République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université A.MIRA-BEJAIA

Faculté des sciences exactes
Département Informatique

THÈSE

Présentée par

Saida KICHOU

Pour l'obtention du grade de

DOCTEUR EN SCIENCES

Filière : Informatique

Option : Réseaux et systèmes distribués

Thème

Approche de Repérage et de Suivi de Compétences dans le Cadre du Web Communautaire

Soutenue le : 17/03/2022

Devant le Jury composé de :

Nom et Prénom	Grade		
Mr SIDER Abderrahmane	**MCA**	Univ. de Bejaia	Président
Mr BOUSSAID Omar	**Professeur**	Univ. de Lyon 2	Rapporteur
Mr MEZIANE Abdelkrim	**Directeur de Recherche**	CERIST	Co-Rapporteur
Mme BENBLIDIA Nadjia	**Professeur**	Univ. de Blida 1	Examinatrice
Mr BOUKHALFA Kamel	**Professeur**	Univ. de USTHB	Examinateur
Mme El BOUHISSI Houda	**MCA**	Univ. de Bejaia	Examinatrice

Année Universitaire : 2021/2022

À ma petite famille : mon cher mari Fouad et mes anges adorés Anis, Yacine, Sirine et Samy

À mes chers parents Abdelmadjid et Rahima

À ma chère grand-mère Adidi et ma tante Akila

À mes chers beaux-parents Said et Wardia

À mes chers frères Mohamed et Azedine

À ma chère sœur Samia, son mari et son ange Iline

À toute ma famille et belle-famille

Remerciements

Merci au bon Dieu le tout Puissant

Je tiens à exprimer mes remerciements les plus sincères et ma profonde gratitude à mon directeur de thèse Mr Omar BOUSSAID, professeur à l'université de Lyon2, de m'avoir donnée l'opportunité de réaliser mon travail de thèse sous sa supervision. Je le remercie de m'avoir encadrée, orientée et dirigée durant toute cette période.

Je tiens également à remercier chaleureusement mon co-directeur de thèse Mr Abdelkrim MEZIANE, directeur de recherche au CERIST pour ses conseils, orientations et compréhension durant tout ce parcours.

Je remercie Mr SIDER Abderrahmane, et Mme EL BOUHISSI Houda, MCA à l'université de Béjaia, Mme Nadjia BENBLIDIA, professeur à l'université de Blida et Mr Kamel BOUKHALFA, professeur à l'USTHB d'avoir accepté de juger cet humble travail.

Je remercie Mr BADACHE Nadjib ex-Directeur du CERIST pour l'opportunité qu'il m'a donnée ainsi que Mr BELBACHIR Hassan, l'actuel Directeur du Centre.

Je remercie particulièrement mon mari d'avoir été à mes côtés, aidée et soutenue tout au long de ce projet de thèse : je reconnais que sans toi je n'aurai pas pu tenir, que Dieu te grade pour moi et pour nos enfants.

Je remercie mes parents, mes enfants et toute ma famille pour leurs prières et encouragements, que Dieu vous garde tous pour moi.

Je remercie également tous mes collègues et amies du CERIST et de la DSISM pour leurs soutiens et encouragements. En particulier, Insaf que je remercie infiniment d'avoir été à mes côtés, sans oublier Mme Mellah pour ses conseils et orientations. Je remercie Houda, Amel, Fatma-Zohra, Lamia.K, Fatma, Zineb, Hadjer, Loubna, Lamia.H, Fouzia, Sahar, Abdeslam, Noureddine et Hocine.

Merci à toute personne ayant contribué de près ou de loin à l'aboutissement de ce modeste travail.

Résumé

Les organisations s'interrogent de plus en plus sur les avantages que peuvent leur procurer les modèles du Web collaboratif grand public (wiki, réseaux sociaux, crowdsourcing, etc.). Les réseaux sociaux sont un des moyens pour diffuser la connaissance et pour innover grâce à l'utilisation des informations qu'ils génèrent.

La problématique générale de notre travail s'inscrit dans le cadre de la recherche d'expertise, un domaine issu de la recherche d'information, qui se focalise sur l'estimation, la découverte, repérage et classement des expertises des utilisateurs. Les indicateurs de l'expertise, appelés souvent sources d'évidence, sont dans la majorité du temps liés aux documents rédigés par ces utilisateurs, leurs activités électroniques liées aux emails, propriétés de pages ou de sites web, etc. Vient s'ajouter ces dernières années leurs activités sociales : commentaires, postes, réponses aux questions, etc.

Dans notre contexte, nous nous intéressons plus particulièrement aux réseaux communautaires formés de professionnels amenés à partager de l'information et de la connaissance (sous diverses formes) autour de projets et/ou dans le cadre d'activités métiers. Ces professionnels sont connus par leurs profils et préférences mais aussi par leurs expertises et compétences, des notions complexes mais nécessaires à capitaliser. Ces notions difficiles à quantifier vu leur dynamicité, complexité et diversité des éléments pouvant apporter un plus à leurs estimations.

Dans ce travail, et pour répondre à notre problématique, nous avons d'abord mené une étude sur le domaine de la recherche d'expertise, nous avons mis en exergue les principaux travaux ayant apporté des plus dans ce domaine. Puis nous avons proposé un ensemble de contributions à l'estimation de l'expertise en se basant sur l'une des activités sociales largement utilisées par les utilisateurs, mais qui n'a pas été profondément exploitée à savoir l'activité du tagging social ainsi que les tweets.

Mots-clés

Compétence, expertise, recherche d'experts, profilage d'expert, topic, topic modeling, tagging social.

Abstract

Organizations are increasingly wondering about the advantages that collaborative Web models (wiki, social networks, crowdsourcing, etc.) can provide. Social networks are one of the means to disseminate knowledge and to innovate through the use of the information they generate.

The general problematic of our work falls within the expertise search, a field resulting from the information retrieval field, which focuses on the estimation, discovery, identification and classification of users' expertise. Expertise indicators, often called sources of evidence, are usually documents written by these users, their electronic activities like emails, page or website properties, etc. In recent years, their social activities have been added: comments, posts, answers to questions, etc.

In our context, we are particularly interested in community networks made up of professionals who share information and knowledge (in various forms) around projects and / or in the context of business activities. These professionals are known by their profiles and preferences but also by their expertise and skills, concepts that are complex but necessary to capitalize. These notions are difficult to quantify given their dynamicity, complexity and diversity of elements that can enhance their estimations.

In this work, and to resolve our problem, we first conducted a study on the field of expertise research; we highlighted the main work that has contributed more in this area. Then we proposed a set of contributions to the estimation of expertise based on one of the social activities widely used by users, but which has not been deeply exploited, namely the activity of social tagging and also tweets.

Keywords

Competency, expertise, expert finding, expert profiling, topic, topic modeling, social tagging.

Table des Matières

Introduction Générale 1
1. Contexte et Motivation 1
2. Problématique 3
3. Contributions 4
4. Organisation 5
5. Publications 7
5.1. Publication Internationale 7
5.2. Communications Internationales 7

Première Partie

Recherche d'Expertise : Rappels et État de l'Art

Introduction 9
Chapitre I: **Recherche et Profilage d'Experts** 11
1. Introduction 11
2. Définitions 12
2.1. Qu'est-ce qu'une compétence 12
2.2. Compétence et expertise, quelle différence ? 13
3. Recherche d'Expertise 14
3.1. Classification des Approches de la Recherche d'Expertise 15
3.2. Sélection des sources d'information 17
4. Repérage ou recherche d'experts 18
4.1. Approches basées sur la RI 19
4.2. Approches basées sur l'analyse des liens 22
5. Profilage de l'Expert 26
6. Tagging et Folksonomies : Avantages et Inconvénients 31
6.1. Définitions 31
6.2. Avantages et inconvénients 32
7. Conclusion 33
Chapitre II: **Le Topic Modeling dans la Recherche d'Expertise** 35
1. Introduction 35
2. Définitions 36
3. Les Techniques du Topic Modeling 38

3.1. Latent Semantic Indexating (LSI) 38
3.2. Probabilistic Latent Semantic Indexing (PLSI) 39
3.3. Latent Dirichlet Allocation (LDA) 40
3.3.1. Les Extensions de LDA 42
4.Principaux travaux sur la recherche d'expertise utilisant le topic modeling 46
5.Conclusion 49
Chapitre III: **Synthèse et Discussion des Travaux sur la Recherche d'Expertise** 50
1. Introduction 50
2. Synthèse des Travaux 51
3. Discussion 55
3.1. Travaux liés à l'expert finding 55
3.2. Travaux liés à l'expert profiling 56
3.3. Sources d'évidence considérées 56
4. Conclusion 58

Deuxième Partie

Contributions

Introduction 61
Chapitre IV: Un Modèle d'Estimation de l'Expertise basé sur le Tagging Social 63
1. Introduction 63
2. Motivations 64
3. Principe Général 65
4. Notation et Terminologie 67
5. Modélisation du Profil de l'Expert 69
5.1. Modèle du Profil Expert 69
5.2. Construction du Profil (Profiling) 70
5.2.1. Prétraitement 70
5.2.2. Construction de la dimension sociale 71
5.2.3. Construction de la dimension topicale (thématique) 74
6. Estimation de l'Expertise de l'Utilisateur 77
7. Recherche de l'Expert (Finding) 80
7.1. Présentation du Modèle Vectoriel 80
7.2. Calcul du Poids d'un Tag 81
7.3. La Fonction d'Appariement 83
8. Conclusion 84
Chapitre V: **Expérimentations et Résultats** 86

1. Introduction 86
2. Collection et Métriques 86
3. Modèles de Base (Baseline) 88
4. Définition des Paramètres 88
5. Résultats Expérimentaux et Analyse 89
5.1. Exp Vs LDA et le modèle de base BM 89
5.2. Exp Vs LDA et les Autres Modèles 92
6. Expérimentation sur le Dataset Delicious 96
7. Expérimentations sur un Dataset de twitter 99
7.1. Acquisition des Données 100
7.2. Prétraitement des données 101
7.3. Tests et Résultats 103
7.4. Modélisation Thématique 103
7.5. Évaluation et Discussion des Résultats 106
8. Conclusion 107
Chapitre VI: **Cas d'Etude : Application au project Algéro-Tunisien « Recommandation des Femmes Artisans »** 109
1. Introduction 109
2. Recommandation Basée sur le Tagging 111
3. Apport du Tagging dans le Cas des Femmes Artisans 112
4. Modélisation du Profil 114
4.1. Représentation du Profil de la Femme Artisan 114
4.1.1. Le Concept Intérêts-Social 116
4.1.2. Le Concept Réputation 116
4.1.3. Le Concept Expertise 117
4.1.4. Les Attributs 117
4.1.5. Les Relations 117
4.2. Représentation du Profil Utilisateur (Fournisseur, Client) 118
5. Approche de Repérage du Profil des Femmes Artisans 119
5.1. Acquisition des Centres d'Intérêt 120
5.2. Acquisition de la Réputation 121
5.3. Enrichissement de la Description du Produit 122
6. Recommandation basée sur les intérêts de l'utilisateur 123
6.1. Calcul de similarité 124
7.Expérimentation 125

7.1 Acquisition des données 125
7.2. Tests et résultats 126
7.3. Synthèse 128
8. Conclusion 129
Conclusion Générale 130
Références 133

Liste des Figures

Figure 1 : Les niveaux d'acquisition de compétences. 14
Figure 2 : Matrice de compétences, (Balog et al., 2012) 15
Figure 3: Classification des techniques de la recherche d'experts.....16
Figure 4 : Classification des techniques selon (Al-Taie et al., 2018). 17
Figure 5 : Exemple d'un document traité avec une technique du topic modeling, (Bietti, 2012) 37
Figure 6: Exemple d'une matrice document-terme. 38
Figure 7: Modèle graphique de LDA, (Blei, Ng, et al., 2003). 41
Figure 8: Une taxonomie des méthodes basées sur LDA et ses extensions, (Jelodar et al., 2019). 43
Figure 9: Principe général de l'approche......66
Figure 10: Exploitation des tags dans l'estimation de l'expertise. 67
Figure 11: Modèle du Profil de l'expert......70
Figure 12: Exemple de graphe construit avec l'approche hybride, (Kichou et al., 2011) 73
Figure 13: Description du processus de LDA 75
Figure 14: Processus de modélisation thématique avec LDA..... 76
Figure 15: Les résultats de LDA (Outputs)......77
Figure 16: Exemple de profondeurs du terme 'Autocar'...... 79
Figure 17 : Représentation des documents dans le modèle vectoriel 80
Figure 18 : Exp Vs BM et LDA (Map et Rprec) 89
Figure 19 : Exp Vs Modèles de base (Rappel-Précision interpolés) pour la requête Java et C# 91
Figure 20: Exp Vs Modèles de base (Rappel/Précision interpolés) pour la requête Android et pour toutes les requêtes 92
Figure 21: Comparaison avec les autres modèles (requêtes Java et C#) 94
Figure 22 : Comparaison avec les autres modèles (requête Android et moyenne ndcg) 95
Figure 23 : Exemple de catégorisation de tags en topics 96
Figure 24: Top-30 des tags pertinents composant le topic 1 97
Figure 25 : Top-30 des tags pertinents composant le topic 9 97
Figure 26: Variation des profondeurs des tags du topic 2 99
Figure 27: Processus de collecte de données Twitter.....100
Figure 28: Processus du prétraitement des tweets.....101
Figure 29: Analyse des sentiments dans les tweets 102

Figure 30: La tokenisation 103
Figure 31: La lemmatisation 104
Figure 32: La création du dictionnaire 104
Figure 33: Topics de l'utilisateur Computer Science 105
Figure 34: Topics de l'utilisateur Cuisine et mets 106
Figure 35: Le Tagging User-User 113
Figure 36: Le Tagging User-Product 114
Figure 37: Graphe Construit avec l'approche hybride 120
Figure 38: Exemple d'une Réputation d'un fournisseur 122
Figure 39: Accueil de la plateforme 126

Liste des Tableaux

Tableau 1: Récapitulatif des travaux existants 54
Tableau 2 : Exemple de tags, leurs popularités et les topics correspondants 82
Tableau 3 : Les expertises résultantes 82
Tableau 4 : Calcul des poids des tags 83
Tableau 5 : Calcul du RSV 84
Tableau 6: Les requêtes utilisées pour l'expérimentation. 87
Tableau 7: Résultats de Map et Rprec pour Exp, LDA et BM. 89
Tableau 8: Comparaison du modèle Exp avec les autres modèles de base (requête Java et Android) 93
Tableau 9: Comparaison du modèle Exp avec les autres modèles de base (requête C# et moyenne nDCG) 93
Tableau 10: Un exemple de topics et leurs fréquences attribués à un ensemble d'utilisateurs 98
Tableau 11: Comparaison des résultats 107
Tableau 12: Dictionnaire des concepts du profil de la femme artisan complété par les 4 nouveaux concepts 116
Tableau 13: Les attributs des nouveaux concepts 117
Tableau 14: Les relations binaires des trois concepts du profil de la femme artisan 118
Tableau 15: Les concepts du profil User 118
Tableau 16 : Les attributs des concepts du profil User 119
Tableau 17: Les relations binaires des concepts du profil User 119
Tableau 18: Extension de l'ontologie métier 122
Tableau 19: Attributs du concept Descripteur-Produit 122
Tableau 20: Les relations ajoutées pour l'ontologie Métier 123
Tableau 21: Résultats du calcul de similarité pour la recommandation des femmes artisans 127
Tableau 22: Résultats du calcul de similarité entre différents utilisateurs 128

Liste des abréviations

KSAO: Knowlege, Skills, Abilities and Other characteristics.
LDA: Latent Dirichlet Allocation.
TREC: Text REtrieval Conference.
HLDA: Hierarchical Latent Dirichlet Allocation
MG-LDA: Multi-Grain Latent Dirichlet Allocation.
LLDA : Labeled Latent Dirichlet Allocation
CQA: Community of question answer.
EBA: Entropy Based Approach.
HITS: Hyperlink-induced topic search.
JCDL: Joint Conference on Digital Libraries.
PSO: Particle Swarm Optimization.
L2R: Learning to Rank.
LSA: Latent Semantic Analysis.
PLSA: Probabilistic Latent Semantic Analysis.
CTM: Correlated Topic Model.
LTN: Latent Topic Network.
HDP: Hierarchical Dirichlet Process.
DTM: Dynamic Topic Modeling.
RTM: Relational Topic Model.
ATM: Author- Topic Model.
MM: Model of Mixture.
CNT: Co-author Network Topic.
STM: Segmented Topic Model.
PLM: Predictive Language Model.
PMI: pointwise mutual information.
MAP: Mean Average Precision.
NDCG: Normalized discount cumulated gain
LMS: LanguageModel Based Scoring.

Introduction Générale

Believe one who has proved it. Believe an expert.
—Virgil (70 BC–19 BC), Aeneid

1. Contexte et Motivation

Un expert est une personne capable de satisfaire certains besoins en information, donner des réponses correctes à des questions spécifiques, les expliquer et même guider l'utilisateur vers d'autres sources d'informations pertinentes. Nous faisons appel à des experts lorsque nous avons besoin de quelqu'un pour nous montrer le bon chemin pour s'attaquer à un problème. Il peut y avoir de gros volumes d'informations disponibles autour du problème à résoudre, l'expert est là pour nous aider à mieux les exploiter (Balog et al., 2012). De nos jours, pour trouver un expert, les entreprises et les organisations étendent leurs activités et investigations aux communautés du web. En effet, la profusion des réseaux sociaux (Facebook, Twitter, Google +, LinkedIn, etc.) a rendu plus facile que jamais les interactions entre utilisateurs. Elle a fait l'objet d'une forte attention de diverses disciplines de recherche qui tendent à étudier les utilisateurs du point de vue, attentes, préférences, compétences, etc.

De manière générale, « Un réseau social est un ensemble d'identités sociales telles que des individus ou des organisations sociales reliées entre elles par des liens créés lors des interactions sociales ». En informatique, les réseaux sociaux (RS) sont des plateformes qui offrent aux utilisateurs la possibilité d'interagir et de former des groupes (amis, famille, etc.), de collaborer, communiquer et créer des communautés autour des participants. Ces réseaux permettent la contribution et la collaboration des utilisateurs, la création de diverses communautés d'utilisateurs, et l'innovation en facilitant les contacts entre professionnels et/ou experts.

Avec l'émergence de tous ces médias sociaux, les utilisateurs sont impliqués dans la production de l'information sur le web. L'information pertinente est devenue de plus en plus précieuse dans le sens où ce n'est pas évident de la retrouver aussi facilement. Mais aussi, trouver l'information crédible issue de connaisseurs ou d'experts d'une thématique particulière. L'un des domaines qui a suscité une attention particulière des

chercheurs est le repérage d'experts ou recherche d'experts dit *expert finding*. Beaucoup de travaux ont été proposés pour retrouver l'expert en se basant sur plusieurs critères, dans un but d'améliorer l'accomplissement d'une tâche donnée, à la suite d'un besoin. Celui-ci peut être une question dans un domaine donné, un emploi dans une spécialité particulière, une tâche spécifique (*reviewing* par exemple).

Néanmoins, la problématique de la recherche d'expertise est loin d'être résolue, et plusieurs pistes liées notamment aux sources indicatrices de celle-ci, enrichies par la prolifération des activités sociales, sont peu explorées.

