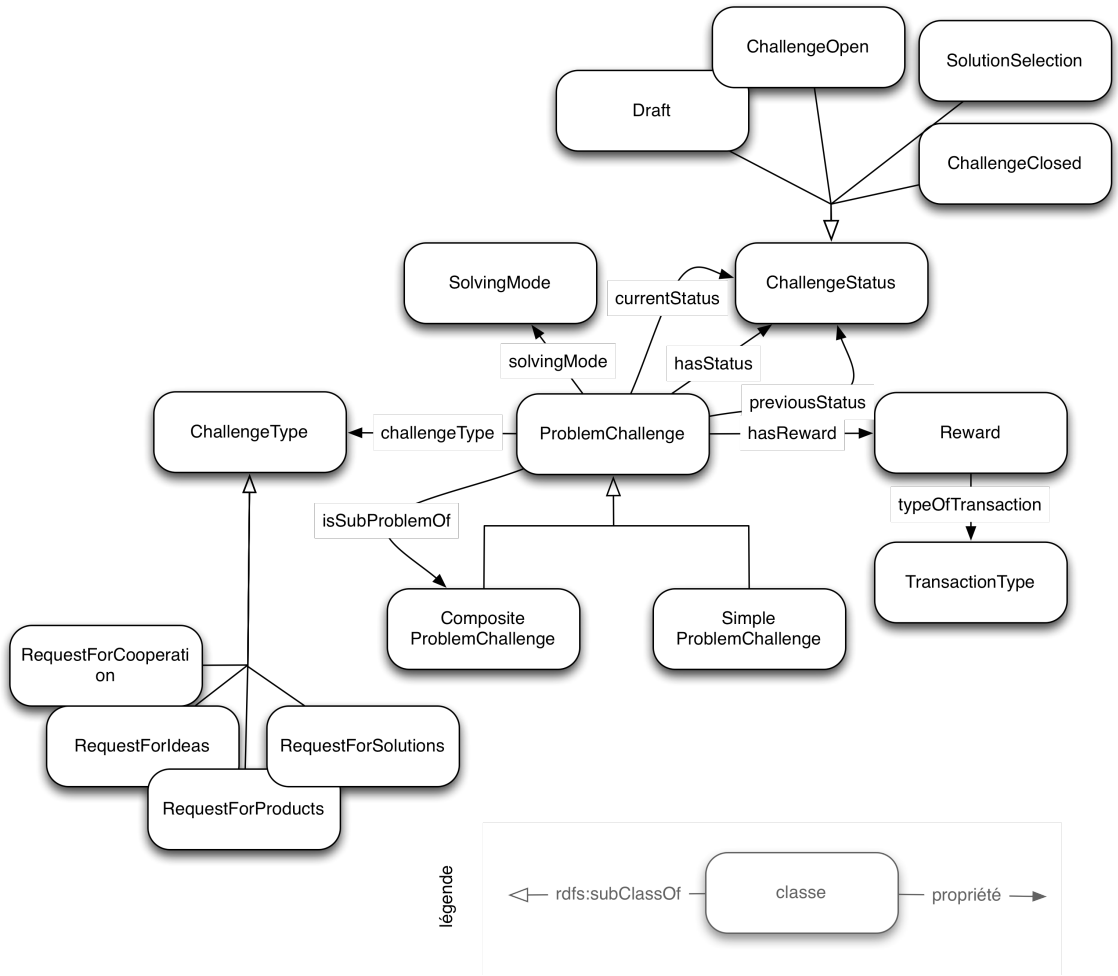


Figure 73 Classes principales de l'ontologie PCO

Les classes principales de l'ontologie sont représentées dans la Figure 73. La notion centrale est celle d'appel aux solutions (classe ProblemChallenge). Deux sous-classes sont disponibles pour représenter des problèmes simples (SimpleProblemChallenge) et complexes (CompositeProblemChallenge) composés de plusieurs problèmes. Ces derniers sont souvent présents chez Hypios. La classe SolvingMode permet, pour un problème donné, de décrire la modalité de résolution prise en charge par la plate-forme.

[151] http://www.innocentive.com/
[152] http://www.ninesigma.com/
[153] http://presans.com/
[154] http://www.innovationexchange.com/
[155] http://www.innoget.com/
[156] http://www.ideaken.com/
[157] http://research.Hypios.com/neologism/pb
[158] http://research.Hypios.com/neologism/node/56

La majorité des plates-formes fournissent deux modes : résolution individuelle ou en équipe. La classe ChallengeType permet de décrire le type d'appel aux solutions en fonction de la nature du résultat qu'attendent les clients. En analysant les problèmes sur différentes plates-formes, nous avons identifié quatre types qui couvrent tous les problèmes analysés :

- **RequestForCooperation**. Le client cherche des collaborateurs pour développer la solution à plusieurs. Il s'attend à des pistes de solutions et autres justificatifs démontrant la qualité du futur collaborateur.
- **RequestForIdeas**. Le client souhaite récupérer le plus grand nombre d'idées sur les possibles solutions, pour ensuite les trier ou bien pour s'en inspirer.
- **RequestForProducts**. Le client cherche des solutions déjà existantes qu'il souhaite acheter.
- **RequestForSolutions**. Le client cherche des solutions à son problème, souvent avec des arguments démontrant l'applicabilité de la solution et parfois même avec un prototype ou autre preuve du concept.

La classe ChallengeStatus permet de décrire l'état actuel de l'appel à solution. La plupart des cycles de vie d'un problème sur ces plates-formes peuvent être décrits avec les quatre phases que nous avons créées : Draft, ChallengeOpen, SolutionSelection, et ChallangeClosed. Ces phases sont représentées en tant que classes. Leurs instances concrètes représenteront des phases concrètes avec leurs dates de début et de fin. A différents moments, la phase actuelle (currentStatus) changera ; la propriété hasStatus permet de suivre toutes les phases. Le fait que currentStatus soit une sous-propriété de hasStatus permet également d'inférer qu'une phase actuelle est aussi une phase.

Finalement la classe Reward permet de décrire le prix proposé pour les solutions acceptées. Plusieurs prix peuvent être proposés dans un seul appel. Il dépend parfois de la satisfaction du client quant à la solution retenue, mais plus souvent de la nature du transfert de propriété intellectuelle (TransactionType). Évidemment les prix donnés en échange du simple droit d'usage d'une idée seront différents de ceux accordés en échange des droits d'exclusivité d'usage des technologies proposées.

Un exemple de problème représenté par l'ontologie PCO est donné dans l'Annexe 1.