L'idée de ce travail naquit dans le cadre d'un projet de recherche et de développement intitulé : « *Towards a new Manner to use Affordable Technologies and Social Networks to Improve Business for Women in Emerging Countries* ». Mené entre l'Algérie représentée par le CERIST (CEntre de Recherche sur l'Information Scientifique et Technique) et la Tunisie représentée par les laboratoires SOIE (Stratégie d'Optimisation et Informatique Intelligente) et MIRACL (Multimedia, InfoRmation Systems and Advanced Computing Laboratory). Les principaux objectifs du projet sont de promouvoir le commerce des femmes artisans de ces deux pays, en améliorant leurs processus de production, leurs approvisionnements, la commercialisation de leurs produits et promouvoir aussi l'échange entre elles. Proposer également une utilisation des nouvelles technologies adaptée à leurs profils (capacités cognitives et physiques). Il a été aussi question de repérer une compétence dans un métier particulier et la recommander à une personne susceptible d'être intéressée par cette compétence.

Les travaux de cette thèse s'inscrivent dans le domaine de la recherche d'expertise (*expertise search*) connu également par repérage d'expertise (*expertise locating*) ou identification d'expertise (*expertise identification*). Après plusieurs années de travail dans ce domaine, deux tâches principales liées à la recherche d'expertise se sont distinguées : la recherche d'experts (*expert finding*) s'intéressant à la question : Quels sont les experts dans le Topic X ? et le profilage d'expert (*expert profiling*) qui s'intéresse à la question : Quels sont les topics d'expertise d'une personne Y ? Ces deux tâches sont considérées comme les deux faces d'une même monnaie, elles ont toutes les deux le même objectif : Repérer la compétence.

2. Problématique

L'expertise et la compétence sont deux termes utilisés pour décrire des professionnels qualifiés. Que ce soit pour les organisations ou les particuliers, retrouver une compétence souhaitée reste un défi. En effet, ceci est dû à plusieurs raisons : (1) la notion de compétence est complexe, difficile à quantifier et même à définir et formaliser. L'expertise peut être définie comme la compétence ou la connaissance qu'une personne a (ou acquiert) dans un domaine particulier, (Carchiolo et al., 2015). Elle a été aussi définie comme étant : « les connaissances, les compétences, les capacités et autres caractéristiques (KSAOs, Knowlege, Skills, Abilities and Other characteristics » nécessaires à une performance efficace dans les emplois en question », (Campion et al., 2011). (2) la diversité des sources d'évidence indicatrices de la compétence ou de l'expertise. Certains se fient aux déclarations des utilisateurs, d'autres aux différentes recommandations qui leurs sont associées. Alors que d'autres ne croient pas à cela et observent le comportement de ces utilisateurs et leurs productions (documents, emails, postes, etc.). (3) La compétence est dynamique, des changements peuvent en survenir, en effet « la compétence est dynamique. Elle reconstruit de manière dynamique les différents éléments qui la constituent (savoirs, savoir-faire pratiques, raisonnement*) »* (Parlier, 1994).

La principale question posée est : Comment localiser cette compétence ? Le domaine de la gestion des connaissances s'est développé dans le but d'utiliser les connaissances d'une organisation aussi bien que possible. L'un des premiers objectifs fut de développer des systèmes d'information susceptibles d'appuyer la recherche d'expertise. Les approches initiales étaient principalement axées sur la façon d'unifier les bases de données disparates et dissemblables de l'organisation en un seul entrepôt de données qui pourrait être facilement exploité (Balog et al., 2012; Yimam & Kobsa, 2000). Les premiers outils reposaient sur l'auto-évaluation par les personnes de leurs compétences par rapport à un ensemble prédéfini de mots-clés.

Malgré les progrès accomplis jusqu'à présent, la question de savoir comment fournir un accès efficace à l'expertise continue d'être abordée de différents points de vue. Les informations utilisées comme indicateurs d'expertise d'individus ont été, pour une longue période et d'une grande majorité de travaux, limitées aux documents rédigés (production scientifique, rapports techniques, etc.) par ces individus et leurs

emails. Ce type d'information peut ne pas être disponible et généralement concerne une catégorie d'individus (chercheurs et académiques). Par ailleurs, la profusion du web social et communautaire permettant aux utilisateurs de s'organiser en communautés, de partager leurs savoirs, et s'exprimer librement sur le web n'a fait que favoriser la production de données utiles pouvant envelopper beaucoup d'informations sur l'utilisateur.

Nous nous intéressons dans ce travail en particulier à l'utilisation des sources d'informations peu explorées à savoir les activités sociales liées au Tagging. En effet, le tagging social est l'un des moyens les plus populaires permettant aux utilisateurs de produire de l'information sur le web. Ces informations peuvent être un indicateur de leurs intérêts et même de leurs expertises et compétences.

3. Contributions

Dans ce travail, nous avons d'abord effectué une étude sur les différents travaux liés au domaine de la recherche d'expertise. Nous avons exploré les sous-domaines qui lui sont liés et présenté les notions fondamentales de compétence et ses relations avec la notion d'expertise. Sans oublier de parler de la compétence en entreprise et sur le web. Par la suite, nous avons présenté un état de l'art lié aux travaux les plus importants du domaine. La dernière notion que nous avons évoquée est celle de la modélisation de sujets ou thématiques (*topic modeling*) vu sa forte corrélation avec le domaine de la recherche d'expertise car celle-ci est liée à un topic (thématique, sujet) particulier. Un ensemble de travaux exploitant le *topic modeling* pour la recherche d'experts a été présenté. La notion du tagging social a aussi été évoquée. Nous avons par la suite récapitulé et discuté les différents travaux.

En plus de l'étude bibliographique et l'analyse des travaux de l'état de l'art, notre contribution peut être résumée dans les points suivants :

- Proposition d'un nouveau modèle d'estimation de l'expertise en se basant sur les activités du candidat liées au tagging social où nous avons proposé un modèle de l'expert comportant deux principales dimensions. Le modèle d'estimation de l'expertise utilisateur se base sur les profondeurs des tags utilisés en exploitant une technique du topic modeling (Latent Dirichlet

Allocation (LDA)) pour déduire les topics d'expertise des utilisateurs avec le niveau de chaque expertise ;

- Proposition d'un modèle de recherche d'expert permettant l'évaluation du modèle d'estimation de l'expertise ;
- Étude du cas de Twitter en vue de la création d'un profil thématique de l'utilisateur en expérimentation ;
- Intégration des activités du tagging social dans le cas des femmes artisans. En effet, cette contribution vise à faire bénéficier les femmes artisans des avantages du tagging social, pour une meilleure visibilité de leurs profils et leurs productions. Deux types de tagging ont été proposés : le tagging *user-product* et le tagging *user-user* ;
- Proposition d'une approche de recommandation basée sur le profil utilisateur pour le cas des femmes artisans. Nous avons proposé d'une manière triviale, des recommandations de femmes artisans pour des clients susceptibles d'être intéressés par les productions de celles-ci. Des recommandations de fournisseurs pour les femmes artisans en prenant en considération leurs réputations basées sur les activités du tagging de ces derniers.
- Les contributions ont été évaluées en menant des expérimentations sur plusieurs collections.

4. Organisation

Ce mémoire de thèse est composé de cette présente introduction générale, de deux parties principales et d'une conclusion générale dans laquelle nous présentons les principales synthèses ainsi que les perspectives de nos travaux.

La première partie intitulée : « Recherche d'Expertise, Rappels et état de l'art » est consacrée à la présentation des concepts de base de la recherche d'expertise ainsi qu'une présentation détaillée des travaux relatifs à ce domaine. Elle est organisée en trois chapitres :

Dans **le premier chapitre** nous introduisons les principales notions liées à la compétence, ses définitions, et son importance en entreprise et sur le web. Par la suite, nous étudions de plus près le domaine de la recherche d'expertise, en présentant ses

deux principales tâches : la recherche d'experts et le profilage d'experts. Les travaux les plus importants de chaque tâche sont aussi présentés, suivis d'une conclusion.

Le deuxième chapitre est consacré à la modélisation thématique (topic modeling). Cette notion a un apport important dans le domaine de la recherche d'expertise. Elle est très liée à l'expertise car celle-ci concerne une thématique particulière. Nous avons donc présenté les différentes techniques citées et adoptées dans la littérature. Un ensemble de travaux sur la recherche d'expertise ayant recours au topic modeling a été également présenté.

Le troisième chapitre est consacré aux synthèses et discussions sur l'ensemble des travaux présentés dans les deux chapitres précédents.

La deuxième partie intitulée : « Contributions » est consacrée à la présentation de nos contributions pour ce domaine de la recherche d'expertise. Cette partie est organisée en trois chapitres.

Dans **le quatrième chapitre** nous présentons notre contribution fondamentale exploitant les relations sémantiques entre les tags ainsi que leurs profondeurs pour l'estimation de l'expertise utilisateur. La contribution consiste également en une proposition d'un modèle de l'expert et un modèle de recherche.

Le cinquième chapitre est consacré aux différentes expérimentations et études de cas pour l'évaluation et la validation de nos propositions.

Dans **le sixième chapitre** nous présentons une adaptation de notre modèle au cas du projet de coopération Algéro-Tunisien, où nous avons proposé deux types de tagging social visant une meilleure visibilité de ces femmes artisans et leurs produits. Le tagging *user-product* permet aux femmes artisans et aux différents utilisateurs d'annoter (tagguer) les produits. Le tagging *user-user* permet aux utilisateurs de se tagguer mutuellement. Une approche de recommandation des femmes artisans et des fournisseurs est également présentée.

5. Publications

5.1. Publication Internationale

- Kichou, S., Boussaid, O., & Meziane, A. (2020). Tag's Depth-Based Expert Profiling Using a Topic Modeling Technique, in *International Journal on Semantic Web and Information Systems (IJSWIS*), 16(4), 81-99. doi:10.4018/IJSWIS.2020100105

5.2. Communications Internationales

- Kichou, S., Meziane, A. (2015): User Profile Extraction Based on Social Tagging. Case Study: Handicrafts Women in Emerging Countries, *Conférence sur les Avancées des Systèmes Décisionnels, ASD'2015*. Tanger, Maroc, Septembre 2015.
- Kichou, S., Mellah, H., Boussaid, O., & Meziane, A. (2015): Recommendation Based on User's Interest. Case Study: Handicrafts women in Emerging countries, in *4th International Conference on Software Engineering and New Technologies, ICSENT*. Istanbul, Turkey, December 2015.
- Kichou, S., Mellah, H., Boussaid, O., & Meziane, A. (2016): Handicraft women Recommendation Approach based on User's Social Tagging Operations, *in 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)* (pp. 618-621). Omaha, USA, December 2016.

PREMIÈRE PARTIE

RECHERCHE D'EXPERTISE : RAPPELS ET ÉTAT DE L'ART

Introduction

Nous sommes tous passés, un jour ou l'autre par un entretien d'embauche, ce moment difficile où nous faisons face à cet étranger qui nous rafale de questions. Nous tenons, malgré notre stress, à montrer nos capacités, qualités et compétences. Ces compétences que nous avons acquises au fil du temps, de nos études, nos stages, nos différentes activités de la vie, etc. Les temps ont changé, avec les nouvelles technologies, il suffit de mettre votre curriculum vitae (CV) dans un site spécialisé pour que vous receviez les offres d'emploi relatives à votre profil.

En effet, l'entreprise peut avoir recours au web pour une éventuelle recherche de compétences, notamment avec l'explosion de celui-ci, ainsi que des réseaux sociaux qui représentent une masse de données non négligeable, pouvant être d'un apport important aux entreprises. Nous citons à titre d'exemple, *LinkedIn*[1], un site offrant la possibilité aux différents utilisateurs de présenter leurs expertises, expériences, etc., et permettre aux organisations de contacter les personnes dont le profil convient aux attentes de celles-ci. Les Communautés de Question Réponse, communément connues par CQA, permettent aux utilisateurs de poser leurs questions et attendre les réponses. Une réponse d'un expert serait certainement d'une grande crédibilité et utilité. Ceci est important pour les utilisateurs, mais aussi pour les organisations qui peuvent repérer un expert à travers ses réponses.

Vous pouvez donc avoir une activité assez riche sur le web, où vous participez peut-être à répondre à des questions, résoudre des problèmes dans un domaine donné, annoter des ressource, etc. et vous serez repéré.e en tant qu'une personne ayant des compétences ou expertise particulières.

Ces notions de compétences et d'expertise ne sont pas aussi simples qu'elles en ont l'air. Dans cette première partie, nous explorerons le domaine de la recherche d'expertise. Nous évoquerons en premier lieu, les notions de compétence et d'expertise. Par la suite, nous présenterons de plus près ce qui se fait dans la recherche d'experts, nous parlerons de l'apport de l'aspect social pour ce domaine. En dernier lieu, nous

[1] https://www.linkedin.com/

présenterons le domaine de la modélisation de sujets (*topic modeling*) fortement lié au domaine de la recherche d'expertise.

Chapitre I
Recherche et Profilage d'Experts

1. Introduction

La recherche d'information (RI) qui est une discipline visant à retourner la bonne information suite à un besoin utilisateur, a fait l'objet de plusieurs progrès ces derniers temps. Le but étant d'aider l'utilisateur à retrouver l'information voulue, de la manière la plus simple, flexible et adéquate possible. Les gens recherchent non seulement des documents, mais des personnes pouvant avoir des connaissances sur des thématiques qui les intéressent, (Hertzum & Pejtersen, 2000). À cet effet, une branche de la RI, a vu le jour il y a maintenant plusieurs années, à savoir la recherche d'expertise (*expertise search*). Celle-ci est primordiale pour repérer la bonne personne pouvant résoudre un problème donné, effectuer une tâche dans une spécialité donnée ou répondre à une question pointue dans un domaine particulier.

Notre intérêt se porte sur le web communautaire, qui constitue de nos jours une source importante d'information pour les organisations, les institutions et les individus. Avec l'explosion des réseaux sociaux, le besoin de moyens de recherche d'information raffinés est de plus en plus important.

Ces dernières années, de nombreuses études ont été réalisées pour définir la recherche d'expertise en tant que tâche supérieure de la RI, (Nobari et al., 2020). En effet, elle a fait l'objet de plusieurs études. (Balog & De Rijke, 2007) avait pour la première fois introduit la manière de réaliser la tâche du profilage de l'expert, faisant de la recherche d'expertise un ensemble de deux tâches principales : la recherche, découverte ou repérage d'experts et profilage de l'expert.

Dans ce chapitre, nous allons présenter quelques définitions de la notion de compétence, la notion d'expertise. Nous allons par la suite explorer le domaine de la recherche d'expertise ainsi que toutes les notions qui lui sont liées. Nous présenterons également les travaux de recherche réalisés dans la recherche de l'expertise. Avant cela, nous expliquerons les classifications de ceux-ci selon les techniques utilisées et les

algorithmes adoptés. Nous jugeons aussi important d'évoquer les notions de Tagging et de folksonomies. Nous terminerons par une conclusion.

2. Définitions

Plus que jamais, les entreprises évoluent dans un environnement économique où chacun cherche à renforcer et à trouver de nouveaux éléments de compétitivité. De nouveaux facteurs comme l'optimisation des compétences et l'organisation humaine viennent compléter la compétitivité des entreprises tout en reconfigurant peu à peu, mais sûrement, les liens avec les clients, partenaires et fournisseurs. Dans cette section, nous donnons les définitions principales existantes dans la littérature sur la compétence et l'expertise.

2.1. Qu'est-ce qu'une compétence

D'après (Roos, 2006) le concept de compétence, s'il est largement répandu, est loin de faire l'objet d'une définition consensuelle et partagée de la part des différents auteurs qui travaillent sur ce sujet, que ce soit les sociologues comme les chercheurs en sciences de gestion. En effet, « la question de la compétence fait enjeu ».

La polysémie de ce terme s'explique en partie par sa généalogie, puisqu'il trouve son origine en sociologie, avant d'être repris et enrichi par des gestionnaires, cogniticiens, etc.

Il existe un grand nombre de définitions totalement différentes, ce qui montre la difficulté d'établir un consensus autour de ce sujet. Toutefois, un élément essentiel semble être plus ou moins reconnu de tous et dissipe légèrement le flou qui entoure cette notion, à savoir que la compétence est éminemment contextuelle. En effet, dans (MOREAU, 2010) les auteurs affirment que : « tant qu'elle n'est pas passée à l'épreuve de situation effective, la compétence reste virtuelle, potentielle ou en devenir ». La compétence a un caractère socialement construit et ne prend sens que par rapport à une situation de travail.

Une autre définition donnée par (Paddeu, 1999): « la compétence est l'ensemble de connaissances, de savoir-faire et de comportements structurés en fonction d'un but dans un type donné de situation de travail ».

(Le Moigne, 2007) dit que « la compétence est un concept abstrait nécessairement dénué de toute réalité tangible. La compétence n'existe que par les représentations que nous en construisons : la carte n'est sans doute pas le territoire, mais le territoire compétence n'a de réalité que par les cartes, ou les modèles, que nous en établissons ».

La compétence individuelle est « la capacité d'une personne à réaliser une tâche donnée » (Jukic & Huljenic, 2007), ceci implique l'existence d'un prérequis physique et mental ainsi que la connaissance nécessaire pour l'accomplissement de la tâche.

2.2. Compétence et expertise, quelle différence ?

La plupart des définitions de l'expertise et de la compétence dans la littérature se réfèrent au terme habilité (Fazel-Zarandi, 2013). Une compétence est définie comme : une capacité qui a été acquise par l'entraînement ou la formation et une capacité à effectuer des tâches mentales ou physiques avec un résultat déterminé, (Marrelli, 1998). En d'autres termes, une habilité est une capacité acquise qui permet d'effectuer certaines activités.

L'expertise Selon (Roqueplo, 1997) est « L'expression d'une connaissance formulée en réponse à une demande de ceux qui ont une décision à prendre, en sachant que cette réponse est destinée à être intégrée à un processus de décision ».

(Mcdonald & Ackerman, 1998) considèrent l'expertise comme la concrétisation des connaissances et des compétences au sein des individus. (Sarahelen Thompson, Mark L Waller, 1988) définissent l'expertise comme la possession d'un vaste corpus de connaissances et de compétences procédurales. D'après (Ferracci, 2012) l'expertise peut se résumer à un axiome : savoir + expérience.

Quant à (Swanson, 2007), pour lui l'expertise est le niveau optimal auquel une personne est capable et/ou attendue de performer dans un domaine spécialisé de l'activité humaine.

Bien que ces définitions reflètent la compréhension commune de l'expertise, elles ne sont pas très utiles pour distinguer un expert d'un non-expert car les non-experts peuvent également avoir une bonne quantité de connaissances, de nombreuses années d'expérience.

Il existe cinq niveaux d'acquisition de compétence (*skill acquisition*) reconnus, et connus sous le nom du « *Model of Skill Acquisition* » illustré dans la *Figure 1* ci-dessous.

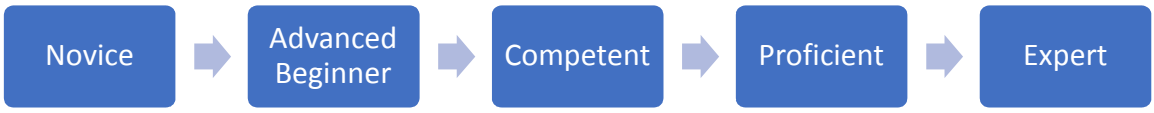


Figure 1 : Les niveaux d'acquisition de compétences, (Fazel-Zarandi, 2013).

Les notions de *compétence* et d'*expertise* sont tellement fortement liées et chevauchantes, que l'évocation de l'une ou de l'autre devient identique. En outre, les termes expertise et expert sont plus utilisés dans ce domaine. Nous allons donc, dans la suite de notre travail, parler plus d'expertise et d'expert pour désigner une compétence.

3. Recherche d'expertise

La recherche d'expertise est une branche de la RI, son objectif est de relier l'humain aux domaines d'expertise et vice versa. Elle a été principalement étudiée dans le domaine de la gestion des connaissances où l'objectif est d'utiliser les connaissances humaines au sein d'une organisation aussi bien que possible, (Balog et al., 2012). À ses débuts, la recherche d'expertise était liée pour longtemps à une organisation bien spécifique. Composée de deux principales tâches : le repérage ou recherche d'experts, et le profilage d'expert, (Rybak et al., 2014) . La première tâche consiste à répondre à la question : « quels sont les experts dans la thématique ou le topic X ». Quant à la deuxième répond à la question : « Quels sont les thématiques ou topics d'expertise d'un candidat Y ». Les systèmes de recherches d'experts utilisent des données fournies explicitement ou implicitement, pouvant informer sur l'expertise de l'utilisateur dans le but d'identifier les experts appropriés. Les données explicites sont généralement celles fournies par l'utilisateur lui-même, celles-ci peuvent être erronées, insuffisantes ou non-précises. En d'autres termes, les gens peuvent ne pas être conscients d'avoir une certaine compétence à un niveau de maitrise donné, ou ils peuvent mentir sur leurs descriptions de ce qu'ils ont contribué ou accompli, (Fazel-Zarandi & Fox, 2011). Le profilage d'expert est basé sur les mêmes principes de création d'associations entre personnes et topics d'expertise. Dans cette tâche, le système doit retourner la liste des topics dont la personne est probablement experte.

Ces deux tâches sont considérées comme deux faces d'une même pièce, (Balog et al., 2012). Ceci est clairement vu dans le cas d'une matrice de compétences (illustrée dans la *Figure 2*), proposée pour visualiser les différentes valeurs où les lignes sont les candidats, et les colonnes sont les domaines d'expertise (*knowledge areas*). Le repérage d'experts consiste à remplir une colonne, étant donné un domaine d'expertise (Remplir les candidats probables d'être experts dans ce domaine). L'expert profiling consiste à remplir les lignes (étant donné un candidat, on remplit les domaines d'expertise). Nous allons détailler ces deux tâches et les travaux de recherche qui leurs sont liés respectivement dans les sections 4 et 5.

	Area 1	Area 2	Area 3	.....	Area n-1	Area n
Candidate 1	●		●			●
Candidate 2		●	●		●	
Candidate 3	●	●			●	
⋮						
Candidate m-1			●		●	●
Candidate m	●	●				

Figure 2 : Matrice de compétences, (Balog et al., 2012)

3.1. Classification des approches de la recherche d'expertise

L'explosion des réseaux sociaux et les communautés a conduit à définir de nouveaux défis en matière de recherche et de profilage d'experts. En particulier, plusieurs approches ont été proposées pour améliorer le processus d'estimation de l'expertise. Ces méthodes se focalisent sur l'extraction automatique d'informations liées à l'expertise des utilisateurs à partir de collections de documents hétérogènes. En général, les collections considérées contiennent des documents provenant de résultats de recherche en ligne ou hors ligne, tels que des publications, des rapports techniques, des courriers électroniques, le contenu des sites de communautés de questions réponses (CQA) et des pages web, (Lin et al., 2017). Nous allons parler de la sélection des sources dans la section suivante.

Ces méthodes sont, en effet, classées selon l'algorithme utilisé dans le processus de repérage ou de profilage de l'expert. La classification la plus citée dans la littérature,

et celle que nous adoptons dans ce travail, est proposée par (Omidvar et al., 2014), (voir la *Figure 3*).

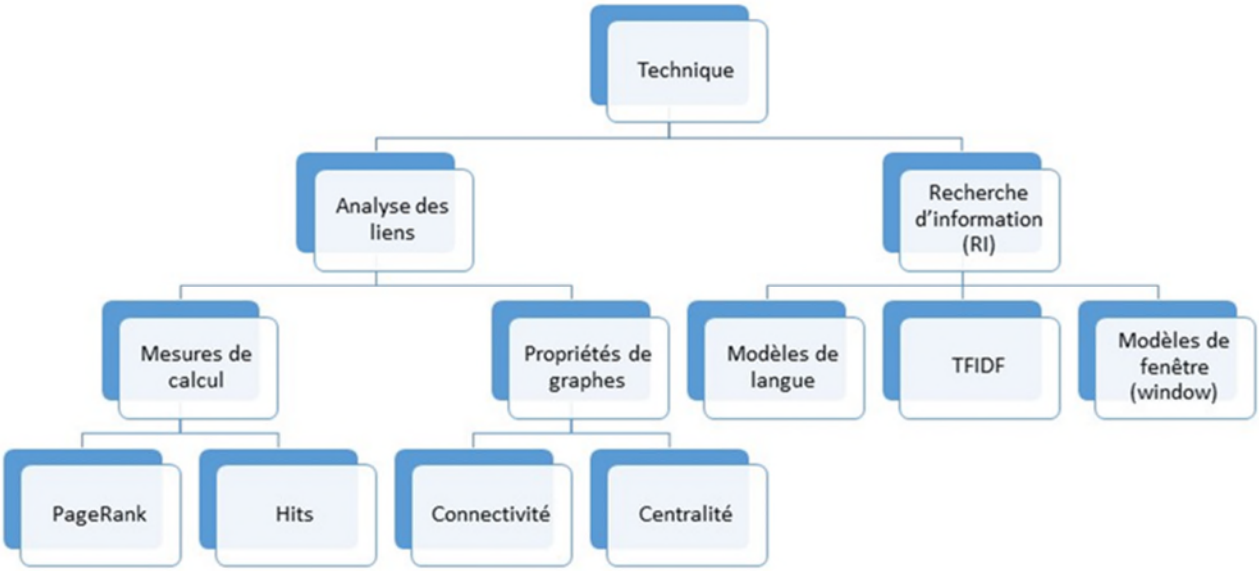


Figure 3: Classification des techniques de la recherche d'experts, (Omidvar et al., 2014)

Les auteurs considèrent deux catégories principales : les méthodes basées sur l'analyse des liens et les méthodes basées sur la recherche d'informations (RI). Cependant, d'autres classifications ont été proposées comme dans (Momtazi & Naumann, 2013; Serdyukov et al., 2011), où les auteurs classent les méthodes selon des approches basées sur des documents, des profils, des fenêtres et des graphiques. Une autre classification est donnée dans (Al-Taie et al., 2018), comme illustré dans la Figure 4 : techniques basées sur l'apprentissage automatique (*machine learning*) et techniques basées sur les graphes. Par ailleurs, les auteurs dans (Yuan et al., 2020) ont classé les méthodes en méthodes basées sur l'autorité et celles basées sur le sujet (*topic*).

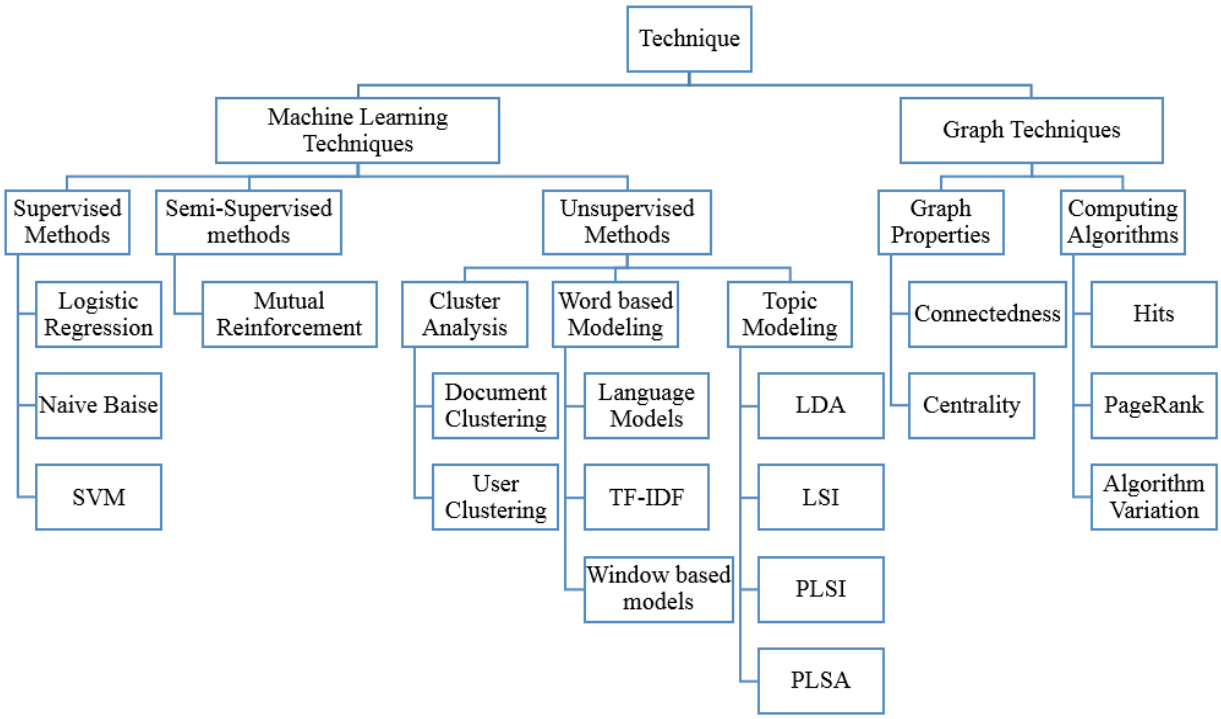


Figure 4 : Classification des techniques selon, (Al-Taie et al., 2018).

Nous adoptons la classification générique de (Omidvar et al., 2014), tout en faisant une nette distinction entre les approches proposées dans la recherche d'experts et celles proposées dans le profilage. Cette nouvelle classification que nous proposons est l'un des premiers apports de notre travail de recherche.

3.2. Sélection des sources d'information

Pour estimer l'expertise d'une personne dans un domaine précis, nous devons acquérir les données pertinentes de celle-ci. En général, les sources principales de données (appelée sources d'évidence) dans les systèmes de recherche d'experts actuels proviennent des trois types suivants : (1) les méta-bases de données, qui stockent les données personnelles, les coordonnées et les compétences professionnelles, et doivent être maintenues régulièrement. (2) Des collections de documents, qui contiennent des documents provenant de résultats de recherche en ligne ou hors ligne, tels que les publications, rapports techniques, courriers électroniques et pages Web. (3) Les réseaux de référence (*referral networks*), où l'expert est recommandé par quelqu'un qui connaît les connaissances et les compétences de celui-ci, (Lin et al., 2017) (cas par exemple de LinkedIn). Nous nous intéressons dans ce travail au deuxième type des sources de données.

Étant donné que la construction manuelle d'une base de données nécessite de nombreux efforts pour compléter et mettre à jour les données (type 1 des sources de données), les recherches se sont focalisées sur l'extraction automatique des informations sur les candidats experts à partir de collections hétérogènes de documents, à savoir le contenu des rapports et des publications. Ainsi que des communications (par exemple, courrier électronique, questions/réponses et commentaires sur les forums et les CQA), et des pages Web (par exemple, pages d'accueil et hyperliens dans les pages).

L'activité réseaux sociale des utilisateurs a été également utilisée comme source d'évidence, en plus des questions/réponses sur les CQA, et les commentaires nous citons les tags. Les tags sont surtout exploités autant que topic d'expertise, nous allons citer dans la section 5 quelques travaux qui se sont basés sur les tags. Nous revenons aussi sur les définitions, les avantages et les inconvénients du tagging social dans la section 6.

4. Repérage ou recherche d'experts

L'expert finding est souvent appelé découverte, recherche ou repérage d'experts. Il porte sur la tâche de trouver la bonne personne ayant des compétences et connaissances appropriées. Au sein d'une organisation, il peut y avoir de nombreux candidats possibles qui pourraient être des experts pour un sujet donné. Pour une requête donnée, le problème est d'identifier lesquels de ces candidats sont susceptibles d'être des experts. C'est la tâche qui a suscité le plus d'intérêt par rapport à la tâche du profilage de l'expert.

TREC (Text REtrieval Conference) Enterprise 2005 (utilisant une collection de W3C) a défini l'expert finding comme : « Étant donné une requêté liée à une thématique (topical query), trouver la liste des personnes W3C étant expertes dans le domaine de cette thématique », (Balog et al., 2012). Formellement, si nous avons une requête q, il s'agit d'estimer le niveau d'expertise d'un candidat expert e en calculant le score(e,q) qui sera le critère de classement décroissant des candidats.

Comme mentionné auparavant, les approches proposées pour la recherche d'experts sont classées en deux catégories : les approches basées sur la RI et celles basées sur l'analyse des liens ou la théorie des graphes. Il faut dire qu'il n'existe pas

dans la littérature, un regroupement de ces approches tel qu'il est proposé dans ce travail, que ce soit par catégorie (RI/analyse des liens) ou par tâche (*Finding/profiling*).

4.1. Approches basées sur la RI

Ces approches sont dites basées sur la RI, du fait qu'elles s'appuient sur les modèles de RI populaires comme les modèles vectoriels, probabilistes et modèles de langue (ML). Par ML on désigne une fonction de probabilité *P* qui assigne une probabilité *P(s)* à un mot ou à une séquence de mots s dans une langue. Une fois cette fonction définie, il est possible d'estimer la probabilité d'une séquence de mots quelconque dans la langue, ou d'un point de vue génératif, d'estimer la probabilité de générer cette séquence de mots à partir du modèle de la langue, (Boughanem et al., 2004). Les ML sont des modèles génératifs probabilistes, visant à modéliser une langue, utilisés principalement dans la linguistique informatique. Ces derniers sont très utilisés dans ce domaine de la recherche d'experts, des modèles formels ont été proposés dans (Balog et al., 2006). Dans le premier modèle, les auteurs construisent une représentation du candidat (c'est-à-dire, construisent un modèle du candidat) en utilisant les documents associés à celui-ci, et à partir de ce modèle la requête est générée. Dans le second modèle, la requête et le candidat sont considérés comme conditionnellement indépendants, et leur relation est résolue via les associations document-candidat. Les principales sources d'évidence utilisées dans ces travaux sont des collections hétérogènes de documents, regroupant les principales publications, les emails et les pages d'accueil. Les auteurs ont également proposé dans (Balog et al., 2009) un modèle de langue pour évaluer les experts, où deux stratégies pour les repérer sont modélisées et évaluées dans le contexte des systèmes de recherche d'entreprise dans un environnement intranet.

Des ML ont été également proposés dans (Balog et al., 2012). L'idée de base de ces modèles consistent à classer les candidats experts par la probabilité $p(ca \mid q)$, qui représente la probabilité qu'un candidat soit un expert sur le sujet q. Les auteurs ont relevé deux types de modèles probabilistes génératifs : les modèles de génération de candidats et modèles de génération de sujets.

Les modèles de génération de candidats calculent généralement, directement la probabilité $p\ (ca \mid q)$, tandis que les modèles de génération de sujet appliquent le théorème de Bayes pour induire la probabilité comme ci-dessous :

$$p(ca\backslash q) = (p(q\backslash ca).p(ca))/p(q) \qquad (1)$$

où $p\ (ca)$ est la probabilité d'un candidat et $p\ (q)$ est la probabilité d'une requête. Pour une requête donnée, $p\ (q)$ est considérée comme une constante, donc pendant le processus de classement, (1) peut être simplifiée comme :

$$p(ca\backslash q) = (p(q\backslash ca).p(ca)) \qquad (2)$$

Par conséquent, si l'on croit a priori que le candidat ca est un expert $(p\ (ca))$, $p\ (ca \mid q)$ est proportionnelle à $p\ (q \mid ca)$.

L'idée principale derrière les modèles de langue en RI est de considérer chaque document comme un échantillon de langue, généré par un modèle de langage spécifique. Le ML considère la pertinence du document pour une requête donnée comme une probabilité de génération de la requête par le ML du document. Supposons que nous avons un corpus de documents $D = \{d1, \ldots, dn\}$, θd est le ML d'un document d. Une requête représentée comme une séquence de termes $Q = \{t1, t2, \ldots, tm\}$. La probabilité de génération de la requête Q par d est calculée comme suit :

$$p(Q\backslash \theta d) = \prod_{t \in Q} p(t\backslash \theta d)^{n(t,Q)} \qquad (3)$$

Avec $n\ (t, Q)$ est le nombre de fois où t se produit dans Q, ce paramètre est souvent négligé car un terme t se produit rarement plus d'une fois dans la demande.

La probabilité $p\ (t \mid \theta d)$ est calculée avec le maximum de vraisemblance donné dans la littérature. Une technique de lissage (smoothing) est appliquée pour garantir qu'il n'y a pas de probabilités nulles. Le terme lissage fait référence à l'ajustement de l'estimateur du maximum de vraisemblance d'un modèle de langue afin qu'il soit plus précis.

Dans (Macdonald & Ounis, 2006) la recherche d'experts est considérée comme un problème de vote. Chaque profil de candidat est constitué d'un ensemble de documents associés au candidat et représentant son expertise. Une recherche *Adhoc* de documents, à partir d'une requête, est ensuite réalisée. Chaque document restitué est assimilé à un vote implicite pour le candidat associé. Les techniques de fusion de données, telles que CombSum ou RRF sont adaptées et appliquées pour déterminer le

score final d'un candidat en réponse à une requête. Les expérimentations menées sur la collection *TREC Enterprise 2005* ont montré que la qualité des documents en termes de structure est importante dans l'amélioration des résultats.

Dans (Petkova & Croft, 2008) plusieurs modèles de langues ont été proposés pour représenter les connaissances d'un expert, ces modèles sont issus de documents qui lui sont associés. L'approche se base sur la collecte des sources d'expertise à partir de plusieurs sources hétérogènes, en étudiant les associations entre documents et experts (le nom de l'expert est utilisé pour trouver les documents qui lui sont associés).

Le travail de (Moreira et al., 2011) a été dédié pour les publications académiques, où les auteurs se sont basés sur la technique de l'apprentissage supervisé pour classer les experts. Pour estimer l'expertise, l'approche est basée sur : (1) la similarité entre les termes de la requête et ceux des documents liés au candidat, les auteurs ont utilisé Okapi BM25, qui fait référence à la mesure de pondération BM25 utilisée dans la RI, pour calculer la pertinence du document, (2) les informations du profil (indépendamment de la requête) : le nombre de publications de conférences et de revues, durées entre publications, la moyenne de publications par an, etc., (3) le graphe de co-citation et co-auteur. Ce dernier élément est divisé en nombre de citations et l'index académique. Pour les citations, les auteurs ont utilisé le nombre total, la moyenne et le nombre maximum des citations des publications liées au candidat et à la requête. Plusieurs index ont été utilisés (H-index et ses extensions).

Par ailleurs, les auteurs dans (Neshati et al., 2017) ont introduit le nouveau problème du *future expert finding*, ou la recherche d'expert pour un temps futur, et prédisent le classement des experts dans le temps futur en proposant un cadre d'apprentissage et en utilisant la similarité des sujets (*topics*), les sujets émergents, le comportement des utilisateurs et la transition des sujets comme indicateurs de l'expertise.

Les travaux de Balog, ont été étendus par (Gharebagh et al., 2018), où les auteurs ont proposé deux modèles pour repérer et classer les experts en utilisant les réponses acceptées comme source d'évidence de l'expertise. Les candidats sont classés en 3 catégories : (1) les non experts, (2) les *T-shaped*, sont ceux qui ont des connaissances profondes dans un domaine donné, et superficielles dans d'autres, (3) les *C-shaped*, sont ceux qui ont des connaissances profondes dans différents domaines. Des

expériences sont réalisées sur trois collections réelles de tests générées à partir des données publiées dans *Stack Overflow*[2]. Les auteurs ont proposé une approche basée sur l'entropie (*EBA, Entropy Based Approach*), où les utilisateurs en forme de C (*C-shaped*) ont une entropie beaucoup plus élevée qu'en forme de T (*T-shaped*) en raison de la corrélation directe entre la diversité et l'entropie. Une version étendue de cette approche est également proposée par les auteurs (*XEBA*). Les deux approches améliorent le classement des utilisateurs en forme de T.