8.3. Modélisation des hypothèses

Pour décrire le fonctionnement des différents systèmes de rapprochement des éléments Web, il était nécessaire de modéliser les hypothèses de ces systèmes et la nature de la pertinence des éléments qu'ils essayent de rapprocher. Nous avons donc conçu l'ontologie RHO (en. *Relevance Hypothesis Ontology*), dont les classes principales et leurs relations sont présentées sur la Figure 74. La notion centrale de cette ontologie est la classe RelevanceHypothesis, qui sert à décrire et distinguer la nature des différentes notions de pertinence. Généralement, la pertinence est calculée entre un élément initial (ou un ensemble d'éléments initiaux) et les éléments considérés comme potentiellement pertinents. Une hypothèse de pertinence peut opérer sur différents types d'éléments, et les propriétés rho :typeOfInitialElements et rho :typeOfRelevantElements permettent de définir le type de ces éléments (initiaux et ceux considérés pertinents).

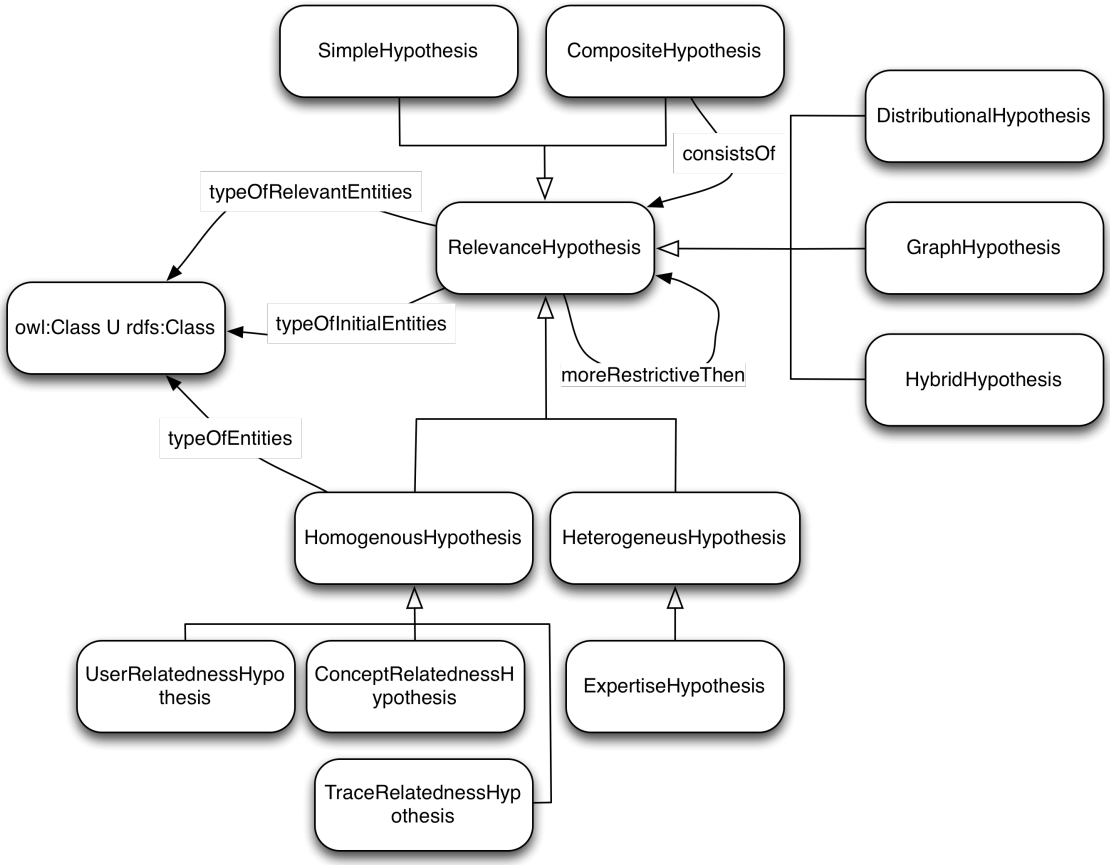


Figure 74 Classes principales de l'ontologie d'hypothèses de pertinence

Une hiérarchie de classes d'hypothèses introduit encore un point de distinction parmi celles-ci. Une branche de cette hiérarchie permet ainsi de distinguer des hypothèses selon le type de rapprochement dans lequel elles interviennent. Les hypothèses de type ExpertiseHypothesis interviennent notamment dans le rapprochement des utilisateurs et concepts ; UserRelatednessHypothesis participe dans le calcul de la proximité des utilisateurs ; ConceptRelatednessHypothesis dans la découverte des concepts pertinents, alors que TraceRelatednessHypothesis joue un rôle important dans la mise en correspondance des traces utilisateurs. Ces types sont également catégorisés en HomogeneusHypothesis pour des hypothèses qui calculent la proximité des entités d'un même type, et en HeterogenousHypothesys pour celles qui opèrent sur des entités hétérogènes.

Une autre branche de la hiérarchie de classes d'hypothèses permet de distinguer la nature de l'hypothèse en tant que distributionnelle, basée sur le graphe ou hybride. Si cette classification provient de la classification standard des mesures de proximité sémantique utilisées pour des concepts, elle est également applicable à d'autres proximités, telles que la proximité des utilisateurs.

Il est également possible de distinguer des hypothèses simples (SimpleHypothesis) de celles composées de plusieurs autres hypothèses (CompositeHypothesis).

Une hypothèse concrète peut donc aussi être l'instance de plusieurs de ces classes. Il est ainsi possible d'imaginer une hypothèse simple distributionnelle utilisée pour calculer la proximité des utilisateurs. Une telle hypothèse serait l'instance des classes : UserRelatednessHypothesis, SimpleHypothesis et DistributionalHypothesis.

Finalement, grâce à la propriété rho :moreRestrictiveThen il est possible de comparer deux hypothèses par leurs niveaux de restrictivité. Une telle nécessité a été évoquée dans la partie §4.4.3 pour faciliter les utilisateurs à choisir des hypothèses d'expertise. Cette notion de restrictivité peut s'appliquer aux autres types d'hypothèses.

L'ontologie RHO est disponible sur le Web[159]. Cette ontologie est notamment utilisée par nos différents outils pour décrire les hypothèses utilisés pour effectuer nos rapprochements en vue de la mise à disposition de résultats de ceux-ci dans les formats structurés.

Ci-après nous présentons quelques exemples d'hypothèses en RDF à l'aide de nos ontologies. Ces exemples concernent notammfrent une hypothèse d'expertise et une hypothèse de proximité conceptuelle. D'autres exemples d'hypothèses peuvent être trouvés dans l'Annexe 2 et Annexe 3.

8.3.1. Hypothèses d'expertise

La notion d'hypothèse d'expertise a initialement été introduite dans le cadre de nos travaux préliminaires dans une ontologie indépendante, EHO (en. Expertise Hypothèsis Ontology), disponible sur le Web[160]. La description du processus de sa conception et des questions de compétence sont également consultables[161]. Si cette notion s'inscrit bien dans le cadre général des hypothèses de pertinence, elle a ses propres spécificités, définies dans EHO. Par exemple, comme nous avons pu le voir dans le Chapitre 4, un point de distinction important parmi les hypothèses d'expertise est le type (ou les types) de trace qu'elles utilisent pour effectuer une inférence sur les compétences potentielles d'un utilisateur. La propriété eho :hasTraceType permet de spécifier ce type de trace utilisé par l'hypothèse.