En utilisant un modèle de langue, les auteurs de (Al-Barakati & Daud, 2018) ont introduit l'influence du lieu (journal ou conférence) pour mieux trouver un expert. Selon les auteurs, un candidat qui publie dans des lieux spécifiques à un sujet, soit des revues ou des conférences, sera un expert sur un sujet par rapport à une autre publication dans des lieux multi-sujets.

Dans la majorité des cas de ces approches, les sources d'expertise utilisées sont les collections hétérogènes de documents incluant les emails, les rapports techniques, les publications scientifiques, et les pages web.

4.2. Approches basées sur l'analyse des liens

Les techniques basées sur l'analyse des liens sont principalement utilisées dans les sites communautaires de questions/réponse (*CQA*), selon les relations existantes entre les questions et les réponses. La communication par e-mail est également utilisée comme indicateur d'expertise. Ces techniques se concentrent sur l'analyse de la structure des liens entre les individus, plutôt que sur le contenu des réponses (Faisal et al., 2019). Elles sont basées sur le calcul de l'autorité de l'utilisateur en adaptant les algorithmes *PageRank* et *HITS (Hyperlink-induced topic search)*. En fait, l'intuition derrière le raisonnement adopté, est que l'expertise d'un candidat sur une certaine thématique dépend non seulement du fait que le candidat possède des connaissances pertinentes, mais aussi de son importance sociale ou de son influence dans la communauté.

Par exemple, les communications par e-mail ont été exploitées par (Campbell et al., 2003) pour proposer un modèle d'analyse des e-mails des utilisateurs et créer un graphe d'expertise à l'aide de Hits. Le processus suit trois étapes : (1) collecter tous les

[2] https://stackoverflow.com/

emails liés à une thématique, (2) analyser les e-mails entre chaque paire de personnes pour lesquelles il y avait une correspondance pertinente pour créer un graphe d'expertise et (3) analyser le graphe d'expertise pour obtenir des notes pour tous les expéditeurs et destinataires. Dans ce travail, la correspondance et le contenu du mail sont considérés. Cependant, l'expérience a considéré un petit réseau de 15 personnes.

Dans le même sens, (Fu et al., 2007) a proposé un processus de propagation d'expertise d'un candidat connu expert vers un autre candidat. Les auteurs se sont basés sur la force des associations entre les candidats pour construire un réseau social, utilisant un premier critère : le contenu des pages web. En effet, selon les auteurs, les candidats dont le nom se produit simultanément dans des contextes bien précis, peuvent partager les mêmes intérêts. La force de l'association est calculée donc sur la base de la fréquence de co-occurrence des candidats dans les pages web. Le deuxième critère est les correspondances par mail. La connexion entre deux candidats est donc calculée selon que ceux-ci apparaissent ensemble ou non dans les : *'from' 'to'* et '*cc*' d'un email. La performance de cette technique est sensible à la sélection des non-experts. Un ensemble de données d'entreprise a été utilisé, où les documents contenaient des informations de qualité. Ce n'est pas évident que cette même approche puisse être appliquée en ligne où la qualité de l'information est moindre et la taille du réseau communautaire est très vaste. L'utilisation des emails comme source d'évidence est adoptée généralement dans les cas d'entreprises où le consentement des candidats est possible. Dans le web communautaire, d'autres sources d'évidence sont adoptées telles que les activités sociales des candidats : réponses, commentaires, posts, tags, etc.

La production scientifique ou les documents rédigés ont été adoptés comme source d'évidence dans les approches basées sur la RI. Cette source est également exploitée dans ces approches basées sur l'analyse des liens, en se focalisant notamment sur les liens de co-auteurs. Comme dans le cas de (Rodriguez & Bollen, 2008) où les auteurs ont introduit une approche pour automatiser l'attribution de relecteurs d'articles dans le contexte de conférence avec évaluation par des pairs (*peer review*). La problématique attaquée consiste à trouver un équilibre entre identifier un relecteur qualifié tout en évitant les conflits d'intérêts entre auteurs et relecteurs.

L'approche est basée sur une modélisation du réseau des co-auteurs et un algorithme de propagation par essaims particulaires (*particle swarm*[3]). Les expérimentations menées sur les données de l'édition 2005 de la *ACM/IEEE Joint Conference on Digital Libraries (JCDL)* ont montré les limites de l'approche en termes de détection des conflits d'intérêts, penchant de ce fait pour une application durable du processus traditionnel d'affectation des relecteurs.

Les liens sociaux basés sur les relations de *posting/replying* (les postes liés aux questions-réponses dans les communautés du web) ont été largement adoptés comme sources d'évidence. (Jun Zhang et al., 2007) ont proposé de créer les réseaux d'expertise des candidats appelés réseaux *post-reply* en considérant chaque utilisateur participant comme un nœud, et lier l'ID d'un utilisateur commençant un fil de post à l'ID d'un répondant. *ExpertiseRank* proposé et basé sur PageRank, calcule l'autorité du candidat en fonction des réponses qu'il donne et pour qui il les donne (si le candidat ayant posé la question a déjà une certaine expertise, l'expertise de celui-ci sera propagée via le réseau de l'expertise). L'expérimentation est réalisée sur un forum, cependant, les auteurs n'ont pas expliqué comment ils catégorisent les postes en question en thématiques ou domaines.

Alors que les auteurs de (Kardan et al., 2011) ont proposé l'algorithme *SNPageRank* qui conduit à classer les utilisateurs en fonction de leur expertise sur *Friendfeed*[4] . *SNPageRank* est une adaptation de *PageRank*, qui dans ce cas, permet de calculer l'importance du candidat dans le réseau en se basant sur ses liens. Le lien entre deux candidats interprète le nombre de postes entre eux. La validation est réalisée en comparant les résultats avec le classement effectué au préalable par trois experts.

[3] Comme les réseaux de neurones, les algorithmes génétiques , le Particle Swarm Optimization (PSO) est un algorithme bio-inspiré. Il repose sur les principes d'auto-organisation qui permettent à un groupe d'organismes vivants d'agir ensemble de manière complexe, à partir de "règles" simples : La cohésion, L'alignement et La séparation.

[4] FriendFeed était un agrégateur en temps réel qui affichait des mises à jour de médias sociaux, de réseaux sociaux, de blogues et d'autres sources RSS. Crée en 2007, acheté par Facebook en 2009 et disparu en 2015.

Dans la même idée d'adaptation de *PageRank*, (Jiao et al., 2009; Wang et al., 2013) ont proposé l'algorithme *ExpertRank* pour classer les experts. En plus des liens sociaux entre candidats, les auteurs se sont basés aussi sur le contenu des postes dans la communauté. De ce fait, l'algorithme proposé est une combinaison de : (1) la pertinence de l'expertise basée sur le profil du candidat construit à partir de tous les documents (postes) dont il est l'auteur. Et (2) l'autorité du candidat calculée en considérant les liens du *posting* (poser la question) et *replying* (répondre à la question). Cette approche présente un inconvénient au niveau de la création du profil du candidat regroupant tous les documents du candidat (les postes peuvent appartenir à des thématiques différentes).

(Carchiolo et al., 2015) ont proposé également une approche d'évaluation de l'expertise afin d'améliorer la qualité des recommandations en sélectionnant celles fournies par des utilisateurs considérés comme experts dans le même contexte que celles-ci. Les utilisateurs notent les produits en écrivant des avis pour fournir des opinions aux visiteurs, les utilisateurs notent les avis des utilisateurs du point de vue utilité (utile ou non utile). L'approche proposée est spécifique au site Web Epinion[5]. Par ailleurs, dans (Xie et al., 2016), les auteurs ont proposé un modèle de recherche d'expert contextuel spécifique au sujet (à la thématique) (*TSCEFM : Topic-Specific Contextual Expert Finding Model*), dans lequel les auteurs s'appuient sur HITS dans le calcul de l'autorité, LDA pour extraire les caractéristiques liées aux thématiques, et les algorithmes SVM pour l'apprentissage de la fonction de notation (score) de l'expert. Les résultats démontrent que la méthode est faisable et peut trouver plus précisément des experts se conformant à une certaine thématique et à une situation contextuelle dont les utilisateurs ont besoin.

Une nouvelle technique de recherche d'experts dans les réseaux sociaux exploitant un modèle de données hypergraphique pour les réseaux sociaux multimédias (MSN) est présentée dans (Amato et al., 2019). Tout d'abord, les auteurs ont proposé un modèle MSN composé de trois types d'entités représentées par des nœuds (de la structure graphique) : les utilisateurs, les objets et les thématiques. Les annotations exprimant les différentes actions telles que le partage d'une photo, un commentaire, etc. La construction de ce modèle repose sur : (1) la construction de la structure hypergraphique, la distribution des thématiques et l'apprentissage de la

[5] https://epinionglobal.com/en/

similarité. Afin de classer les nœuds experts, des mesures de centralité sont définies pour calculer l'influence de l'utilisateur dans la communauté. Les auteurs exploitent le concept du voisinage (*neighborhood*) parmi les utilisateurs via l'algorithme des voisins les plus proches (*λ-Nearest Neighbors*) dans le MSN. Les résultats obtenus sur une collection de données de *Last.FM* montrent l'efficacité de l'approche proposée.

5. Profilage de l'expert

Il n'existe pas de terme bien défini de *l'expert profiling* dans la littérature en langue française, on l'appelle souvent profilage de l'expert. La présente tâche s'intéresse à trouver l'ensemble des thématiques (*topics, knowledge areas*) dont un candidat est expert, et donner son niveau de compétence dans chacune de ces thématiques. (Balog & De Rijke, 2007) introduisent le profil thématique et le définissent comme étant : « Un relevé des types et domaines de compétences et de connaissances de cette personne, ainsi qu'une identification des niveaux de compétence dans chacun ». Le résultat du profilage est une liste classée de domaines de compétence, le niveau estimé des compétences dans chacun d'eux, ainsi que la source adoptée pour cette estimation. La tâche de déterminer les profils d'experts est naturellement décomposée en deux étapes. Elles sont définies comme suit :

- Découvrir et identifier les domaines de connaissances (*Knowlegde areas, Ka*) possibles (que nous appelons dans notre travail, sujet, thématique ou *topic*).
- Mesurer la compétence de la personne dans ces domaines.

Les auteurs se sont positionnés à la deuxième étape, et suppose un ensemble de *Ka* pour un candidat donné (prédéfini) où le profil est définit comme suit :

$$Profile\ (ca) = score(ca, ka_1), score(ca, ka_2) \dots \dots, score(ca, ka_n) \quad (4)$$

Pour estimer le score, les auteurs proposent deux méthodes :

Méthode 1

La première méthode consiste à considérer que la compétence d'une personne peut être représentée par un score sur des documents pertinents dans un *Ka* particulier.

Pour chaque *Ka*, un sous-ensemble de documents D_{ka} est obtenu en utilisant les *n* premiers documents récupérés pour la requête *Ka*. Une somme des pertinences des documents pertinents associés à l'utilisateur est calculée de la sorte :

$$score(ca, ka) = \sum_{d \epsilon D_{ka}} relevance\ (d, ka) A(d, ca) \qquad (5)$$

Où $relevance\ (d, ka)$ est la pertinence du document par rapport au *Ka*, calculée en utilisant un modèle de langue. Et $A(d, ca)$ est l'association ou non de l'utilisateur au document (par nom ou email) ayant les valeurs *1* si l'utilisateur est associé au document et *0* sinon.

Les travaux de Balog sont une référence dans le domaine, néanmoins nous pouvons citer quelques limites :

1. L'ensemble des *Ka* n'est pas découvert, les auteurs se basent sur un ensemble prédéfini qui peut ne pas converger avec les *Ka* des candidats ;
2. L'association des documents aux utilisateurs peut avoir des ambiguïtés, telles que la ressemblance des noms (appartenance du nom à plusieurs candidats, le nom peut être incomplet) ;
3. Un candidat peut être concerné par une partie seulement du document (pour les documents ayant plusieurs auteurs) ce qui peut fausser l'association du candidat aux *Ka* pour lesquels le document est pertinent.

Méthode 2

La deuxième méthode est une approche complétement différente. Un ensemble de mots clé (*KW)* représentant le plus possible le document est constitué en utilisant TF-IDF, noté *KW(d)*. L'ensemble de *KW* de documents liés au *Ka* (D_{Ka}), est obtenu en utilisant les top *n* documents retournés pour la requête *Ka* sur la collection de documents W3C. le résultat est :

$$KW_{ka} = \bigcup_{d \epsilon D_{ka}} KW(d) \qquad (6)$$

Les *KW* liés au candidat *ca* sont obtenus sur les documents associés à ce dernier :

$$KW_{ca} = \bigcup_{d \epsilon D, A(d,ca)=1} KW(d) \qquad (7)$$

Après avoir obtenu les *KW* des documents liés au *ka* et au candidat *ca* le *score(ca,ka)* est donné par :

$$score(ca, ka) = |KW_{ka} \cap KW_{ca}| / |KW_{ka} \qquad (8)$$

Avec cette deuxième méthode également les *ka* ne sont pas découverts et les problèmes liés à l'association du candidat au document demeurent présents.

(Fazel-Zarandi & Fox, 2011) ont proposé un modèle de profilage d'expert utilisant un énoncé de compétences qui stipule que « l'individu *p* possède des compétences de niveau de maitrise d'au moins *1* ». Chaque ressource humaine aura un profil composé d'un ensemble de telles déclarations de compétences. Les auteurs ont proposé qu'à tout moment un énoncé de compétences puisse être dans un et un seul des quatre états : *démontré*, *suggéré*, *inconnu* ou *réfuté*. Le statut de l'information est « *inconnu* » pour l'auto-déclarations, il peut être « *suggéré* » ou « *réfuté* » à mesure que l'information devient disponible jusqu'à ce qu'elle soit démontrée.

De plus, les auteurs dans (Berendsen et al., 2013) ont proposé une extension d'une collection de tests pour la recherche d'experts, sur la base des propositions de (Balog & De Rijke, 2007). Les auteurs ont utilisé les systèmes de profilage existants pour générer un profil d'expert et obtenir des commentaires d'experts sur la qualité des domaines de connaissances générés par le système.

Des approches de profilage d'expertise temporelle ont été proposées dans (Y. Fang & Godavarthy, 2014; Rybak et al., 2014). Des modèles probabilistes et des modèles de langue prédictifs ont été adoptés pour modéliser la dynamique de l'expertise. (Y. Fang & Godavarthy, 2014) a proposé un modèle prédictif sur les distributions des domaines d'expertise sur les prochaines publications de l'expert. (Rybak et al., 2014) ont proposé un profil hiérarchique sou forme d'arbre, où les nœuds sont les topics et les arcs sont les relations hiérarchiques (les auteurs ont utilisé la classification d'ACM) entre les topics, les poids des nœuds représentent les niveaux d'expertise. Les travaux sont basés sur les documents rédigés.

(Liang, 2018) ont également proposé un modèle de suivi de l'expertise des utilisateurs, utilisant un algorithme de profilage en continu (SPA) pour résoudre le problème du profilage dynamique des utilisateurs.

D'autres travaux se sont intéressés aux activités sociales des candidats principalement les tags, nous citons ci-dessous quelques approches.

(Budura et al., 2009) ont développé une méthode probabiliste pour construire le profil d'expertise des utilisateurs en fonction de leurs tags. L'intuition derrière est que les ressources marquées par les utilisateurs reflètent leurs intérêts, qui sont corrélés à

l'expertise des utilisateurs. L'approche est basée sur le calcul de la corrélation entre les tags et les compétences (c-à-d l'expertise auto-déclarée). La méthode proposée est basée sur une série de tests appliqués sur un ensemble de données réelles de l'application d'IBM-Internal pour répondre aux questions suivantes : (1) Les termes qui définissent les compétences sont-ils également utilisés comme tags dans un système de bookmarking social ? (2) Pour un utilisateur donné, existe-t-il un chevauchement entre son jeu de tags et les compétences qui définissent son profil d'expertise ? (3) En général, les tags les plus populaires du système de bookmarking social se réfèrent-ils à des domaines d'expertise ? (4) Pour un utilisateur donné, ses tags les plus populaires sont-ils aussi ses compétences ? Les auteurs ont fait quelques observations. Globalement, le comportement du tagging des utilisateurs est lié à des sujets d'expertise. De tous les termes utilisés par un utilisateur pour le tagging, très peu sont également inclus dans son ensemble de compétences. Globalement, la popularité d'un tag est un bon indicateur que ce tag représente une compétence. Les tags les plus populaires d'un utilisateur ne correspondent pas à ses compétences. Les auteurs ont également proposé d'utiliser un modèle de langue pour le cas du profilage d'expertise, et par simulation, ils ont proposé un modèle de notation pour représenter la pertinence entre les compétences (requêtes en ML) et un ensemble de tags (documents en ML).

Par ailleurs, (Serdyukov et al., 2011) considère qu'un échantillon représentatif d'utilisateurs de l'entreprise a déjà décrit son expertise en utilisant des tags. Le but du système proposé est d'attribuer automatiquement des tags à partir du vocabulaire contrôlé de tags créé par l'ensemble initial d'utilisateurs à d'autres employés de l'entreprise. Les auteurs ont utilisé une approche basée sur un modèle de langue pour obtenir une estimation de la probabilité de pertinence $P(e, t)$ du tag t par rapport à l'expertise personnelle de l'employé e, compte tenu d'un flux de preuves S (par exemple, un ensemble de documents rédigés). Les sources considérées sont les documents d'entreprise écrits ou liés aux employés, des documents issus du Web, les listes de discussions internes à l'entreprise et les clics issus des fichiers de consultation (logs) du moteur de recherche intranet de l'entreprise.

Dans (Ribeiro et al., 2015), les auteurs ont considéré le problème de profilage d'expertise comme un problème de recommandation de tags (*people tagging*). L'approche vise à générer une liste représentative de tags pour décrire les sujets d'expertise des chercheurs. Les auteurs ont utilisé différentes combinaisons

d'algorithmes de recommandation de tags (L2R) et de sources d'évidence de l'expertise d'un chercheur. Les principaux indicateurs d'expertise utilisés sont les titres des publications les résumés et les mots-clés. Les tags ne sont pas exploités comme un indicateur social de l'expertise du candidat.

Pour un objectif de recommandation de films, les auteurs de (Faggioli et al., 2019) ont proposé un système de recommandation basé sur le profil utilisateur créé avec ses tags regroupés en sujets utilisant l'algorithme du regroupement agglomératif hiérarchique (*hierarchical agglomerative clustering algorithm*). Le système utilise des concepts théoriques du jeu afin de trouver une description à la fois assez générale et capable de rassembler les facettes spécifiques de l'utilisateur. La motivation derrière l'utilisation du clustering agglomératif hiérarchique n'est pas explicitement décrite par les auteurs, tandis que d'autres algorithmes plus appropriés pour la modélisation de sujets tels que LDA, LSA (*Latent Semantic Analysis*), PLSA (*Probabilistic Latent Semantic Analysis*) peuvent également être utilisés.

(Dehghan et al., 2019) ont proposé une méthode pour construire des arbres d'expertise pour les candidats experts. L'arbre d'expertise est composé de trois niveaux : le troisième niveau est représenté par les tags, le second par les *skill area*, le premier niveau (racine) par le domaine de tous les *skill area*. La source d'évidence considérée est les réponses acceptées.