Un exemple d'hypothèse d'expertise décrite par cette ontologie est donné sur la Figure 75. Une représentation textuelle de celle-ci est donnée par la propriété eho :hasTextualStatement. Les types de traces utilisateurs que cette hypothèse exploite sont exprimés par la propriété eho :hasTraceType. Dans ce cas, l'hypothèse tient compte des publications scientifiques de l'utilisateur. L'hypothèse est appliquée en effectuant une requête SPARQL qui permet d'évaluer ses conditions sur un graphe de données. Cette requête peut être instruite via la propriété eho :hasRuleURL permettant de spécifier un document Web la contenant.

[159] http://research.Hypios.com/neologism/rho

[160] http://research.Hypios.com/neologism/eho

[161] http://research.Hypios.com/neologism/node/85

```
@prefix swrc: <http://swrc.ontoware.org/ontology-07#>.
@prefix eho: <http://ontologies.Hypios.com/eho#>.
@prefix swc: <http://data.semanticweb.org/ns/swc/ontology#>.
@prefix dct: <http://purl.org/dc/terms/> .

<http://example.org/hypothesis1> a eho:ExpertiseHypothesis ;
      rho:hasTextualStatement "If a user wrote a scientific publication
on topic X than he is an expert on topic X" ;
       eho:hasTraceType swrc:Publication, swrc:InProceedings, swc:Paper ;
      eho:hasRule :r1_1 .
:r1_1 a eho:ExpertiseHypothesisRule;
      eho:hasFormat :sparqlQueryFormat ;
      eho:hasRuleURL            <http://Hypios.com/expert-search/example-
rules/rule_1_1> ;
       eho:hasParameter :p1 1.
```

Figure 75 Un exemple d'hypothèse d'expertise représentée par EHO, format Turtle

Le système hySemEx, décrit dans la partie §4.4, utilise ce type de description des hypothèses d'expertise. Il peut les importer en format RDF, les proposer à un utilisateur et les exécuter sur un EDS choisi grâces aux requêtes SPARQL fournies dans les descriptions.

8.3.2. Hypothèses de proximité conceptuelle

L'exemple présenté sur la Figure 76 propose la description RDF d'une hypothèse de proximité conceptuelle, faite à l'aide de l'ontologie RHO. La description, assez simple, permet principalement de distinguer des hypothèses par leur type. Dans ce cas nous présentons l'hypothèse qui est au cœur de notre approche de recommandation de concepts par la fonction de distance transversale que nous avons décrite dans le chapitre 5.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix db: <http://www.dbpedia.org/ressource>.
@prefix rho: <http://Hypios.com/ontologies/rho#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.

<http://example.org/hypothesis2> rdfs:label "The Transversal Proximity
Hypothesis";
                                  rdf:type
rho:ConceptRelatednessHypothesis,                rho:GraphHypothesis,
rho:SimpleHypothesis.
```

Figure 76 Un exemple d'hypothèse de la proximité conceptuelle représentée par RHO

8.4. Modélisation des proximités

Pour permettre l'échange des résultats de rapprochement effectués par les différents outils, y compris ceux développés dans le cadre de cette thèse il est nécessaire de les présenter dans une forme structurée et descriptive. Si les éléments ontologiques décrits préalablement permettent déjà de décrire des entités rapprochées ainsi que la nature des rapprochements, la représentation des résultats de rapprochements i.e. des proximités reste un défi que nous abordons ici.

Dans cette partie nous illustrons l'utilisation de l'ontologie RHO pour décrire des rapprochements sur deux exemples. Un autre exemple est disponible dans l'Annexe.

8.4.1. Proximité expert-concept

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix db: <http://www.dbpedia.org/ressource>.
@prefix rho: <http://Hypios.com/ontologies/rho#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.

<http://www.example.org/proximities> rho:forInput [
a rdf:Seq;
rdf:_1 <http://www.example.org/Milan_Stankovic>.];

                                                rho:proximities [
a rdf:Seq;
rdf:_1 [a rho:WeightedResource; rho:resource db:Semantic_Web; rho:weight
0.567.];
rdf:_2 [a rho:WeightedResource; rho:resource db:Social_Web; rho:weight
0.567.];
rdf:_3 [a rho:WeightedResource; rho:resource db:Twitter; rho:weight
0.567.].];
                                                rho:hypothesisUsed
<http://example.org/hypothesis1>;
                                                rho:systemUsed
```

Figure 77 Un exemple de résultats de rapprochement de concepts et experts décrits par l'ontologie RHO

Un exemple de résultat de rapprochement de concepts et experts, conséquence de l'application d'une hypothèse d'expertise, est présenté à la Figure 77. Le code RDF illustre des propriétés introduites pour relier des éléments à des résultats. Par exemple la propriété rho:forInput permet de décrire les entrées (dans ce cas, l'utilisateur pour lequel les concepts d'expertise sont sollicités). La propriété rho:proximities permet de décrire des ressources considérées proches par l'hypothèse (données par la propriété rho:hypothesisUsed). Ces ressources sont représentées sous la forme d'une liste ordonnée des entités de type rho:WeightedResource qui permet de décrire une ressource par son URI et son importance numérique dans le classement donné.

Notre système hySemEx est capable de produire ce format de résultats de rapprochements et de les transmettre aux autres systèmes.

8.4.2. Proximité concept-concept

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix db: <http://www.dbpedia.org/ressource>.
@prefix rho: <http://Hypios.com/ontologies/rho#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.

<http://www.example.org/proximities> rho:forInput db:Semantic_Web;