Pour pallier au problème lié à la différence du langage utilisé entre les candidats experts dans *Stack Overflow* et les entreprises qui cherchent des compétences en vue d'un emploi, (Nobari et al., 2020) a proposé une méthode pour traduire les domaines de compétence (*skill area*) en leurs principaux mots. Les tags donnés associés aux questions sont considérés comme les *skill area*, comme ils peuvent être considérés étant des requêtes exprimant un besoin particulier en termes de compétence. La source d'évidence de l'expertise du candidat est dans ce cas le contenu des réponses de celui-ci. L'approche proposée est basée sur le plongement de mots ou le plongement lexical (*word embedding*) qui vise à réduire les *skill area* (tags associés aux questions et requêtes) d'une part et les termes de documents (dans ce cas les réponses aux questions) en un même espace.

6. Tagging et folksonomies : avantages et inconvénients

Marquage collaboratif, étiquetage collaboratif, tagging social, annotations sociales ou tagging collaboratif, différentes appellations désignant toutes ce phénomène qui est apparu ces dernières années et qui ne cesse de gagner en popularité sur le web, (Kichou et al., 2011). Marquer un contenu par des termes descriptifs est une manière d'organiser ce contenu pour une navigation future, un filtrage ou une recherche. Le tagging collaboratif est devenu un moyen de plus en plus courant pour le partage et l'organisation du contenu web. D'autres concepts découlent du tagging telles que la folksonomie et l'indexation collaborative.

6.1. Définitions

Tagging social ou collaboratif : On désigne par tagging collaboratif le processus qui consiste à associer un ou plusieurs "tags" (mot clé) à un document numérique (page web, photo, vidéo, billet de blog) dans un environnement multi utilisateurs. Selon (Golder & Huberman, 2005) le tagging collaboratif décrit le processus par lequel plusieurs utilisateurs ajoutent des métadonnées à un contenu partagé.

Tag *:* Un tag ou étiquette est un mot clé librement choisi par un utilisateur pour décrire un objet dans le web (document texte, image, fragment d'un document). Un tag peut être vu simplement comme étant un jeu de mots-clés librement choisi par les utilisateurs. Et le fait que ces derniers ne soient pas spécialistes de l'information, leurs tags ne suivent aucune indication formelle. Cela signifie que ces mots peuvent être catégorisés avec n'importe quel mot définissant une relation entre la ressource et un concept issu de l'esprit de l'utilisateur. Un nombre infini de mots peut être choisi, dont quelques-uns sont issus de représentations évidentes tandis que d'autres ont peu de signification en dehors du contexte de l'auteur du tag, (Guy & Tonkin, 2006).

Folksonomie *:* La Folksonomie est un terme anglais introduit en 2004 par Vander Wall, exprimant l'idée d'une classification (Taxonomie) faite par les utilisateurs (*Folks*). Une folksonomie est le résultat de la collecte de données du Tagging pour un groupe donné, elle est donc liée à un site communautaire bien particulier : par exemple,

la folksonomie de *Flickr* est différente de celle de Youtube. Cependant elle est souvent confondue avec son processus de création qu'est le Tagging collaboratif.

Les folksonomies sont des séries de métadonnées créées en collectif par les utilisateurs pour catégoriser et retrouver les ressources en ligne (Broudoux, 2006). Une folksonomie est différente d'une taxonomie car, d'une part, elle n'est pas contrainte par des relations hiérarchiques, et d'autre part, elle n'est pas conçue par des experts. Il ne s'agit pas non plus d'une ontologie. Une ontologie est un ensemble structuré de concepts, alors qu'une folksonomie ne possède qu'une structure émergente, floue, et non contraignante (exemple un utilisateur peut utiliser un tag dans un sens totalement différent des autres utilisateurs).

6.2. Avantages et inconvénients

La liberté du choix des tags offre aux folksonomies un certain nombre d'avantages, cette même caractéristique est à double tranchant et provoque des limites : (Mathes, 2004), résume ce paradoxe en une autre phrase paradoxale : « Une folksonomie représente en même temps ce qu'il y a de meilleur et de pire dans l'organisation de l'information ». Un vocabulaire non contrôlé, ne peut qu'avoir des limites :

Ambiguïté : on parle d'ambiguïté quand un même tag dénote deux concepts différents : le terme orange pour le fruit, la couleur ou la société française de télécommunication.

Hétérogénéité : c'est lorsqu'un un tag se présente en différentes formes. L'hétérogénéité englobe la variation d'écriture : New York et New_York, les synonymes : 'mac' et 'macintosh', ou télévision et TV, l'utilisation ou non du pluriel : 'flower' et 'flowers', et le multilinguisme : 'chat' et 'cat'.

Info-pollution ou *Spamming* : certains utilisateurs malveillants, peuvent nuire en inondant les contenus de tags inadéquats. Effet indésirable notamment pour les sites du e-commerce.

Or les folksonomies marchent bien, certes elles ne présentent pas que des inconvénients ; en quoi donc consiste leurs forces ?

Peu coûteuse : Puisque la folksonomie est réalisée par les utilisateurs finaux et non pas par des professionnels.

Tagging continu, folksonomie dynamique : Elle est mise à jour automatiquement et en permanence (au fur et à mesure de l'activité des utilisateurs).

Folksonomie intuitive : puisqu'elle est le fruit de collaboration de simples utilisateurs.

Améliore les résultats de recherche : Avec une folksonomie on peut tomber sur des documents qu'un moteur de recherche classique aurait pu ignorer (documents non-indexés par le moteur).

Utilisée pour la veille : Un point fort des folksonomies et la possibilité de leur utilisation pour le suivi (*tracking*) : Technorati utilise le *tracking* de termes précis, on retrouve les blogs où sont employés ces termes.

7. Conclusion

Si les définitions ne manquent pas de ce terme générique qu'est la compétence, un consensus se dégage tout de même autour de l'idée que celle-ci est une capacité à agir en utilisant ou combinant différentes ressources qui ont pu être acquises par la formation ou l'expérience.

Nous avons évoqué dans ce chapitre la notion de la compétence, ses différentes définitions, et sa relation avec la notion d'expertise. Nous avons par la suite introduit les tâches principales liées à la recherche d'expertise, à savoir la recherche d'experts et le profilage d'expert. Nous avons vu les différents travaux dans ce domaine, essayé de mettre en évidence les limites de ceux-ci. Nous sommes revenus sur les différentes notions liées au tagging social, ses avantages et inconvénients comme introduction à notre proposition qui se base en grande partie sur les activités du tagging social comme principale source d'évidence de l'expertise.

Notre exploration de l'état de l'art du domaine de la recherche d'expertise a mis en évidence plusieurs limites des travaux existants. L'utilisation des documents rédigés comme principale source d'évidence, même si ceux-ci sont d'une grande crédibilité pour l'expertise du candidat, reste source de certains problèmes. L'association du document à l'auteur, en est le premier. En effet, les noms des auteurs peuvent être

ambigus, incomplets, etc. De plus, pour les cas du web communautaire, des experts peuvent ne pas avoir de production scientifique ou autre documents rédigés. Mais aussi, les opportunités de l'utilisation des emails sont très réduites, elles ne sont possibles qu'avec consentement du candidat.

En outre, pour ce qui est du web communautaire et avec la profusion de l'information sociale, les travaux de la littérature ont exploité en grande majorité les réponses aux questions ainsi que les liens engendrés. Toutefois les tags sont les moins exploités en tant que source d'évidence malgré leur force à décrire et représenter une personne. Mais aussi leur capacité à donner de l'information sur les éventuelles expertises des candidats.

Les solutions que nous allons proposer sont dédiées aux cas du web communautaire. Elles se basent sur les tags comme principales source d'évidence, ceci en exploitant le *topic modeling* pour découvrir les éventuelles thématiques d'expertise. De ce fait, nos propositions appartiennent à la catégorie d'approches basées sur la RI et touchent les deux tâches de la recherche d'expertise à savoir la recherche et le profilage d'experts.

Nous allons par ailleurs, dans le prochain chapitre présenter l'apport du *topic modeling* ou la modélisation de sujets dans le domaine de la recherche d'expertise.

Chapitre II
Le Topic Modeling dans la Recherche d'Expertise

1. Introduction

La progression des techniques du web et de gestion de big data en termes de stockage et d'échange sur internet, a induit une explosion des réseaux sociaux du point de vue volume de données et nombre d'utilisateurs. Le fonctionnement quotidien et permanent des réseaux sociaux tel que Twitter a changé carrément l'image du web 2.0 et lui a infligé de nouveaux défis.

Avec ce volume important de données, les sites de réseautage social ont pris un grand champ de recherche pour plusieurs acteurs de la société. Ces domaines de recherche concernent principalement le suivi de campagnes publicitaires, l'étude de marché, l'analyse de nouveauté, l'analyse des comportements humains, l'identification des personnes influentes, détection des maladies, et détection de l'opinion publique concernant un sujet donné, etc.

L'un des objectifs les plus critiques de l'analyse de données est de déterminer les caractéristiques que ces données indiquent. Dans l'analyse de texte, cela signifie souvent de déterminer quels événements ou concepts un document traite. Cette information est claire pour une personne lisant le document, mais un programme donne uniquement le texte tel qu'il est écrit sans connaitre le sujet derrière chaque document (Vayansky & Kumar, 2020). Afin d'automatiser cette tâche en un programme, les *data scientist* utilisent, entre autres, des techniques appartenant au *topic modeling*.

L'extraction des informations pertinentes à partir de ces sites communautaires est l'un des objectifs majeurs des recherches dans ce domaine. Les statistiques génératives et les modèles de distribution du texte apportent d'importantes contributions à l'analyse statistique de grandes collections de documents telles que

celles qu'on peut collecter à partir des réseaux sociaux, (Blei, Ng, et al., 2003). L'un des moyens utilisés est le *topic modeling* (ou la modélisation de sujets).

Le topic modeling, que nous appellerons aussi dans ce document modélisation de sujets ou modélisation thématique, est un ensemble de techniques d'apprentissage automatique non supervisées qui est capable d'étudier un ensemble de documents, de détecter des modèles de mots et de phrases qu'ils contiennent et de regrouper automatiquement des groupes de mots et des expressions similaires qui caractérisent le mieux un ensemble de documents. C'est l'une des techniques les plus puissantes de l'exploration de texte, pour la fouille de données ou la découverte de données latentes, ainsi que pour la recherche de relations entre les données et les documents texte (Jelodar et al., 2019). Il est appliqué dans divers domaines tels que le génie logiciel, les sciences politiques, les sciences médicales et linguistiques, etc. Il faut mentionner aussi que les approches basées sur le *topic modeling* dans la recherche d'expertise font partie des approches puissantes pour surmonter le problème du manque de vocabulaire et devraient également surpasser les approches basées sur les candidats et les documents, déjà citées dans le chapitre précédent.

Nous allons dans ce chapitre présenter les notions liées au *topic modeling*, son utilité et ses différentes utilisations dans le domaine de la recherche et le profilage d'experts.

2. Définitions

La modélisation thématique a été développée à l'origine dans les années 80 et s'articule autour du domaine de « modélisation probabiliste générative » (Liu et al., 2016). Ce type de modélisation suppose que les variables observées interagissent avec des paramètres non observés ou latents dans une relation probabiliste spécifique qui génère les données dans un ensemble de documents (Steyvers & Griffiths, 2007).

En apprentissage automatique et en traitement automatique du langage naturel, un topic model est un modèle probabiliste permettant de déterminer des sujets ou thèmes abstraits dans un document. La Figure 5 ci-dessous montre un exemple d'un document traité par le topic modeling.

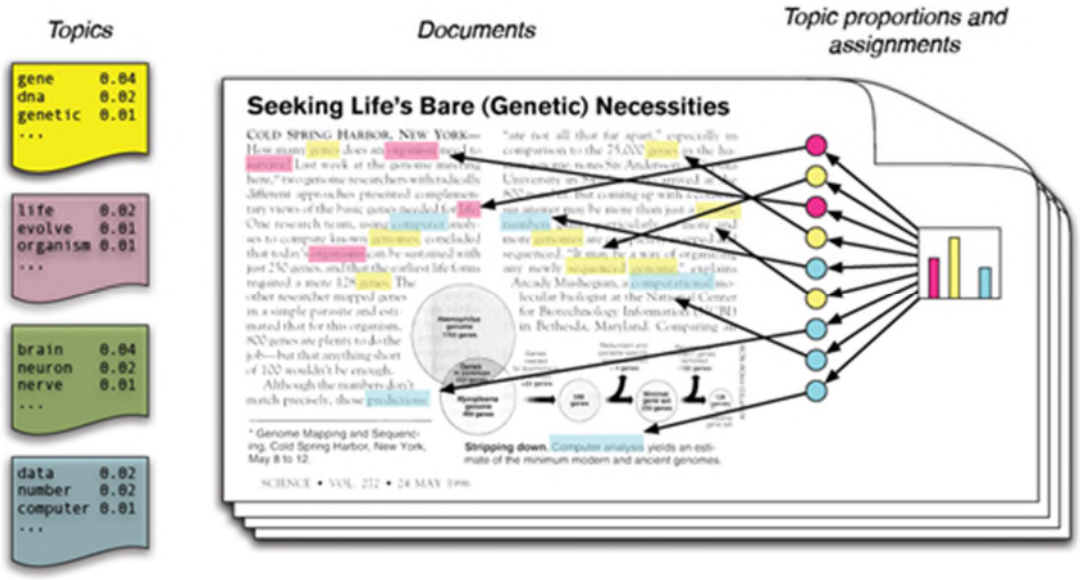


Figure 5 : Exemple d'un document traité avec une technique du topic modeling, (Bietti, 2012)

La modélisation de sujet fait référence au processus de division d'un corpus de documents en deux :

1. Une liste des thèmes couverts par les documents du corpus ;
2. Plusieurs ensembles de documents du corpus regroupés selon les thèmes qu'ils couvrent.

L'hypothèse posée est que chaque document comprend un mélange statistique de sujets, c'est-à-dire une distribution statistique de sujets qui peut être obtenue en regroupant toutes les distributions pour tous les sujets couverts. Les méthodes de modélisation de sujets essaient de déterminer quels sujets sont présents dans les documents du corpus et à quel point cette présence est forte.

La modélisation thématique fournit des méthodes pour organiser, comprendre, rechercher et résumer de grandes archives électroniques :

1. Découvrir les thèmes cachés de la collection ;
2. Annoter les documents selon ces thèmes ;
3. Utiliser des annotations pour organiser, résumer et rechercher.

L'idée fondamentale de la modélisation de sujets suppose qu'il existe une couche de sujets cachée $Z = \{z_1, z_2, z_3, \ldots, z_t\}$ entre le mot *token* et les documents, où z_i désigne un sujet latent et chaque document d est un vecteur de N_d mots w_d. Une collection de documents D est définie par $D = \{w_1, w_2, w_3, \ldots, w_d\}$ et chaque mot w_d est choisi parmi un vocabulaire de taille V. Pour chaque document, une distribution de 'mixture' de

sujets est échantillonnée et un sujet latent *Z* est choisi avec la probabilité que ce sujet donne le document.

3. Les Techniques du topic modeling

Il existe différentes méthodes de modélisation de sujets. Nous présentons les approches les plus pertinentes.

3.1. Latent Semantic Indexating (LSI)

Le modèle LSI (*Latent Semantic Indexing*) proposé dans (Deerwester et al., 1990), initialement pour introduire l'aspect sémantique et améliorer le retour des documents pertinents en remédiant aux problèmes de synonymie et polysémie.

LSI appelée aussi LSA suppose que les mots dont le sens est proche apparaîtront dans des morceaux de texte similaires (hypothèse distributionnelle). Une matrice contenant le nombre de mots par document (les lignes représentent les documents et les colonnes représentent les termes) est construite à partir d'un gros morceau de texte et une technique mathématique appelée décomposition en valeurs singulières (SVD) est utilisée pour réduire le nombre de lignes tout en préservant la structure de similarité parmi les colonnes.

LSI est une approche vectorielle qui exploite les co-occurrences entre termes, réduit l'espace des termes, en regroupant les termes co-occurrents (similaires) dans les mêmes dimensions. Les documents et les requêtes sont alors représentés dans un espace plus réduit, composé de concepts de haut niveau. Le score utilisé est le TF-IDF, la Figure 6 ci-dessous montre un exemple de matrice documents-termes de dimension m*n.

Terms

Documents	T1	T2	T3	…	Tn
D1	0,2	0,1	0,5	…	0,1
D2	0,1	0,3	0,4	…	0,3
D3	0,3	0,1	0,1	…	0,5
…	…	…	…	…	…
Dm	0,2	0,1	0,2	…	0,1

Figure 6: Exemple d'une matrice document-terme, (Deerwester et al., 1990).

(Deerwester et al., 1990) affirme que les fonctionnalités dérivées de LSI, qui sont des combinaisons linéaires des fonctionnalités TF-IDF d'origine, peuvent capturer certains aspects des notions linguistiques de base telles que la synonymie et la polysémie.

LSI adopte aussi la représentation en sac de mots des documents texte pour l'extraction des mots avec une signification similaire (AlSumait et al., 2009). Comme toute autre technique, LSI présente des avantages et des inconvénients. Pour ses avantages : elle est facile et rapide à implémenter, et côté RI, elle donne des résultats meilleurs qu'un simple modèle vectoriel. Cependant, puisqu'il s'agit d'un modèle linéaire, il pourrait ne pas fonctionner correctement sur les ensembles de données avec des dépendances non linéaires. Elle suppose une distribution gaussienne des termes dans les documents, ce qui peut ne pas être vrai pour tous les problèmes. LSA implique SVD, qui est intensif en calcul et difficile à mettre à jour à mesure que de nouvelles données apparaissent.

3.2. Probabilistic Latent Semantic Indexing (PLSI)

Le modèle PLSI (Probabilistic Latent Semantic Indexing) introduit dans (Hofmann, 1999)**,** a été proposé principalement pour améliorer et donner une base probabiliste au modèle précédent (LSI). Le cœur de PLSI est un modèle statistique qui a été appelé modèle d'aspect. Ce dernier est un modèle variable latent pour les données générales de co-occurrence qui tente de définir un modèle génératif de donnée.

Le modèle du PLSI peut être défini comme suit :

C'est un modèle statistique des données de co-occurrence qui associe un groupe de variables non-observées $z \epsilon Z=\{z_1,z_2,.....z_k\}$ à chaque occurrence du terme $t \epsilon V$ dans un document $d \epsilon D$.

Si R_d est la pertinence d'un document (égale *1* si pertinent et *0* sinon), *q* est la requête, l'objectif est de calculer la probabilité d'une requête sachant le document : $P(q \setminus R_d=1)$. Cela revient à calculer pour chaque document, la probabilité que chaque mot *w* provienne (ou soit pertinent pour) ce document, c'est-à-dire $P(w \setminus Rd=1)$. Puis calculer la probabilité conditionnelle des mots de la requête *q*.

PLSI est considérée générative uniquement au niveau des mots mais pas au niveau des documents, ce qui est considéré comme sa principale limitation, elle ne peut

permettre de faire une prédiction pour un nouveau document ajouté au corpus. Par conséquent un nouveau modèle a été proposé, à savoir le *Latent Dirichlet Allocation* (LDA).

3.3. Latent Dirichlet Allocation (LDA)

L'approche la plus adoptée dans la littérature pour la modélisation thématique est l'algorithme d'allocation latente de Dirichlet (Blei, Ng, et al., 2003), qui est un modèle probabiliste génératif. L'idée de base est que les documents sont représentés comme des mélanges aléatoires sur des sujets latents, où chaque sujet est caractérisé par une distribution sur des mots. LDA a une finalité essentielle de classification, elle permet d'associer un contexte à un document à partir de mots contenus dans ce document, lesquels mots pris individuellement pourraient appartenir à des contextes différents. Par exemple, le terme «java» peut faire référence au langage de programmation ou au café. L'analyse des mots proches de ce mot dans une page permet de dire si la page ou le paragraphe est lié à des langages de programmation ou du café. Les termes «Ipad», «PC» et «ordinateur portable» cachent également une relation sémantique existante, en effet les trois termes appartiennent au sujet «ordinateur» (Zhu et al., 2011).