                                        rho:proximities [
a rdf:Seq;
rdf:_1       [a       rho:WeightedRessource;       rho:resource       db:
Resource_Description_Framework; rho:weight 0.567.];
rdf:_2  [a  rho:WeightedRessource;  rho:resource  db:Web_Ontology_language;
rho:weight 0.567.];
rdf:_3   [a   rho:WeightedRessource;   rho:resource   db:SPARQL;   rho:weight
0.567.].];
                                                    rho:hypothesisUsed
<http://example.org/hypothesis2>;
                                                    rho:systemUSed
<http://Hypios.com/hyProximity>.
```

Figure 78 Un exemple de résultats de rapprochement de concepts décrit par l'ontologie RHO

Similairement à l'exemple précédent, la description des proximités entre concepts résultant de l'application d'une hypothèse de pertinence conceptuelle prend la même forme, présentée dans l'exemple sur la Figure 78.

Ce format permet aux utilisateurs de hyProximity d'exporter les résultats de leurs rapprochements de concepts, puis de les réutiliser.

8.5. Un scénario d'usage des données

Clairement, l'utilité des résultats des rapprochements et des distances calculées pour les effectuer dépasse le cadre de la situation particulière pour laquelle ils ont été faits. Imaginons un scénario, représenté visuellement par la Figure 79, dans lequel un système effectuerait un rapprochement de concepts, par les méthodes décrites dans le Chapitre 5, afin de trouver des concepts pertinents pour un problème. Le résultat d'une telle opération serait des concepts pertinents, mais aussi des valeurs de proximités, calculées par une fonction de proximité particulière.

Imaginons ensuite un utilisateur *u1* qui travaillerait sur une solution pour le problème de l'exemple précédent. L'utilisateur *u1* cherche des collaborateurs (scénario similaire à celui traité dans le Chapitre 7) qui complèteraient ses compétences. Si la recherche de collaborateurs est basée à la fois sur la similarité conceptuelle avec le problème en question et la similarité sémantique, la proximité des concepts calculée dans la phase précédente pourra de nouveau être utile. Il est donc utile de publier des résultats intermédiaires dans la première phase en tant que distances entre concepts.

En raison d'une potentielle intégration des proximités pré-calculées dans d'autres calculs, il est très important de pouvoir associer la description des mesures de proximité et algorithmes utilisés afin de pouvoir estimer la compatibilité de ces résultats avec le calcul souhaité. Nos ontologies sont donc conçues dans cet esprit et permettent la comparaison des métriques de distance par leurs différentes caractéristiques. Une pratique de partage des résultats en ce format très riche pourrait réduire le temps de calcul tout en gardant un esprit distribué et ouvert, si cher au Web.

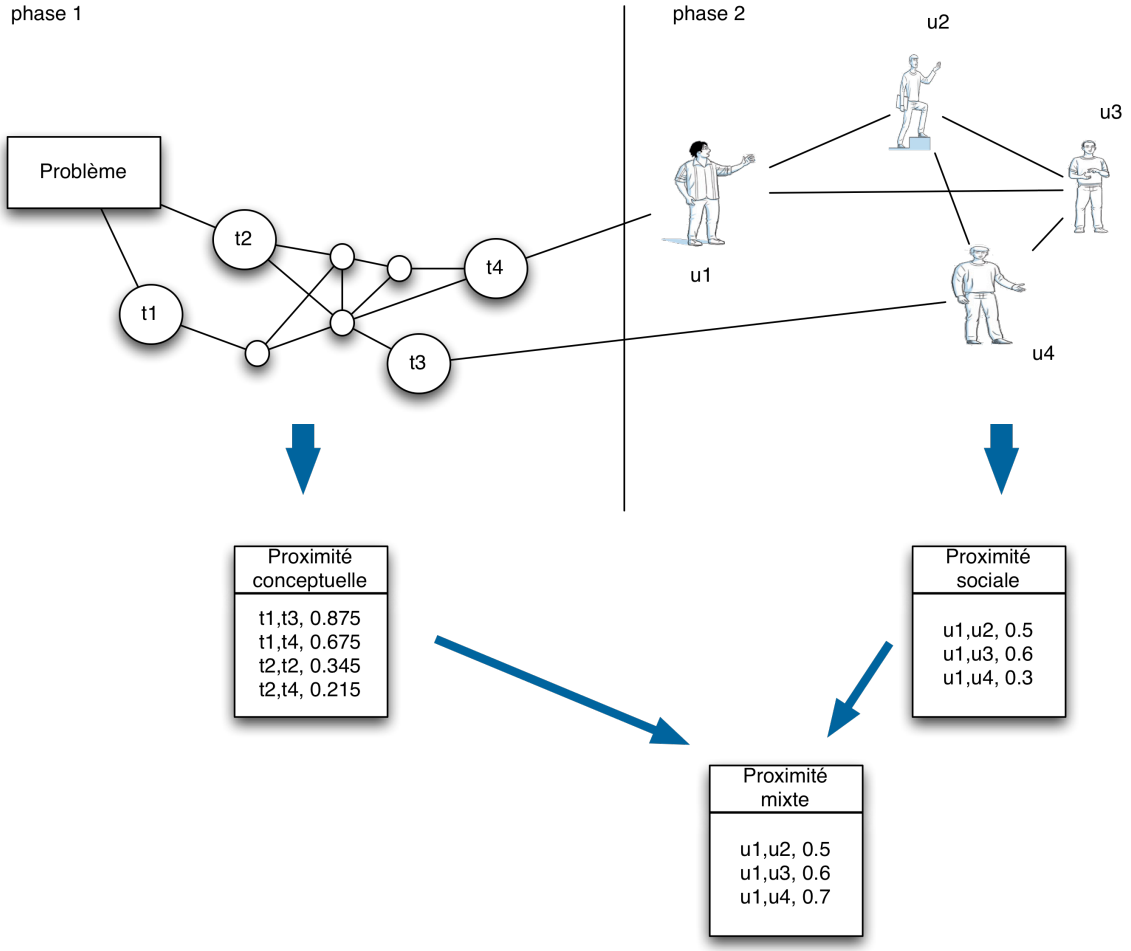


Figure 79 Scénario de réutilisation des proximités

8.6. Conclusions

Dans ce chapitre nous avons présenté les ontologies correspondant aux éléments informationnels (problèmes, utilisateurs, traces et concepts) qui participent aux scénarios liés à l'innovation. Guidé par le principe de réutilisation, nous avons toujours préféré l'utilisation des ontologies existantes pour décrire ces éléments, comme par exemple dans le cas de l'ontologie FOAF qui couvre largement les besoins de représentation des utilisateurs en format RDF. Dans d'autres cas comme la description des problèmes d'innovation, nous avons été amené à créer des ontologies. Nos travaux de modélisation ne se limitent pas à des éléments initiaux, mais cherchent également à couvrir l'ensemble des processus et des résultats de nos scénarios. De telles descriptions peuvent permettre la réutilisation des résultats de rapprochements sans trop de problèmes. C'est dans ce but que, nous avons présenté à la fin du chapitre un scénario de

réutilisation de résultats qui met en avant l'utilité des descriptions riches et structurées des ontologies.

Si l'objectif principal de ce chapitre était de présenter les différentes modélisations, il récapitule également les différents éléments du processus de rapprochement qui nous ont permis de construire une nouvelle vue sur le rapprochement lui-même.

Bibliographie

[Chang et al., 2006] Chang, E., Hussain, F. K., & Dillon, T. (2006). On the Move to Meaningful Internet Systems 2006: *OTM 2006 Workshops*. (R. Meersman, Z. Tari, & P. Herrero, Eds.) (Vol. 4278, pp. 1724-1733). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from http://dl.acm.org/citation.cfm?id=2148751.2148861

[Halpin & Hayes, 2010] Halpin, H., & Hayes, P. J. (2010). When owl : sameAs isn ' t the Same : An Analysis of Identity Links on the Semantic Web. *Proceedings of Linked Data on the Web 2010*, on WWW2010.

[Limpens et al., 2009] Limpens, F., Monnin, A., Laniado, D., & Gandon, F. (2009). NiceTag Ontology: tags as named graphs. *International Workshop in Social Networks Interoperability*. Retrieved from http://hal.inria.fr/hal-00434129/

[Miles & Brickley. 2005] Miles, A., & Brickley, D. (2005). SKOS Core Vocabulary Specification. W3C. Retrieved from http://www.w3.org/TR/2005/WD-swbp-skos-core-spec-20051102/

9. Conclusions

> *Dans ce chapitre nous présentons nos conclusions générales. Ce chapitre donne notamment l'occasion de rappeler nos hypothèses de départ et d'estimer si celles-ci ont pu être confirmées par nos travaux (§9.1). Ce chapitre est aussi dédié à une analyse d'applicabilité de nos travaux dans d'autres cas d'utilisation (§9.2). Nous y présentons également des pistes pour travaux futurs, et des problématiques émergentes dans le domaine d'innovation sur le Web qui pourraient inspirer des thèses dans l'avenir (§9.3). Finalement nous présentons notre vision de l'avenir du Web Social Sémantique et des sujets de recherches qui pourraient émerger dans l'avenir de cette discipline (§9.4).*

9.1. Conclusions par rapport aux hypothèses de départ

A la fin d'une thèse, il est pertinent de rappeler nos hypothèses de départ et d'essayer de comprendre si nos travaux ont permis de les confirmer. Dans le chapitre 3, nous avons introduit l'idée que le Web Social Sémantique, de par sa structure riche, facilite les rapprochements et par conséquence améliore la pertinence des informations que les utilisateurs reçoivent du Web. Plus précisément, nous avons proposé l'hypothèse suivante :

> *En s'appuyant sur la diversité des types de ressources et des types de liens entre eux, il est possible de concevoir des systèmes capables de rapprocher deux éléments Web de manière utile par rapport à une situation particulière.*

Nous avons essayé de confirmer cette hypothèse générale dans des scénarios liés à l'innovation sur le Web. Ce contexte impose des besoins spécifiques que nous avons introduit dans §3.4.3.1, notamment la sérendipité et la découverte. Dans le chapitre 4, nous avons tenté de rapprocher des utilisateurs aux concepts de leurs compétences,

définis dans le LOD. Nos travaux ont montré une utilité du graphe de données qui émerge sur le Web pour cette tâche de rapprochement. Ce graphe, de par sa richesse des utilisateurs de types différents, introduit une flexibilité importante dans la recherche d'experts. En effet, il permet d'appliquer différentes approches de recherche d'experts, dont les ressources sont abondantes dans la littérature, à partir d'informations disponibles sur le Web. Nous avons également montré que grâce à la structure riche de données sur le graphe Web, il est possible d'orienter la recherche d'experts vers une approche bien adapté à un sujet d'intérêt. Ceci nous a permis de confirmer une autre hypothèse : au sein de différentes communautés de pratique, différents types de traces peuvent être considérées comme crédibles en raison de différents canaux de communication utilisés par des experts.