La *Figure 7* ci-dessous illustre le modèle graphique de LDA. α et β sont appelés *hyperparamètre*s. α donne les proportions moyennes de chaque sujet pour chacun des documents, c'est un vecteur de réels positifs de taille k (k est le nombre de topics). β est une matrice de dimension $k*V$, avec V la taille du vocabulaire associé au corpus. Où $\beta_{i,j}$ est la probabilité d'appartenance du terme w_j au topic z_i ($\beta i,j = p(w_j | z_i)$). L'algorithme Espérance-maximisation (EM) est souvent utilisé pour l'estimation de β.

M est le nombre de documents du corpus, N est le nombre des termes dans un document. Les paramètres θ, z, w sont appelés paramètres *latent*s.

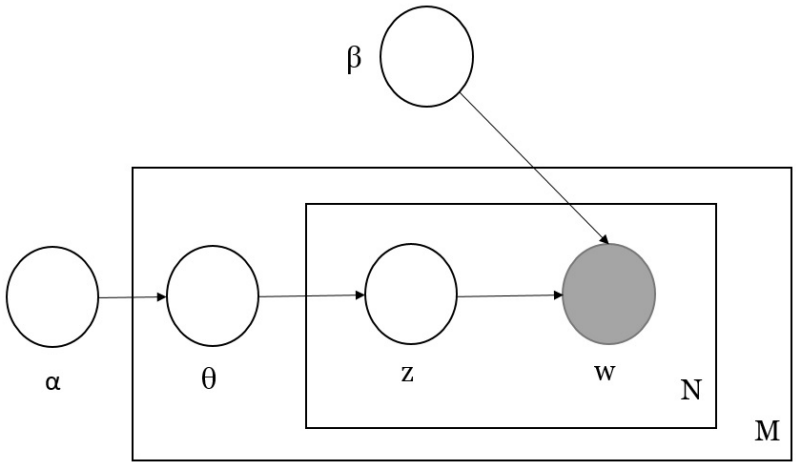


Figure 7: Modèle graphique de LDA, (Blei, Ng, et al., 2003).

Les trois niveaux de LDA sont comme suit :

- Les hyperparamètres α et β sont des paramètres globaux définis pour le corpus tout entier, les autres variables sont générées à partir de ceux-ci.
- Le paramètre θ est au niveau document. Il représente la proportion exacte des topics dans chaque document.
- Les paramètres z et w sont au niveau terme (word). Les z_i; $i \in [1, K]$, sont les topics associés à chaque mot w_i d'un document.

La distribution des termes $p(w|\theta,\beta)$ est donnée par:

$$p(w|\theta,\beta) = \sum_z p(w|z,\beta)p(z|\theta) \quad (9)$$

Un processus génératif pour un document est défini comme suit :

1. Pour Chaque document d
2. Générer $\theta^{(d)} \sim$ Dirichlet(α)
3. Pour Chaque $i \in [1, N^{(d)}]$
4. Générer $z_i \in [1, K] \sim$ Multinomiale($\theta^{(d)}$)
5. Générer $w_j \in \{1, \ldots, V\} \sim$ Multinomiale(β_{zi})
6. Fin Pour Chaque
7. Fin Pour Chaque

LDA modélise un document comme un mélange de plusieurs sujets avec des proportions différentes en répondant aussi à la synonymie et la polysémie. Par conséquent, LDA franchis éventuellement les inconvénients des autres modèles de sujet (tel que LSI et PLSI) en raison de ses propriétés génératives et la possibilité d'attribuer plusieurs sujets pour chaque document.

Le modèle LDA suppose que les mots de chaque document proviennent d'un mélange de sujets, chacun d'entre eux représente une distribution sur le vocabulaire du contenu de ce document. L'apprentissage d'une hiérarchie de sujets à partir d'une collection de documents est l'un des défis les plus complexes. Étant donné une collection de documents, chacun contenant un ensemble de mots, le défi dans ce cas n'est pas seulement de découvrir les sujets dans les documents, mais encore d'organiser ces sujets dans une hiérarchie. Une extension du modèle LDA, nommée LDA hiérarchique (HLDA) a été développée à cette fin dans (Blei, Griffiths, et al., 2003). LDA suppose que les documents sont inchangeables, dans certaines situations cette hypothèse est trop restrictive, et cela revient à l'évolution des sujets avec le temps. En sachant que plusieurs collections de documents telles que les revues, les courriers électroniques et les articles de presse, etc ; présentent tous un contenu évolutif (Blei & Lafferty, 2006). Une autre limite de LDA est son incapacité de modéliser la corrélation entre les sujets où elle est parfois très utile. Ces limites ainsi que d'autres cas particuliers ont provoqué l'émergence de plusieurs extensions de LDA.

3.3.1. Les Extensions de LDA

Il existe de nombreuses extensions de LDA pour correspondre à des applications particulières ou pour corriger une limitation spécifique du modèle. La Figure 8, présente une taxonomie de méthodes de modélisation de sujet basé sur LDA.

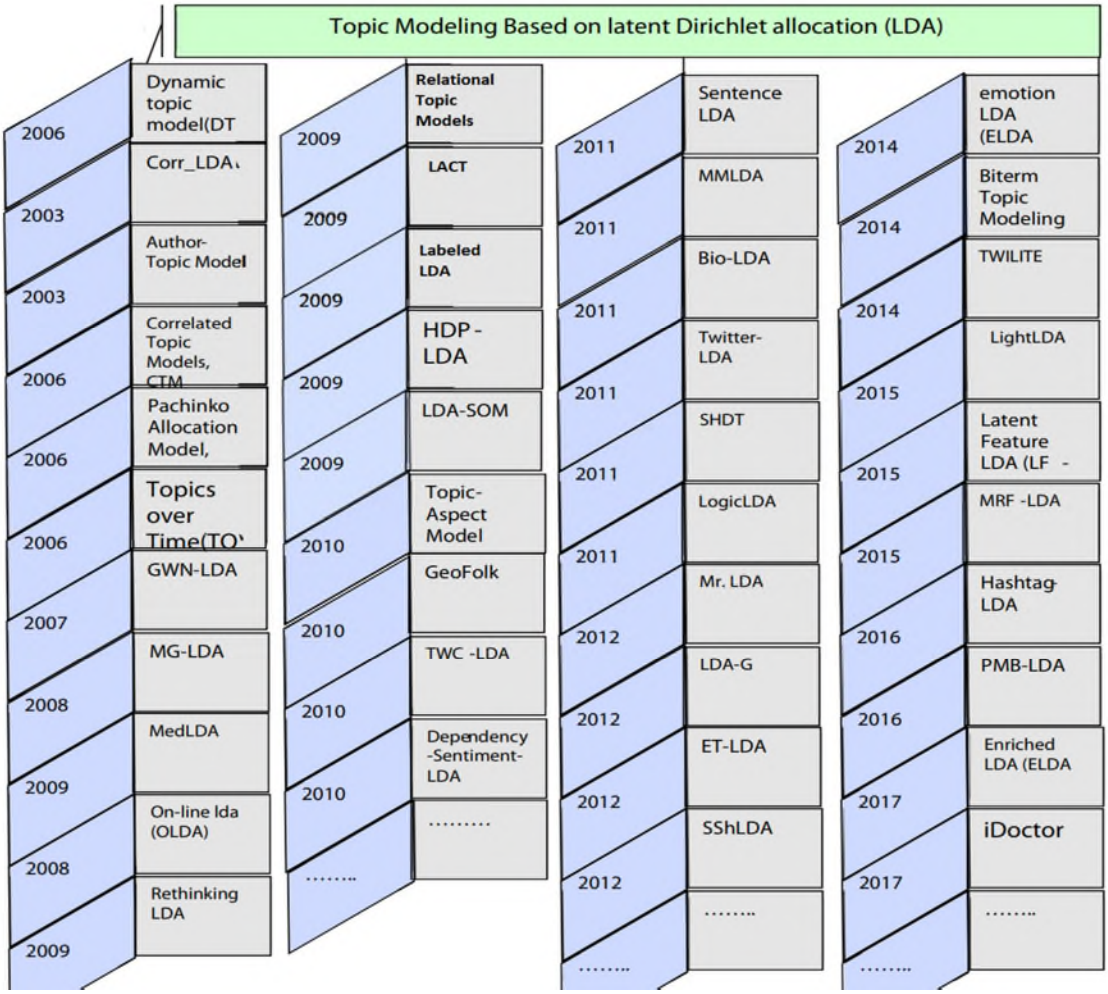


Figure 8: Une taxonomie des méthodes basées sur LDA et ses extensions, (Jelodar et al., 2019).

3.3.1.1. Structure du sujet

Comme nous l'avons déjà cité, l'une des limites du modèle LDA est son incapacité à modéliser la corrélation entre les sujets. Pour cette fin les auteurs dans (Blei et al., 2007) ont proposé le modèle de sujet corrélé (CTM), où le partitionnement du document est remplacé par une distribution logistique normale (loi logit-normale) pour paramétrer et apprendre la corrélation entre les sujets. Nous trouvons aussi pour ce même problème le modèle de réseau à thème latent (LTN) (Foulds et al., 2015).

3.3.1.2. Approches non paramétriques

Une autre limitation du modèle LDA est le nombre de sujets *K* qui doit être réglé manuellement. Plusieurs extensions non paramétriques de LDA apprennent de manière adaptative le nombre de sujets des documents. Par exemple dans le processus

hiérarchique de Dirichlet (*Hierchical Dirichlet Process*, HDP) (Teh et al., 2005), une liste de sujets est mise à jour de manière itérative en ajoutant ou en supprimant des sujets de la liste. Nous trouvons aussi d'autres structures plus complexes que les listes pour modéliser les sujets. Le processus de restaurant chinois imbriqué (nCRP) (Blei et al., 2010), et le processus de Dirichlet hiérarchique imbriqué (nHDP) (Paisley et al., 2014) consiste à apprendre une arborescence (non paramétrique) de sujets à partir de document. Dans cet arbre, le sujet racine est le plus générique tandis qu'un sujet feuille est très spécifique.

3.3.1.3. Sujets multi-échelles

La structure des documents n'est pas prise en charge dans le modèle LDA. Il existe une extension de LDA sous le nom de LDA multi-grains (MG-LDA) (Titov & McDonald, 2008), où deux échelles de sujets sont apprises. Une liste de sujets « généraux » décrivant un document entier et une autre liste de sujets « locaux » décrivant des phrases. Par conséquent, ce modèle prend en compte la structure des documents tout en conservant la représentation pratique du sac de mots pour les documents et des phrases.

3.3.1.4. Dynamic Topic Modeling (DTM)

Une autre approche de la modélisation de sujets appelée la modélisation dynamique de sujets. Dans cette approche le corpus de documents est divisé en segments séquentiels de telle sorte que l'hypothèse d'échangeabilité des documents est mieux envisagée dans ce modèle en supposant que seuls les documents de chaque segment sont échangeables. Par la suite, le DTM doit être appliqué sur le corpus segmenté pour permettre aux distributions de sujets d'évoluer d'un segment à l'autre ce qui donne un modèle hiérarchique. Ces types de modèle de sujets dynamiques procurent un aperçu qualitatif du contenu d'une grande collection de documents et fournissent également des modèles quantitatifs et prédictifs d'un corpus de documents séquentiels (Blei & Lafferty, 2006).

3.3.1.5. Labeled LDA (LLDA)

LLDA est un modèle graphique probabiliste basé sur LDA conçu par (Ramage et al., 2009). Il modélise un document dans un corpus comme une mixture de sujets et de mots générés par le sujet, et construit une correspondance un à un entre les sujets latents et les étiquettes, à partir de laquelle une distribution d'étiquette-terme (c'est-à-

dire terme-catégorie) pourrait être établie par un apprentissage, où une étiquette représente une classe.

LLDA suppose que chaque document a un ensemble d'étiquettes connues (par exemple les tags donnés par des utilisateurs). Le modèle est construit avec ces documents étiquetés en effectuant une correspondance entre les topics proposés par LDA classique et les étiquettes. Ceci permet à LLDA d'apprendre les correspondances terme-étiquette.

LLDA est un algorithme supervisé permettant l'application manuelle des étiquettes. Il n'utilise que les topics correspondants à l'ensemble des étiquettes. Dans ce cas, le nombre de topics n'est autre que le nombre des étiquettes uniques.

L'inconvénient majeur de cette version, est qu'elle nécessite des documents étiquetés au préalable.

3.3.1.6. Relational Topic Model (RTM)

Proposé dans (Chang & Blei, 2009), basé sur LDA, c'est un modèle de documents et des liens entre eux. Les liens expriment les différentes citations entre documents. Comme pour LDA, le document est d'abord exprimé en une distribution en topics, Les liens entre les documents sont ensuite modélisés comme variables binaires, une pour chaque paire de documents.

3.3.1.7. Author-Topic Model (ATM)

Proposé par (Rosen-Zvi et al., 2012), ce modèle est une extension de LDA. Il permet d'inclure les informations des auteurs et considère chaque auteur comme une distribution multinomiale à travers des termes. Il est mieux adapté dans les cas de documents rédigés par plus d'un auteur.

3.3.1.8. Twitter-LDA

Les paramètres d'origine de LDA, où chaque mot a une étiquette de sujet, peut ne pas fonctionner correctement avec twitter, cela revient à la forme courte et bruyante des tweets. De plus, un seul tweet est plus susceptible de concerner un seul sujet. Par conséquent twitter-LDA a été proposé pour résoudre ce problème. T-LDA aborde également la nature bruyante des tweets, où il capture les mots d'arrière-plan dans les tweets (W. X. Zhao et al., 2011).

3.3.1.9. Hashtag-LDA

C'est une approche de recommandation des hashtags personnalisés, introduite en fonction des informations d'actualités latentes dans les microblogs non étiquetés. Ce modèle renforce l'influence des hashtags sur la génération des sujets latent en modélisant à la fois les hashtags et les mots dans les microblogs (F. Zhao et al., 2016).

4. Principaux travaux sur la recherche d'expertise utilisant le topic modeling

L'introduction du *topic modeling* pour la recherche d'experts a été effectuée dans (Jing Zhang et al., 2008), où les auteurs ont discuté la limite de la non prise en compte des informations sémantiques des approches précédentes, et a proposé un modèle de mixture (MM) basé sur PLSA. MM a utilisé une couche de topics latents entre les documents, les auteurs et la requête.

(Daud et al., 2010) a proposé *Temporel-expert-topic* où il intègre les techniques du *topic modeling* et prend en considération la dynamicité et le changement dans le temps des topics. Ceci est réalisé pour le cas des conférences scientifiques. Il intègre simultanément l'information du temps et l'influence des conférences[6].

Dans (Zeng et al., 2010) les auteurs ont exploité le réseau des co-auteurs des documents pour proposer le *Co-author Network Topic* (CNT). Ce dernier peut apprendre les distributions de topics ainsi que les expertises des candidats à partir de grande collection de documents. L'effet des relations de co-auteurs a été étudié sur la modélisation des topics et des expertises. L'approche est notamment très utile lorsqu'un auteur cherche des collaborateurs (un encadreur, une équipe de recherche, etc.). L'approche a été évaluée sur le contenu DBLP de six conférences scientifiques.

Pour orienter une question aux experts adéquats en vue d'avoir les meilleures réponses dans les communautés de question/réponse, (Riahi et al., 2012) ont proposé de créer pour chaque utilisateur un profil en combinant les questions pour lesquelles il a donné de meilleures réponses (c.-à-d. ses réponses ont été choisies comme meilleures réponses à la question). Basé sur ce profil, des relations entre le candidat et de

[6] L'influence des conférences sont les informations liées à la conférence (issues de l'appel à communication par exemple), aux papiers acceptés et leurs auteurs dans le but d'améliorer les résultats de recherche d'experts liés aux thématiques de la conférence voulue.

nouvelles questions sont calculées en utilisant un nombre de méthodes : modèle de langue, TF-IDF, LDA et *Segmented Topic Model* (STM).

Les auteurs dans (Momtazi & Naumann, 2013) ont exploité ces techniques du *topic modeling* pour faire ressortir les topics d'expertise probables à partir d'une collection de documents utilisée comme indicateur d'expertise. LDA a été utilisée comme technique du *topic modeling*. Les sujets extraits présentent les connexions entre les candidats experts et les requêtes des utilisateurs. Dans une deuxième étape, les topics sont utilisés comme un pont pour trouver la probabilité de sélectionner chaque candidat pour une requête donnée. Les candidats sont ensuite classés en fonction de ces probabilités. Les expérimentations sont réalisées sur TREC. Les résultats sont jugés satisfaisants par rapport aux approches utilisant les techniques de la RI pour le classement des experts.

Les travaux dans (Kavitha et al., 2014) sont basés sur plusieurs éléments pour classer les experts, le graphe des citations, les informations du profil (nombre de publications dans le *topic* par exemple) et le *topic modeling* pour déduire les distributions des publications en *topics*, suivi d'un calcul de probabilités d'un candidat à être expert dans le topic en question (exprimé par une requête). L'algorithme *Lambda Rank* est utilisé pour le classement des experts.

(Xie et al., 2016) déjà cité dans la section des approches basées sur l'analyse des liens (section 3.2), a proposé un modèle de recherche d'expert contextuel spécifique au sujet (à la thématique) (TSCEFM : *Topic-Specific Contextual Expert Finding Model*). LDA a été utilisé pour l'extraction des topics. Le modèle d'apprentissage utilisé est Lambda.

Pour (Olieman et al., 2016), générer des profils thématiques d'utilisateurs consiste à identifier les sujets dans les contributions individuelles des utilisateurs, à agréger ces sujets en profils en fonction des interactions entre utilisateurs et contenu (par exemple rédiger, commenter, modifier, aimer). Une approche de généralisation de topics qui forme des groupes de sujets basés sur leur association avec des sujets plus larges dans la catégorie graphe de Wikipédia a été proposée.

(Neshati et al., 2017) ont repris l'idée de (Momtazi & Naumann, 2013) pour induire les *topics* d'expertise probables à partir des documents en utilisant LDA. Rappelons que les auteurs ont introduit le nouveau problème du *future expert finding*

et prédisent le classement des experts dans le temps futur en proposant un cadre d'apprentissage.

Les auteurs dans (Liang, 2018) ont étudié le problème du profilage dynamique des utilisateurs dans le cadre du flux de textes courts. Un modèle de sujet de suivi des intérêts collaboratifs a été proposé. Les intérêts des utilisateurs dans les flux ont été modélisés en fonction de sujets latents. Par conséquent, les intérêts dynamiques de chaque utilisateur u à la période de temps t peuvent être représentés par une distribution multinomiale $\theta_{t,u}$ à travers des sujets, $\theta_{t,u}$ étant une distribution de Dirichlet. La proposition a été comparée à certains travaux de l'état de l'art à savoir TF-IDF, LDA, ATM, DTM, *Predictive Language Model* (PLM), etc. Les résultats montrent que le modèle proposé est capable de profiler les intérêts dynamiques des utilisateurs au fil du temps pour les flux de textes courts.