Dans le Chapitre 5 nous avons montré comment les différents types de liens entre des concepts peuvent être utilisés pour concevoir une notion de proximité de concepts, comme par exemple dans le cadre de la diffusion de problèmes. Cette richesse de types de liens apporte des résultats concluants pour la découverte des concepts. Ils sont à la fois pertinents et inattendus, surtout par rapport aux approches de l'état de l'art qui se basent sur la cooccurrence et n'utilisent que la couche de documents. Nous avons également montré comment ce rapprochement de concepts peut être utilisé au sein d'une plate-forme de résolution de problèmes en ligne. Ces rapprochements nous ont permis de stimuler la visibilité des problèmes au sein de communautés directement concernées et qui sont susceptibles de pouvoir transposer des connaissances d'un « domaine » à l'autre, i.e. de stimuler la sérendipité.

Dans le Chapitre 6 nous avons montré l'utilité de l'introduction des nouveaux liens entre des traces utilisateurs au sein du graphe Web, notamment pour la création des profils d'experts et le suivi des intérêts des internautes. La structure du graphe, riche, cohérente et détaillée, permet donc d'améliorer le traitement de l'information disponible sur le Web dans certains scénarios.

Le Chapitre 7 concerne la recommandation de collaborateurs. Il est dédié à l'exploitation des liens sociaux entre utilisateurs, ainsi que l'exploitation des liens établis à travers des concepts dont les utilisateurs sont mutuellement intéressés. Dans ce cas de figure, nous avons montré qu'une notion de proximité sémantique de concepts, dérivée de la structure de graphe Web, a des effets bénéfiques pour la recommandation de collaborateurs potentiels.

Dans le Chapitre 8, nous démontrons comment les descriptions complètes de rapprochements effectués dans le cadre des scénarios traités dans les chapitres précédents, peuvent êtres utiles dans le cadre de la réutilisation des résultats de rapprochements pour de nouveaux processus de rapprochement.

Toutes ces conclusions indiquent l'utilité de la richesse des données du graphe Web appliquée à des scénarios liés à l'innovation. Cette utilité dépasse souvent, comme nous l'avons montré dans nos évaluations, l'utilité de traitement de la couche de documents du Web. Nous pouvons donc constater que notre hypothèse de départ a été confirmée pour ces scénarios, ce qui était l'un des objectifs de cette thèse, mais également pour le contexte industriel. La généralité de l'hypothèse de départ laisse certainement quelques défis pour des travaux visant à confirmer dans d'autres scénarios, non liés à l'innovation. Nous abordons ces défis dans les paragraphes qui suivent.

9.2. D'autres champs d'applicabilité de nos travaux et les travaux futurs

Il est clair que le cadre conceptuel de rapprochement que nous avons présenté dans le chapitre 3 est bien plus général que ses applications concrètes dans cette thèse. Il est donc possible d'appliquer le principe du rapprochement d'éléments du graphe Web, en tenant compte de la richesse structurelle du Web Social Sémantique, dans d'autres cas d'utilisation tels que :

- Les systèmes de diffusion de contenu pour mieux cibler un public large;
- La recherche d'experts à partir des sources fermées, i.e. au sein d'entreprises, pour mieux orienter la recherche en fonction d'une situation particulière;
- L'elargissement des requêtes dans les systèmes de recherche pour stimuler les découvertes et la sérendipité;
- Les systèmes de recommandation pour diversifier les recommandations et les rendre plus utiles en favorisant les découvertes pertinentes, etc.

Par la suite, nous illustrons quelques exemples d'applications possibles liées à ces cas d'utilisation et nous présentons nos plans pour des travaux futurs.

9.2.1. La recherche d'experts au sein d'entreprises

Hypios a établi au cours de son existence de nombreux contact commerciaux avec ses clients. Face à la vision des solutions technologiques qui font partie des prestations d'Hypios, les clients ont exprimé leur besoin d'appliquer la démarche d'Hypios au sein de leur entreprises Les grands groupes ont souvent des opérations dispersées dans plusieurs pays et ils rencontrent souvent des difficultés dans l'attribution des tâches et des projets aux experts internes. Ils manquent souvent de visibilité sur les compétences réelles de leurs employés. Ce problème est encore plus grave dans le cas de l'acquisition d'une entreprise par une autre. Les compétences des équipes acquises sont souvent difficiles à connaître et à gérer.

De nombreuses solutions de gestion, de connaissances, et de réseaux sociaux d'entreprises existent dans la recherche et dans l'industrie sous la forme de produits commerciaux. Mais à notre connaissance, il n'y a pas encore eu de démarche proposant une recherche pointue d'experts, une démarche qui est finalement similaire à celle que nous avons proposée dans le Chapitre 4. Il serait donc intéressant d'appliquer nos solutions orientées sur la découverte et la sérendipité des données internes des entreprises. Ces données, très différentes des données qui se trouvent sur le Web de par leur nature confidentielle, sont dotées d'une richesse importante. Une fois les données structurées, il serait envisageable d'appliquer notre approche de recherche d'experts et testr la vraie applicabilité de ce cas d'utilisation. Nous envisageons d'ailleurs d'adapter nos travaux de rapprochements pour les entreprises dans le cadre du développement de futurs produits Hypios. Ils viseront particulièrement le marché des entreprises recherchant des outils d'innovation.

9.2.2. Suggestion de concepts pour des campagnes publicitaires en ligne

Si notre notion de proximité de concepts a été développée particulièrement pour des scénarios d'innovation, et notamment pour la diffusion de problèmes en ligne, il est clair qu'une telle notion peut être utile pour d'autres tâches de diffusion de contenus sur le Web. Les campagnes publicitaires sur le Web visent à exposer des contenus, notamment des pages Web et des publicités, à un public large et adapté. Il paraît donc opportun d'appliquer notre notion de proximité de concepts dans la suggestion de mots-clés à des fins publicitaires. Notre approche se focalise sur des types de liens dans le graphe qui sont susceptibles d'être pertinents par rapport à une situation, ce qui la rend particulièrement intéressante dans le cas de la publicité. En effet, les différences de contenus peuvent nécessiter des notions de proximité différentes selon la nature ciblée ou pas de leur public.

9.2.3. Prédiction de liens sociaux et recommandation de personnes à suivre

Dans le cadre d'une collaboration avec nos collègues anglais, nous avons commencé à explorer une piste d'utilisation de rapprochement de concepts dans le cadre de la prédiction des liens sociaux. Ce travail s'inspire, bien évidemment, de notre rapprochement de collaborateurs. Il vise cependant un public différent. Prédire les liens sociaux qui vont se créer au sein des réseaux en ligne, et surtout pouvoir suggérer des personnes à suivre dans un réseau comme Twitter, sont très important pour faciliter aux utilisateurs le maintien de leur connectivité et l'expansion de leur réseau. Nous croyons qu'un rapprochement exploitant la richesse du Web Social Sémantique pourrait, dans ce cas aussi, avoir des effets positifs.

9.3. Problématiques émergentes dans le domaine de l'innovation sur le Web

Dans cette partie, nous présentons les problématiques, pertinentes pour l'innovation sur le Web, que nous n'avons pas traité dans cette thèse. Nous avons pu nous rendre compte de leur importance, soit dans le cadre de nos propres travaux, soit en observant la pratique d'Hypios, en analysant les expériences cumulées au cours de ces dernières années. Nous espérons que ces thématiques inspireront de futurs travaux.

9.3.1. Soutien aux consultants et métiers intermédiaires

Si l'idée des plates-formes de résolution de problèmes en ligne fut introduite pour accélérer les échanges dans le monde des différents acteurs d'innovation, il est devenu clair que certains acteurs, notamment les entreprises ayant des problèmes, ne sont pas toujours très autonomes dans leur utilisation des plates-formes. C'est pourquoi celles-ci proposent, soit elles-mêmes, soit par intermédiaire de partenaires, des offres d'accompagnement dans le processus d'identification et de formulation de problèmes. Les besoins d'accompagnement s'imposent également après l'acquisition des solutions afin d'assurer leur implémentation dans le processus industriel. Il est donc clair que l'émergence des plates-formes n'a pas complètement éliminé les intermédiaires, mais a

plutôt changé leur rôle en rendant leurs services complémentaires à la fonctionnalité des plates-formes.

Dans la démarche d'accompagnement, une mise en valeur des connaissances existantes sur le Web pour la formulation de problèmes, et leur identification au sein des pratiques des entreprises s'imposent. Il serait particulièrement intéressant de voir si des technologies du Web Social Sémantique peuvent apporter des améliorations aux pratiques actuelles, qui reposent principalement sur un travail manuel des consultants.

9.3.2. Veille dans le cadre de processus d'innovation

La veille concurrentielle et technologique est devenue très importante pour des entreprises qui souhaitent garder leur compétitivité dans le marché actuel très dynamique. Certaines entreprises ont même des départements entiers consacrés à cette activité. La veille est aussi un levier important de l'innovation et peut nourrir le processus d'innovation par l'apport de nouvelles idées. Si des logiciels de veille existent sous forme de crawleurs, l'utilisation de la richesse du Web Social Sémantique pour améliorer leurs performances dans le cadre de l'innovation reste un défi intéressant et à grand potentiel, notamment pour générer des solutions innovantes dans l'avenir.

9.4. Au-delà du Web Social Sémantique

Dans cette partie nous présentons quelques pistes de travaux futurs - les possibles évolutions du Web Social Sémantique qui pourraient le diriger vers un Web encore plus utile. Nos idées de pistes visent principalement à inspirer le débat autour d'un Web d'avenir et sont donc essentiellement provocatrices.

9.4.1. Vers un Web véritablement intelligent

Dans la version initiale du Web Sémantique présentée en 2001 [Berners-Lee et al., 2001], la richesse structurelle du Web Sémantique doit servir principalement aux agents intelligents et indépendants pour mener des opérations à la place des utilisateurs. Ces agents peuvent intégrer les données de plusieurs sources et effectuer divers raisonnements. Ces agents doivent le rôle de véritables assistants personnels qui peuvent, par exemple, gérer l'agenda, effectuer des achats quotidiens en fonction des préférences et besoins de l'utilisateur sans avoir besoin de le consulter, etc. De tels agents pourraient aussi être capable de proposer des vacances de rêve pour l'utilisateur, en lui demandant uniquement de spécifier des dates. Ils seraient donc capables de choisir l'endroit et le type d'activités qui conviennent le mieux à l'utilisateur, et juger du meilleur choix par rapport aux dates spécifiées, au budget et au goût de l'utilisateur. Des données sémantiques de plusieurs sources les aideront dans cette tâche difficile.

A l'exception de quelques assistants virtuels très puissants comme Siri[162] – l'assistant fourni avec les téléphones portables de la marque Apple, nous n'avons pas encore observé l'émergence de tels assistants. Même Siri n'exploite pas encore la richesse du graphe Web et montre une indépendance assez limitée. Les progrès en terme d'inférence à partir de données du graphe Web reste donc toujours un défi à relever.

[162] http://www.apple.com/iphone/features/siri.html

Si l'inférence de nouveaux faits à partir de données sémantiques peut sans doute être utile pour le processus de rapprochement, il reste à explorer comment exactement les rapprochements peuvent rendre des assistants personnels plus utiles et plus indépendants.

9.4.2. Un Web qui « comprend » les émotions et les traits de personnalité

Si le Web d'aujourd'hui donne l'impression d'être dynamique grâce aux applications dites « en temps réel », son dynamisme est purement informationnel. Et pourtant, le Web a souvent en face de lui une personne réelle, qui a des émotions et qui réagit grâce à ses émotions et à son intelligence. En comparaison avec l'aspect émotionnel de l'interaction, le Web reste figé, froid.

Afin de permettre aux applications Web de devenir partenaires avec l'utilisateur et travailler en collaboration avec lui, le Web devrait s'acquitter d'un aspect émotionnel. Il doit être capable de prendre conscience non seulement du contenu informationnel des échanges des utilisateurs mais aussi de leur contenu affectif.

Si la plupart des applications Web tiennent compte des préférences des utilisateurs, rares sont celles capables de deviner le type de personnalité de l'utilisateur. Et pourtant la personnalité influence nos décisions, nos choix, et la nature de l'interaction que nous menons avec le Web. Il paraît donc important d'étudier également cet aspect et concevoir des applications qui s'adaptent mieux à l'utilisateur, à sa personnalité, à son contexte.

Les processus de rapprochement pourraient sans doute bénéficier d'une nouvelle richesse du Web de telle sorte. Celle-ci pourrait notamment aider à mieux caractériser la situation et le contexte affectif de l'utilisateur et y adapter le choix des rapprochements à faire pour servir l'utilisateur au mieux.

9.4.3. Vers un Web de subconscience

Si les différentes phases d'évolution du Web pourraient se comparer à des mouvements artistiques, le Web Sémantique correspondrait sans doute au Cubisme. Le cubisme essaye de représenter dans un plan visuel à deux dimensions la connaissance d'un objet qui dépasse ce qui serait visible sur un plan à une dimension. Cela entraîne des formes étranges, mais riches en connaissance sur l'objet. Le Web Sémantique enrichit le Web de documents d'une manière similaire par la couche de données sémantiques.

La question qui s'impose suite à une telle comparaison est la suivante : pouvons nous imaginer les futures phases du développement du Web comparables aux autres mouvements artistiques? Partons par exemple du surréalisme. Ce mouvement joue avec la subconscience, les rêves, et communique avec le consommateur de l'art en utilisant un registre au-delà du conscient. Pourrions nous interagir avec le Web au niveau subconscient et échanger plus d'informations? L'émergence des interfaces cerveau-machine, déjà utilisées dans les jeux vidéo apporte certainement l'espoir irraisonnable qu'un jour l'ordinateur pourrait lire dans les pensées de l'utilisateur et s'en servir pour communiquer.

* * *

Finalement, il nous semble important de dire un mot sur la réalisation de la thèse en tant que processus intellectuel. Elle a apporté une expérience et des connaissances considérables pour le doctorant. Nous serions contents si, grâce à cette thèse, nous avons réussi, même modestement, à faire avancer la science et à contribuer au corpus de la connaissance humaine. Mais ce qui est certain, c'est que la manière de penser du doctorant a beaucoup évolué. Depuis le début des recherches effectuées pour l'écriture de la thèse, ses visions du monde, de multiples facettes de la réalité, et de la limite du possible ne seront plus jamais les mêmes.

Bibliographie

[Berners-Lee et al., 2001] Berners Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*.

Annexe

Annexe 1 : Exemple d'appel aux solutions en RDF

Dans la Figure 80, nous présentons un exemple de problème d'innovation i.e. d'appel à solutions qui lui est associé en format RDF décrit en utilisant PCO. En plus de la structure des différents éléments, instances des classes précédemment décrites, cet exemple présente certaines propriétés permettant de localiser une page Web consacrée à l'appel à solutions qui l'identifie de manière non ambigüe (pb :challengePage), et de donner une description textuelle du problème à résoudre (pb :shortChallengeDescription). Les concepts pertinents, concepts du problème, sont déclarés en utilisant la propriété dct :subject.