Un autre problème qui a nécessité l'utilisation du *topic modeling* est la différence de vocabulaire entre les requêtes et les documents (tags associés aux questions, et contenus des réponses pour le cas de *Stack Overflow*, par exemple). Ce qui nécessite parfois la proposition des méthodes de traduction. (Dehghan & Abin, 2019) ont proposé un modèle de traduction basé sur le regroupement *clustering* de termes dans deux espaces : espace de requêtes et espace de co-occurrences. Les documents sont d'abord regroupés en topics en utilisant LDA, puis une matrice termes-requêtes (W-Q) est créée ayant comme valeurs, les co-occurrences des termes w et les tags q de la requête dans l'ensemble de documents. L'espace de co-occurrence est créé par la matrice co-occurrence-terme (C-W) ayant comme valeurs les co-occurrences simultanées des termes dans l'ensemble de documents. Les tests sont réalisés sur *Stack Overflow*.

(Nobari et al., 2020) ont proposé une méthode pour traduire les domaines de compétence (skill area) en leurs principaux mots. Les tags donnés associés aux questions sont considérés comme les *skill area*. Avant de proposer leur approche basée sur le plongement de mots (*word embedding*) les auteurs ont repris le modèle de (Momtazi & Naumann, 2013) basé sur le topic modeling. En effet, les sujets extraits sont utilisés comme pont pour connecter les candidats à une requête donnée.

(Rampisela & Yulianti, 2020) ont également utilisé le *word embedding* pour l'expansion de requêtes combiné à BM25, mesure de pondération de termes en RI, pour améliorer la recherche d'experts. Le dataset utilisé est en indonésien. La source

d'évidence utilisée est l'ensemble des résumés de thèses des étudiants du département informatique de l'université d'Indonésie.

5. Conclusion

L'expertise d'un candidat est liée à un domaine particulier, une thématique donnée. Plusieurs travaux de la littérature de la recherche d'expertise exploitent les techniques du *topic modeling* pour la découverte de thématiques dont un candidat est susceptible d'être expert.

Nous avons présenté dans ce chapitre les différentes notions liées à la modélisation thématique. Nous avons cité les principales techniques utilisées en nous focalisant essentiellement sur LDA que nous estimons être le modèle le plus performant, le plus utilisé et cité dans la littérature. Nous avons mentionné un ensemble de travaux qui ont utilisé la modélisation thématique pour rechercher l'expertise, soit en définissant le profil de l'utilisateur qui sera l'indicateur de l'expertise, soit en posant un classement des candidats experts.

Le prochain chapitre est dédié aux synthèses et discussion de l'état de l'art concernant les différents travaux cités dans les deux chapitres précédents.

Chapitre III
Synthèse et Discussion des Travaux sur la Recherche d'Expertise

1. Introduction

L'objectif principal de ce chapitre est de revoir brièvement les différentes catégories de travaux présentées dans cette première partie de la thèse. Discuter également les résultats obtenus, les avantages et inconvénients des solutions proposées. Dans les travaux proposés, les auteurs ont soulevé la difficulté de l'estimation de l'expertise. La dynamicité de celle-ci et la diversité des sources d'évidence font de son estimation une tâche ardue. Nous allons donc discuter les défis qui restent à relever car l'estimation de l'expertise d'une manière précise et complètement efficace est loin d'être accomplie.

Dans le chapitre 1, nous avons divisé les travaux en deux parties, ceux de la recherche d'expert et ceux du profilage d'expert. Nous avons aussi présenté un ensemble de travaux utilisant le topic modeling dans le chapitre 2. Ces derniers sont liés soit à la recherche ou le profilage d'expert. Nous allons donc regrouper et récapituler l'ensemble des travaux présentés.

Par ailleurs, l'une des étapes les plus importantes d'une recherche d'expertise, est la sélection des sources d'évidence. En effet, tout ce qui est produit par l'utilisateur peut servir à définir ses domaines d'expertise, ce qui explique la diversité des sources utilisées. Nous allons discuter les sources d'évidence considérées dans les différents travaux.

Dans ce qui suit, un récapitulatif regroupant tous les travaux sera présenté. Il sera suivi de discussions des travaux du *finding/profiling* séparément. Par la suite, une discussion des sources d'évidence est également présentée. Nous clôturons ce chapitre par une conclusion.

2. Synthèse des travaux

Nous allons dans cette section récapituler les travaux cités avec la description de certains paramètres utilisés dans chaque travail, à savoir : la tâche considérée (*finding/profiling*), la technique utilisée, la source d'évidence adoptée, intégration ou non du *topic modeling* et le domaine de test ou d'application. La catégorie (RI/Analyse des liens) du travail est déduite à partir de la source d'évidence adoptée et la technique utilisée. En effet, les travaux utilisant les documents rédigés utilisent certainement les techniques de la RI, ceux utilisant les emails et les activités sociales peuvent exploiter les deux types de techniques. Nous résumons ceci dans le Tableau 1.

Tâche	Travaux	Sources d'évidence utilisées				Technique utilisée	Topic modeling	Domaine d'application
		Documents	Mails	Activité Sociale	Analyse des liens			
Expert Finding	(Balog et al., 2006), (Balog et al., 2009), (Balog et al., 2012)	X				LM	Non	Organisation
	(Macdonald & Ounis, 2006)	X				Votes	Non	TREC Enterprise/ TREC W3C
	(Petkova & Croft, 2008)	X				LM	Non	Organisation
	(Jing Zhang et al., 2008)	X				PLSA	Oui	Communauté académique
	(Daud et al., 2010)	X				LDA	Oui	DBLP
	(Moreira et al., 2011)	X				L2R4IR (SVM)	Non	Publications Academiques (computer science)
	(Neshati et al., 2017)			(Réponses)		L2R4IR	Oui	Communauté Web
	(Gharebagh et al., 2018)			(Réponses)		LM	Non	Communauté Web
	(Al-Barakati & Daud, 2018)	X				LM	Non	DBLP
	(Campbell et al., 2003)		X		X	Unsupervised clustering/ HITS	Non	Organisation
	(Fu et al., 2007)		X	(contenu de page web)	X	Propagation	Non	TREC Enterprise
	(Rodriguez & Bollen, 2008)	X			X	PSO	Non	ACM/IEEE conference

	(Jun Zhang et al., 2007)				X	PageRank/ HITS	Non	Communauté Web
	(Kardan et al., 2011), (Jiao et al., 2009; Wang et al., 2013)				X	Modèle vectoriel/PageRank	Non	Communauté Web
	(Carchiolo et al., 2015)			(reviews)	X	Random walk	Non	Epinion
	(Xie et al., 2016)				X	HITS/CAM /SVM	Oui	Communauté Web
	(Amato et al., 2019)				X	HITS	Non	MSN
	(Zeng et al., 2010)	X			Coauteur-document-network	MRF	Oui	DBLP
	(Momtazi & Naumann, 2013)	X				LDA	Oui	TREC Enterprise
	(Kavitha et al., 2014)	X				LM/PageRank/LDA	Oui	Communauté académique
	(Rampisela & Yulianti, 2020)	X				Word embedding/ BM25	Non	Communauté académique
Expert Profiling	(Balog & De Rijke, 2007)	X				LM/TF-IDF	Non	TREC Enterprise
	(Budura et al., 2009)			(Tags)		LM	Non	Communauté Web
	(Serdyukov et al., 2011)	X		(Tags automatiques)		LM/ Logistic regression	Non	Organisation
	(Fazel-Zarandi & Fox, 2011)			(activité journalière online ou offline+ auto-déclarations)		Ontologies	Non	Organisation

	(Riahi et al., 2012)			(Questions-réponses)		LM/LDA/STM	Oui	Communauté Web
	(Berendsen et al., 2013)	X				LM	Non	Organisation
	(Rybak et al., 2014)	X				LM	Non	DBLP
	(Y. Fang & Godavarthy, 2014)	X				Predictive LM	Non	Communauté académique
	(Ribeiro et al., 2015)	X				L2R	Non	Communauté académique
	(Olieman et al., 2016)			(comments, like…)	X	ERD	Oui	Wikipédia
	(Liang, 2018)			(Tweets)		LDA	Oui	Communauté Web
	(Faggioli et al., 2019)			(tags)		Game theory	Oui	Communauté Web
	(Dehghan et al., 2019)	X				LDA/ Clustering	Oui	Communauté Web
	(Nobari et al., 2020)			(Réponses)		LDA/ Word embedding	Oui	Communauté Web

Tableau 1: Récapitulatif des travaux existants

(LM): Language Model, (L2R4IR): Learning to rank for Information Retrieval, (CAM): context-aware model, (MSN) Multimedia social network, (MRF) : Markov Random Field, (tags automatiques: veut dire que le système suggère un ensemble de tags à l'employé, (Logistic regression) : utilisé pour classer les tags. (PSO): Particle Swarm Optimization. (STM) Segmented Topic Model. (ERD) Entity Recognition and Disambiguation.

3. Discussion

3.1. Travaux liés à l'expert finding

Les travaux de recherche sur *l'expert finding* sont divisés en deux grandes catégories : ceux basés sur les techniques de la RI où on peut trouver les approches basées sur le profil et celles basées sur les documents ; et ceux basés sur l'analyse des liens. Comme nous pouvons voir sur le Tableau 1, certaines approches pouvant être appelées hybrides du moment qu'elles combinent des techniques de la RI et l'analyse des liens (cas utilisant par exemple les documents rédigés ou le contenu des emails ainsi que les connections résultantes).

L'objectif premier de ces approches était d'automatiser la tâche de recherche d'experts au sein de l'entreprise, et de trouver les correspondances entre les candidats et les documents rédigés. Par la suite, l'objectif est étendu vers le web communautaire. Les techniques de la RI principalement utilisées sont les modèles de langue, les modèles probabilistes et vectoriels ainsi que le *topic modeling* tels que : LDA, LSI et PLSA.

D'un autre côté, les approches basées sur l'analyse des liens se basent, dans la majorité des travaux, sur des algorithmes adaptés à partir d'anciens algorithmes basés sur la théorie des graphes, tels que *PageRank* et *Hits*. Ce qui est évident, est que généralement, le lien seul ne peut être un indicateur d'expertise s'il n'est pas combiné au contenu.

Les sources d'évidence utilisées dans la majorité des travaux sont les documents rédigés par les candidats. Cet indicateur présente cependant un inconvénient majeur : les noms d'experts peuvent être ambigus, incomplets et appartenir à des personnes différentes. Par ailleurs, un autre problème peut survenir lorsque les documents sont longs et que chaque auteur n'est concerné que par une partie spécifique du document.

Les travaux utilisant le *topic modeling* dans la recherche d'experts, visent à intégrer l'aspect sémantique et découvrir les distributions de documents à travers les topics, et catégoriser les termes composant les documents dans les topics découverts.

3.2. Travaux liés à l'expert profiling

Dans cette catégorie d'approches, l'objectif est de définir l'ensemble des domaines d'expertise du candidat avec son niveau d'expertise dans chaque domaine. Ces domaines sont appelés parfois *topics* ou *knowledge area* (Ka). Les approches proposées sont majoritairement basées sur la RI et le topic modeling. Les sources d'évidence utilisées sont basées sur les documents rédigés, les emails, les réponses aux questions, etc.

Ces dernières années, un ensemble de travaux se basent sur l'activité sociale du candidat. Nous avons cité des travaux qui se basent sur les activités du tagging social pour estimer l'expertise. En effet, les tags sont exploités en tant que topics d'expertise cités par le candidat, soit pour décrire ses compétences, ou soit dans certains cas pour décrire une source particulière telle qu'une question (cas de *Stack Overflow*), ou même pour décrire une autre personne. Cependant, ils ne sont pas exploités comme un indicateur social de l'expertise du candidat.

L'aspect sémantique a été pris en considération après l'intégration des techniques du *topic modeling*. L'objectif est de découvrir les *topics* considérés dans les documents. Les modèles du *topic modeling* visent à établir des distributions des documents sur les *topics* et les *topics* sur les termes. Ces techniques sont utilisées par les approches de la recherche d'experts et celles du profilage.

3.3. Sources d'évidence considérées

L'information produite par les candidats, telle qu'elle soit, peut être considérée en tant qu'une source d'évidence de leurs intérêts et leurs expertises : les collections hétérogènes de documents incluant les emails, les rapports techniques, les publications scientifiques, les pages web. Mais également les liens construits à partir de leurs activités permettent de définir leur influence et importance sociale.

Comme il a été mentionné, la source d'évidence la plus considérée dans ces travaux de recherche est l'ensemble des documents rédigés par les candidats (rapports techniques, publications scientifiques, cours…). Le problème imminent est l'association document-candidat. Ce problème n'a pas reçu beaucoup d'importance dans la communauté de recherche. Toutefois, un nombre de techniques a été appliqué pour estimer la force de l'association : (1) les techniques basées sur un ensemble de

documents (*set-based*), (Macdonald & Ounis, 2006) où le candidat est associé aux documents comportant son nom ou son adresse email ; (2) les techniques basées sur la fréquence (frequency-based), (Balog et al., 2006; H. Fang & Zhai, 2007; Petkova & Croft, 2008) où la force de l'association candidat-document est proportionnelle au nombre de fois que l'identifiant du candidat (nom ou adresse email) se produit dans le document.

Les techniques proposées se basent donc sur le nom ou l'adresse email du candidat, leurs apparitions dans un ensemble de documents, ou leurs fréquences d'apparition dans le même document. Toutefois, le problème d'ambiguïté des noms persiste toujours. L'utilisation d'une adresse email n'est pas toujours possible (il existe des documents scientifiques sans adresse email des auteurs). La fréquence d'apparition peut être traduite par les citations, certains types de documents peuvent ne pas contenir de citations (un cours de mathématiques présentant des éléments de base peut ne pas contenir de citations). Sans oublier le cas des documents longs, où le candidat peut être l'auteur d'une partie uniquement (un seul topic par exemple), et donc ses topics d'expertise ne convergent pas avec les autres topics du document auquel il est associé.

Malgré les limites soulevées, l'utilisation des documents rédigés comme sources d'évidence est certainement un choix pertinent vu que l'expertise d'un candidat se traduit généralement par ce qu'il écrit d'une manière professionnelle. Ceci est fortement utilisé dans des cas bien particuliers tels que : la recherche scientifique, l'attribution des relecteurs, la constitution des comités de programme, etc. Cependant, si nous élargissons ces cas d'utilisation au web communautaire, l'exploitation des documents rédigés n'est pas toujours possible (ce n'est pas tous les utilisateurs qui possèdent une production scientifique ou technique). De plus, le problème d'accès aux données privées peut être relevé pour le cas des emails. En effet, il n'est pas toujours permis d'accéder aux contenus des emails sauf dans des cas précis d'études avec consentement des candidats.

Puisque nous nous intéressons au web communautaire, nos propositions seront orientées vers l'utilisation des activités sociales des candidats comme source d'évidence. Nous exploiterons les activités liées au Tagging social ainsi que les tweets.

L'utilisation des tags ou des tweets comme indicateur d'expertise, peut nous éviter le problème d'association candidat-document. Un ensemble important de tags

ou de tweets pourra remplacer les documents, ceci est surtout favorisé du moment que le document est considéré dans ces cas de travaux de recherche comme un sac de mots (*bag of words*).

4. Conclusion

Nous avons présenté les travaux relatifs à l'expertise. Nous nous sommes focalisés sur les travaux considérant les documents rédigés comme source d'évidence (ils sont majoritaires) et nous avons mis en avant les problèmes rencontrés dans ce cas.

Nous sommes particulièrement intéressés par les approches exploitant l'activité sociale de l'utilisateur afin de déterminer son expertise, notamment l'activité du tagging social. Les travaux ayant exploité les tags, sont en majorité des études sur l'importance d'utiliser ces tags pour déterminer l'expertise, les relations entre tag et topic d'expertise. Cependant il n'y a pas eu de réelles exploitations des activités du tagging en tant qu'indicateurs d'expertise.

La problématique de la recherche d'expertise demeure un problème majeur de la RI. L'apparition et la profusion du web communautaire et des activités sociales ont révolutionné la recherche d'expertise par de nouveaux défis en passant de méthodes traditionnelles à d'autres modernes. Dans ce sens, nous allons opter pour des solutions basées sur le tagging, pour améliorer et généraliser le processus de recherche d'expertise, qui jusqu'ici, est limité soit aux cas académiques et entreprise, soit au cas de CQA. Une généralisation de la recherche d'expertise aux différents cas des communautés du web (différents réseaux sociaux, *bookmarking* social, etc.) seront d'un grand intérêt pour tout utilisateur désirant rechercher une compétence particulière pour un besoin quelconque.

En effet, nous avons constaté que le tagging n'est pas suffisamment exploité en tant que source d'évidence, bien que les méthodes qui l'ont intégrée, aient clairement démontré sa contribution à l'amélioration des performances des expertises. Nous souhaitons donc une estimation de l'expertise de l'utilisateur en fonction de son activité du tagging, car nous pensons qu'un tag peut exprimer le niveau de connaissances de l'utilisateur dans le sujet de la ressource tagguée. Une nouvelle approche sera proposée, pour permettre de mieux utiliser les tags afin d'estimer l'expertise de l'utilisateur en exploitant les tags et les liens sémantiques entre eux.

Cet aspect sémantique est pris en considération par le *topic modeling*. Les différentes techniques du *topic modeling* existant dans la littérature ont pour objectif d'analyser et catégoriser le texte pour pouvoir en déduire les topics traités. Utiliser le *topic modeling* pour modéliser un expert a été rarement réalisé (Riahi et al., 2012). Pour notre cas, l'exploitation du *topic modeling* serait dans le sens du profilage des candidats experts, et non pas la catégorisation des documents.

Nous portons également un intérêt particulier à Twitter, étant un réseau social très connu et très utilisé. Nous voulons connaître les thématiques d'intérêt d'un utilisateur en exploitant ses tweets. La notion du profil thématique n'a pas été beaucoup abordée dans la littérature.

Nous reviendrons également au cas des femmes artisans pour étudier ce que peut apporter le tagging social pour pouvoir améliorer l'échange informationnel entre elles, ainsi que faciliter le repérage de compétences particulières.

Dans la seconde partie de ce mémoire, nous allons entamer la présentation de nos propositions concernant la recherche et le profilage de l'expert.

DEUXIÈME PARTIE
CONTRIBUTIONS

Introduction

Dans la première partie nous avons exploré le domaine de la recherche d'expertise en présentant les différentes notions qui lui sont liées. Les travaux du domaine les plus importants ont été présentés dans les deux premiers chapitres puis synthétisés dans le troisième chapitre. À travers cette étude, nous avons mis en exergue la difficulté et la complexité de la recherche et du profilage de l'expert vue la sensibilité et la dynamicité de la notion d'expertise ainsi que la diversité des sources indicatrices. L'abondance de l'information sociale pouvant être considérée comme évocatrice de l'expertise de l'utilisateur rajoute également une couche à cette complexité. Nous avons néanmoins constaté que parmi ces sources d'information, l'activité de tagging n'a pas été suffisamment explorée dans ce domaine. Or, nous pensons qu'un tag affecté par un utilisateur à une ressource reflète la perception et l'appréciation de l'utilisateur de la ressource en question et que la qualité sémantique de ce tag pourrait nous renseigner sur l'expertise de l'utilisateur dans le domaine de la ressource en question. Il serait donc intéressant d'exploiter cette information afin de déterminer les domaines dans lesquels l'utilisateur est expert ou de manière duale, les experts d'un domaine particulier.

Dans nos travaux que nous présentons dans cette partie, nous exploitons le tagging comme source d'information pour l'évaluation de l'expertise et ceci en proposant premièrement un nouveau modèle d'évaluation de l'expertise et deuxièmement un modèle de recherche d'expert.