```
@prefix dct: <http://purl.org/dc/terms/> .
@prefix gr: <http://purl.org/goodrelations/v1#> .
@prefix pb: <http://www.hypios.com/ontologies/problem/ns#> .

<http://data.hypios.com/problems/104.rdf#problem> dct:subject
      <http://dbpedia.org/resource/Fat>,
      <http://dbpedia.org/resource/Food>,
      <http://dbpedia.org/resource/Saturated_fat>,
      <http://dbpedia.org/resource/Trans_fat>,
      <http://dbpedia.org/resource/Triglyceride> ;
   pb:challengePage <http://www.hypios.com/problems/view/id/104> ;
   pb:currentStatus <http://research.hypios.com/files/problems/104.html#status> ;
   pb:hasReward <http://research.hypios.com/files/problems/104.html#price> ;
pb:publishedOn <http://data.hypios.org/hypios#me> ;
   pb:shortChallengeDescription "We are looking for alternatives to triglycerides to structure oil in food emulsions, without relying
on ingredients that structure the water phase in case of the emulsion."@en .
<http://research.hypios.com/files/problems/104.html#price> pb:Reward
      [ gr:hasCurrency "USD"@en ;
        gr:hasCurrencyValue "15000"@en ;
        pb:transactionType <https://www.hypios.com/ontologies/problem/ns#IPTransfer>
      ] .
<http://research.hypios.com/files/problems/104.html#status> pb:SelectionPhase
      [ pb:selectionDeadline "2011-03-01T00:00:00+00:00"@en
      ] .
```

Figure 80 Exemple de problème en format RDF. L'extrait de code est obtenu à partir d'un problème réel sur le site hypios.com, en traitant sa page Web par le distillateur RDFa[163] pour en extraire du code RDF.

Le site http://data.hypios.com/ donne l'accès aux descriptions de problèmes d'Hypios dans le format RDF à travers un point d'accès SPARQL

Annexe 2 : Hypothèses de proximité sociale

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix db: <http://www.dbpedia.org/ressource>.
@prefix rho: <http://hypios.com/ontologies/rho#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.

<http://example.org/hypothesis3> rdfs:label "The Distributional Social
Proximity Hypothesis";
                                  rdf:type   rho:UserRelatednessHypothesis,
rho:DistributionalHypothesis, rho:SimpleHypothesis ;
```

[163] http://www.w3.org/2007/08/pyRdfa/

Figure 81 Un exemple d'hypothèse de proximité sociale représenté par EHO

Les hypothèses de pertinence utilisées pour la recommandation des co-solveurs dans le chapitre 7, malgré leur nature essentiellement sociale, représentent un cas d'hypothèse composée, intersection de plusieurs hypothèses simples. Nous présentons un exemple d'hypothèse de ce type à la Figure 82.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix db: <http://www.dbpedia.org/ressource>.
@prefix rho: <http://hypios.com/ontologies/rho#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.


<http://example.org/hypothesis4>      rdfs:label      "Co-SolverProximity
Hypothesis";
                                      rdf:type   rho:CompositeHypothesis,
rho:UserRelatednessHypothesis;
                                      rho:consistsOf [
a rho:HeterogenousHypothesis, rho:DistributionalHypothesis;
rho:typeOfInitialEntities pco:ProblemChallenge ;
rho:typeOfRelevantEntities               foaf:Person.],               [a
rho:UserRelatednessHypothesis,
rho:DistributionalHypothesis.],<http://example.org/hypothesis3>.
```

Figure 82 Un exemple d'hypothèse d'expertise représentée par EHO

Annexe 3 : Hypothèses de proximité des traces

Sur la Figure 83 nous présentons la description RDF d'une hypothèse de proximité des traces, faite à l'aide de l'ontologie RHO. Cette hypothèse est utilisée dans les rapprochements entre les tweets et les articles que nous avons présentés dans le chapitre 6. Dans cet exemple nous introduisons également la propriété rho:textualStatement qui sert à décrire, d'une manière compréhensible pour les utilisateurs, la nature de l'hypothèse. Une telle description textuelle peut être utilisée dans des interfaces visuelles pour permettre à un utilisateur de mieux choisir une hypothèse.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix db: <http://www.dbpedia.org/ressource>.
@prefix rho: <http://hypios.com/ontologies/rho#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.


<http://example.org/hypothesis5> rdfs:label "The Hypothesis of Tweet-
Paper proximity";
                                 rho:textualStatement "If a tweet and a
scientific paper can be mapped to one another using classification
techiques relying on their topics and other features, then they can be
consider related";
                                      rdf:type
rho:TraceRelatednessHypothesis,              rho:DistributionalHypothesis,
rho:SimpleHypothesis.
```

Figure 83 Un exemple d'hypothèse de la proximité des traces représentée par RHO

Annexe 4 : Proximité collaborateur-collaborateur

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix db: <http://www.dbpedia.org/ressource>.
@prefix rho: <http://hypios.com/ontologies/rho#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.

<http://www.example.org/proximities> rho:forUser
<http://www.example.org/Milan_Stankovic>;
                                        rho:forProblem
<http://data.hypios.com/problems/21.rdf#problem>;

                                                rho:proximities [
a rdf:Seq;
rdf:_1 [a rho:WeightedResource; rho:resource <http:example.org/Matthew-
Rowe>; rho:weight 0.567.];
rdf:_2          [a          rho:WeightedResource;          rho:resource
<http:example.org/Fabien_Gandon>; rho:weight 0.567.];
rdf:_3          [a          rho:WeightedResource;          rho:resource
<http:example.org/Philippe_Laublet>; rho:weight 0.567.].];
                                                rho:hypothesisUsed
<http://example.org/hypothesis4>;
                                                rho:systemUSed
<http://hypios.com/hyProximity>.
```

Figure 84 Un exemple d'hypothèse d'expertise représentée par EHO

Dans le cas d'utilisation d'une hypothèse composite, comme dans l'exemple présenté à la Figure 84, il est possible de distinguer plusieurs types d'entrées, tels que l'utilisateur (par la propriété rho:forUser) et le problème (rho:forProblem) par rapport auxquelles les collaborateurs sont sollicités.

Résumé 1700 caractères maximum

Cette thèse s'inscrit dans le cadre de travaux relatifs au Web Social Sémantique, dans la perspective de la complémentarité et de la coévolution de deux aspects du Web, l'aspect social et sémantique. Le développement du Web au cours de ces dernières années a fait émerger un énorme graphe de données structurées, sémantiques résultant en partie de l'activité des utilisateurs, le LOD. Nous nous intéressons à l'utilisation de ce graphe afin de faciliter l'accès à l'information présente sur le Web, et ce de manière utile, informative et enrichissante pour l'utilisateur. Cette problématique est notamment étudiée dans les scénarios de l'innovation sur le Web – pratiques visant à utiliser des technologies du Web pour contribuer à l'émergence de l'innovation. Une spécificité de ce contexte, assez peu abordé dans la littérature existante, est sans doute le besoin d'inciter les découvertes inattendues et fortuites. Au delà de la simple pertinence sollicitée dans toute situation de recherche et de recommandation sur le Web, le contexte d'innovation impose une certaine ouverture d'esprit pour permettre à l'utilisateur d'accéder aux informations inattendues mais néanmoins pertinentes, et permet par la même occasion de s'inspirer et de transposer des idées d'un domaine à l'autre. Les travaux présentés dans cette thèse ont donc pour objectif d'aider, de manière directe ou indirecte, les acteurs de l'innovation en ligne (e.g., les entreprises qui cherchent à innover, les experts et les porteurs d'idées) de faire des découvertes. Cet objectif ce décline particulièrement par les travaux de construction d'un système de recherche d'experts, un système de recommandation de mots-clés pertinents pour un problème et un système de recommandation de collaborateurs pour aider à un expert d'affronter des problèmes pluridisciplinaires.

Web Sémantique, Web Social, Innovation sur le Web, Innovation Ouverte, Plate-forme de résolution de problèmes, Traces utilisateurs, Données Liées

Convergence between the Socaial and the Semantic Web

Application to Innovation on the Web

Résumé en anglais

This thesis builds upon the work on the Social Semantic Web, a research perspective on the complementarity and coevolution of two aspects of the Web, the social and semantic one. Web development in recent years has given rise to a huge graph of semantically structured data, partly resulting from user activity. We are particularly interested in the use of this graph in order to facilitate access to information found on the Web, in a useful, informative manner. This problem is particularly studied in scenarios related to innovation on the Web - practices to use Web technologies to contribute to the emergence of innovation. A notable specificity of this context, so far little discussed in literature, is the need to encourage serendipity and discovery. Beyond the simple relevance sought in any search and recommendation situation on the Web, the context of innovation requires a certain openness to allow the user to access information relevant yet unexpected, and should also open opportunities to learn and translate ideas from one domain to another.

The work presented in this thesis therefore aims to assist, directly or indirectly, the innovators online (eg, companies seeking to innovate, experts and carriers of ideas) to make discoveries. We address each of these challenges in different parts of the thesis. This vision is principally implemented through the construction of an expert search system, Hy.SemEx, a system for keyword recommendation allowing to discover unknown relevant keywords, HyProximity, and an approach for recommending collaborators to experts in order to help them face multidisciplinary problems.

Semantic Web, Social Web, Innovation via the Web, Open Innovation, Problem-sovling Platform, User Traces, Linked Data