Le chapitre 4 est consacré à la présentation de notre contribution principale qui consiste en un nouveau modèle de profilage de l'expert pour l'estimation de son expertise basé sur ses activités du tagging social (Kichou et al., 2020). Nous présentons ainsi dans ce chapitre les détails du modèle ainsi que la méthodologie de construction de ses différentes parties. Ce modèle est doublement évalué dans cette thèse : la première évaluation est présentée dans le chapitre 5 qui est une validation académique basée sur des expérimentations menées sur les Datasets *StackOverflow*, *Delicious* et *Twitter* avec des métriques issues de la recherche d'information. Une comparaison des résultats obtenus par notre modèle avec un modèle de base et quelques modèles de la littérature est également présentée avec une discussion approfondie. La seconde

évaluation, présentée dans le chapitre 6, consiste en une mise en pratique du modèle sur un cas réel. Il s'agit d'une instanciation du modèle proposé et son application dans le cas du projet de coopération Algéro-Tunisien consistant en un modèle de recommandation des femmes artisans (Kichou & Meziane, 2015). Dans cette phase nous démontrons par un cas pratique l'efficacité du modèle proposé et son applicabilité à un cas réel et ceci via la proposition d'une approche de recommandation des femmes artisans basée sur les opérations du tagging (Kichou et al., 2016).

Chapitre IV
Un Modèle d'Estimation de l'Expertise basé sur le Tagging Social

1. Introduction

Les communautés du Web sont aujourd'hui l'une des sources de recherche d'experts les plus utilisées par les organisations. Les réseaux sociaux, avec l'énorme volume de données générées par les utilisateurs, sont considérés comme un terrain fructueux pour identifier les compétences et les expertises. Les données sont disponibles sous forme de publication (*posts*), de discussions, de tags, de commentaires, etc. Les informations de l'utilisateur sont transférables sous forme sémantique car il existe des vocabulaires largement utilisés et acceptés pour ces domaines (De Vocht et al., 2017). La recherche de l'expertise humaine a récemment attiré une attention considérable. L'expertise humaine étant difficile à formaliser vu l'absence de règles génériques de sa formalisation (Cifariello et al., 2019), son évaluation est perçue par les chercheurs comme une tâche complexe. Il s'agit d'un domaine de recherche actif dans de nombreuses spécialités telles que l'intelligence artificielle, la gestion des connaissances, le travail coopératif assisté par ordinateur, et autres (Al-Taie et al., 2018).

Rappelons que le but principal de la recherche d'experts est de déterminer et de classer les personnes susceptibles d'être expertes dans un domaine spécifique, alors que l'objectif du profilage d'expert est de déterminer les domaines de compétences dans lesquels une personne est spécialisée (Dehghan et al., 2019). Les systèmes de recherche d'experts utilisent des données fournies implicitement ou explicitement sur l'expertise de l'utilisateur pour identifier les experts appropriés. Cependant, créer des profils d'utilisateurs à base de leurs contenus personnels et leurs commentaires est un défi. De plus, les auto-déclarations de compétences peuvent être incorrectes, inexactes ou insuffisantes.

Une tâche alternative, construite sur le même principe de calcul des associations personnes-sujets, est le profilage expert, dans lequel les systèmes doivent renvoyer une liste de sujets qu'une personne connaît bien ou maitrise (Balog & De Rijke, 2007; Berendsen et al., 2013).

D'importantes sources d'informations sont généralement utilisées pour identifier le niveau d'expertise de l'utilisateur. Les sources de données liées à l'expertise sont souvent sous forme de documents hétérogènes tels que les publications du candidat, les rapports techniques, les courriels et les pages Web. Elles sont extraites à partir des réseaux sociaux (Wang et al., 2013) et des documents (Balog et al., 2006).

Les activités du tagging social ont été intégrées dans plusieurs travaux en tant qu'informations liées à l'expertise des candidats (Budura et al., 2009; Serdyukov et al., 2011; Yang & Manandhar, 2014; Zhu et al., 2011). Le tagging social est apparu dans le web social depuis quelques années, comme support à l'organisation de ressources partagées en permettant aux utilisateurs de catégoriser et de trouver ces ressources. Les données sociales peuvent également être un support pour améliorer la recherche personnalisée (Zhou et al., 2016), en modélisant le profil utilisateur à partir du tagging social. Ce dernier est reconnu pour son potentiel à exploiter la production collaborative d'informations qui prennent en charge un large éventail de mécanismes tels que la recherche sociale (Brusilovsky et al., 2018; Xie et al., 2016), l'extraction de profils (Kasper et al., 2017; Kichou et al., 2013; Kichou & Meziane, 2015) et la recommandation (BellogíN et al., 2013; Font et al., 2013; Kichou et al., 2016). Nous allons ainsi, dans nos travaux, proposer des modèles permettant l'exploitation du tagging dans le profiling et la recherche d'experts.

2. Motivations

Les activités du tagging social des utilisateurs ont été utilisées comme des éléments liés aux domaines d'expertise dans les systèmes de recherche d'experts et ont montré une amélioration considérable des performances de recherche d'experts. Ceci est principalement dû au fait que la description d'un expert n'est pas liée à la ressource elle-même (document, page web, question…), mais plutôt à l'avis de l'expert sur cette ressource via des tags. D'autant plus que cela permet à l'utilisateur d'exprimer librement ses domaines d'expertise et même sans le savoir. Cependant, une

information aussi importante n'est pas prise à sa juste valeur, car les tags ne sont pas considérés comme de vrais indicateurs d'expertise et leur aspect sémantique est complètement négligé. En outre, dans les cas considérant l'activité du tagging, les utilisateurs sont amenés à choisir des tags dans une liste générique prédéfinie (tagging automatique) ce qui fait perdre au tagging social sa première force à savoir la liberté. Dans notre cas, nous nous intéressons à la manière d'améliorer le processus de recherche d'experts et de profilage en intégrant les activités du tagging social d'une nouvelle manière.

L'apport principal de notre approche est double : d'une part, l'exploitation de l'activité du tagging social des utilisateurs dans la définition du profil expert, et d'autre part, l'utilisation des caractéristiques sémantiques et taxonomiques des tags dans l'estimation de l'expertise de l'utilisateur. En plus de définir une nouvelle façon de modéliser le profil expert, l'utilisation de la profondeur du tag fourni par les utilisateurs permet de mesurer leur niveau de connaissance dans les thématiques liées aux ressources correspondantes. Par ailleurs, découvrir les thématiques et distribuer les tags sur les sujets découverts en exploitant un aspect sémantique *latent* permet d'estimer une expertise utilisateur différente pour chaque sujet au lieu d'une seule expertise. À notre connaissance, cette solution n'a pas été proposée auparavant.

Dans ce qui suit, nous allons présenter un nouveau modèle d'estimation de l'expertise basé sur l'exploitation de l'activité du tagging social. La première partie est consacrée à la présentation du modèle de profilage et la seconde partie au modèle de recherche d'experts.

3. Principe général

Les systèmes du tagging social permettent aux utilisateurs de donner un avis sur une ressource en utilisant des tags qui sont une expression succincte. Un utilisateur marque (taggue) uniquement les ressources qui l'intéressent. Cette action de tagging exprime, dans une certaine mesure, la propre perception des ressources par l'utilisateur. En d'autres termes, le tag d'un utilisateur peut exprimer le niveau de connaissances de cet utilisateur dans les thématiques de la ressource considérée. Il peut donc être exploité pour estimer l'expertise de l'utilisateur. Nous pensons que l'utilisation des tags peut améliorer considérablement la recherche d'expertise en

permettant une estimation du niveau de connaissance d'un utilisateur indépendamment de la ressource elle-même. En revanche, les tags utilisés par les utilisateurs ont des liens sémantiques permettant leur regroupement en catégories. Chaque sujet représente une catégorie d'expertise de l'utilisateur, il serait donc intéressant de considérer des sujets (*topics*) plutôt que des tags individuels. La structure générale de la proposition est illustrée dans la *Figure 9* ci-dessous.

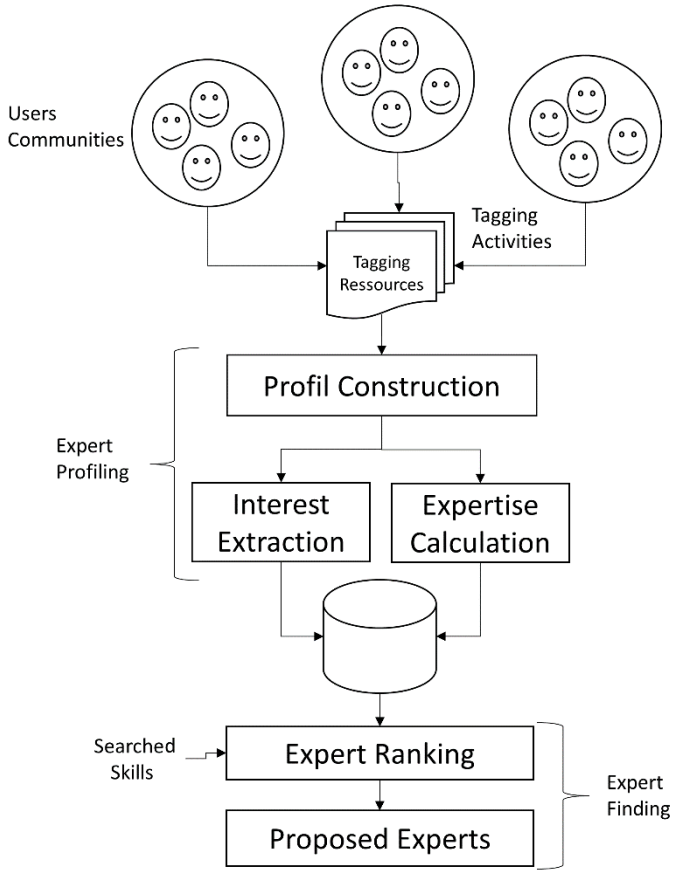


Figure 9: Principe général de l'approche.

Par conséquent, une nouvelle approche est présentée pour permettre d'exploiter ces informations afin de quantifier l'expertise de l'utilisateur quel que soit le type de ressource (voir la Figure 10). Pour cela, nous proposons tout d'abord un modèle du profil expert multidimensionnel permettant une description plus significative de l'activité sociale de l'expert et ce en considérant des sujets plutôt que des tags individuels. Ensuite, nous présentons la manière de construire les différentes dimensions du profil en nous concentrant sur notre approche d'estimation d'expertise statistique. Formellement, notre but consiste à définir la fonction f qui permet

l'estimation de l'expertise *Exp* de l'utilisateur u en fonction de son ensemble de tags T de la manière suivante :

$$Exp_u = f(T_u) \qquad (10)$$

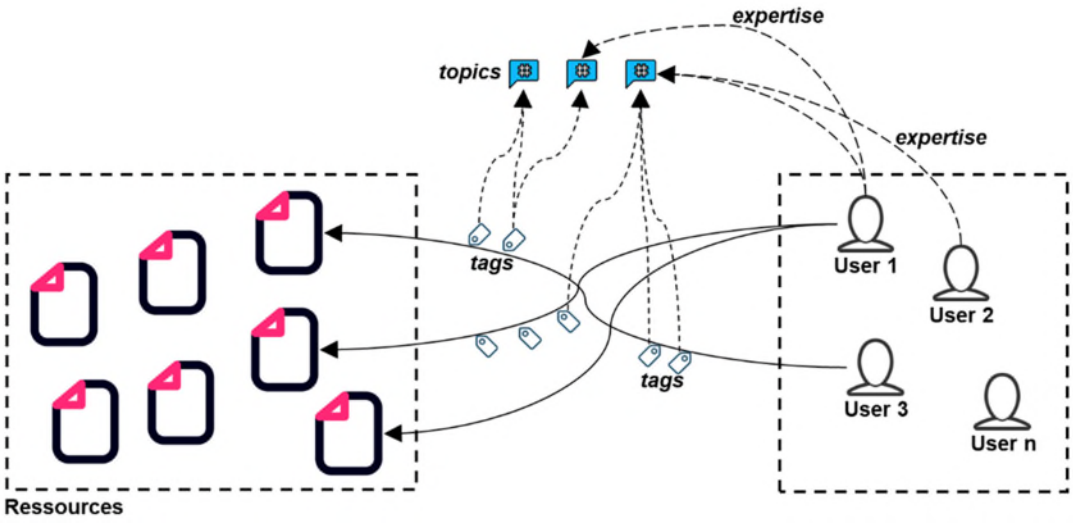


Figure 10: Exploitation des tags dans l'estimation de l'expertise.

Comme illustré dans la *Figure 10*, les tags attribués par l'utilisateur pour la description d'une ou plusieurs ressources sont catégorisés en topics. Les expertises de l'utilisateur sont calculées pour chaque topic.

4. Notation et terminologie

Avant de présenter notre approche, il est important de définir quelques concepts de base.

Définition 1. Un Tag, ou étiquette, est un mot clé librement choisi par un utilisateur pour décrire un objet dans le web (document texte, image, fragment d'un document, question). Il est attribué à ces ressources partagées et considéré comme une métadonnée permettant de décrire un *élément et de le retrouver par navigation ou recherche***.**

En se basant sur la communauté du tagging dans Delicious[7], (Golder & Huberman, 2005) spécifie sept fonctions qu'un tag peut avoir:

[7] https://del.icio.us/

- Identifier de quoi l'objet s'agit-il, son thème ou sujet (vacances, hiver…) ;
- Un tag peut identifier ce qu'est l'objet lui-même (une photo, un blog…) ;
- Identifier à qui l'objet appartient ;
- Description ou détails de tags existants : certains tags n'ont aucune signification seuls. Ils n'ont de sens que quand ils sont associés à d'autres tags. Leur fonction est donc d'apporter plus de détails ou de description à des tags existants. Tel est souvent le cas avec les nombres tel que le 10 de l'expression *top 10*.
- Identifier des qualités ou des caractéristiques du contenu : c'est le fait de dire que tel contenu est comique, *funny* ou horrible… ;
- Auto-référence du tag : dans ce cas, le tag illustre une relation entre l'utilisateur et le contenu. C'est le cas des tags commençant par mon ou *my* : *myphoto*, mon_enfant… ;
- Aide-mémoire : il s'agit de planifier une tâche donnée, à lire, à revoir….

Définition 2. La Folksonomie *F* est un ensemble de données générées par des systèmes de tagging social qui permettent aux utilisateurs *U* de décrire des ressources web partagées *R* avec des tags *T*. La folksonomie peut être formellement représentée comme : $F = (U, R, T, A)$, où $A = \{(u, r, t) \in U \times R \times T\}$ un ensemble de tuples reliant l'utilisateur u à la ressource r via le tag utilisé t, *(Golder & Huberman, 2005)* .

Définition 3. La Personomie concerne un utilisateur donné u, défini comme $P_u = (R_u, T_u, A_u)$, où A_u est une projection des tuples de *A* sur *T* et *R* pour l'utilisateur *u*. $T_u = \{t \mid (t, r) \in A_u\}$ est appelé le vocabulaire des tags de l'utilisateur et $R_u = \{r \mid (t, r) \in A_u\}$ l'espace d'activité des utilisateurs qui est composé de toutes les ressources que l'utilisateur a tagguées, *(Broudoux, 2006)*.

Définition 4. Le sujet (un thème, thématique ou topic) représente sur quoi s'exerce l'activité de l'homme (sujet d'un discours, d'un livre…). Chercher un sujet dans un document revient à poser la question suivante : de quoi ce document s'agit-il ? Dans notre contexte, un sujet $\vec{s}$ représente un vecteur de tags partageant la même sémantique avec des degrés différents.

Un tag peut décrire plusieurs sujets à la fois, son degré de description est quantifié comme la probabilité que ce tag décrive le sujet considéré. Ainsi, un sujet est un vecteur de tags pondéré défini comme suit :

$$\vec{s} = \{(t, p_s(t)) | p_s(t) \ \textit{est la probabilité que t décrive le sujet s}\}$$

L'ensemble des sujets spécifiques à un utilisateur donné noté S_u est appelé ensemble de sujets de l'expert et représente un ensemble de vecteurs tel que chaque vecteur représente un seul sujet. S_u est construit à partir des relations existantes entre les tags de l'expert dans A_u. Ces relations sémantiques latentes sont considérées par LDA (et les autres techniques de *topic modeling*) supposant que les termes similaires, ou très proches dans le sens, ont tendance à se produire ensemble de manière fréquente.

5. Modélisation du profil de l'expert

Dans le but de repérer un expert il est important de construire préalablement une description suffisamment complète de l'utilisateur pour pouvoir déceler son expertise dans un domaine donné. Dans notre approche, nous proposons un profil utilisateur basé sur son activité sociale, notamment son activité de tagging. Nous estimons également qu'il est plus intéressant d'aller au-delà des simples tags en exploitant les liens sémantiques entre ces derniers ultérieurement afin de construire des sujets. Ainsi, l'expertise de l'utilisateur pour chaque sujet est déterminée et intégrée à son profil.

5.1. Modèle du profil expert

Un profil de l'expert est une structure bidimensionnelle composée d'une dimension sociale et d'une dimension thématique (topicale).

Définition 5. La dimension sociale *Soc* d'un candidat expert représente une synthèse de son activité sociale. Elle se compose d'un ensemble de tags pondérés. Le poids w d'un tag t représente la fréquence de son utilisation par l'utilisateur en co-occurrence avec d'autres tags. Formellement, $Soc = \{(t, w) | t \in T_u\}$, où T_u est le vocabulaire des tags de l'utilisateur.

Définition 6. La dimension thématique ou *Top* est un ensemble de sujets pondérés extraits de la dimension sociale dans laquelle le candidat est susceptible d'être un expert. Le poids du sujet pour un expert *Exp* représente l'estimation de l'expertise de

l'utilisateur dans le sujet considéré. Formellement, $Top = \{(\vec{s}, Exp) | \vec{s} \in S_u\}$, où S_u est l'ensemble des sujets de l'expert.

Définition 7. Un profil de l'expert i, noté $\overrightarrow{E_i}$, est un couple de deux vecteurs $\overrightarrow{E_i} = (\overrightarrow{Soc_i}, \overrightarrow{Top_i})$, où $\overrightarrow{Soc_i}$ est le vecteur de tags de l'expert et $\overrightarrow{Top_i}$ est le vecteur des topics d'expertise de l'expert.

Un profil expert peut ainsi être illustré de la manière suivante :

$$\text{User Profile} \begin{cases} \overrightarrow{Soc} = \{(t_1, w_1), (t_2, w_2), \dots, (t_n, w_n)\} \\ \overrightarrow{Top} = \{(s_1, Exp_1), (s_2, Exp_2), \dots, (s_m, Exp_m)\} \end{cases}$$

$$\text{with} \begin{cases} \overrightarrow{s_1} = \{(t_1, p_1(t_1)), (t_2, p_1(t_2)), \dots, (t_l, p_1(t_l))\} \\ \overrightarrow{s_2} = \{(t_1, p_2(t_1)), (t_2, p_2(t_2)), \dots, (t_o, p_2(t_o))\} \\ \vdots \\ \overrightarrow{s_m} = \{(t_1, p_m(t_1)), (t_2, p_m(t_2)), \dots, (t_q, p_m(t_q))\} \end{cases}$$

Figure 11: Modèle du Profil de l'expert.

5.2. Construction du profil (Profiling)

La construction d'un profil de l'expert consiste à construire ses deux dimensions : sociale et thématique. Pour la première, il s'agit de définir le vecteur $\overrightarrow{Soc}$ en extrayant les tags utilisateur et leurs poids. Pour la seconde, il faut déduire les sujets de la première dimension en construisant les vecteurs de chaque sujet et calculer l'expertise de l'utilisateur pour chaque sujet. Il faut noter qu'avant de construire les dimensions, les tags doivent subir un prétraitement composé de plusieurs étapes, surtout que ces derniers sont introduits par les utilisateurs d'une manière complètement libre.

5.2.1. Prétraitement

Pour faciliter le travail et ne garder que les tags significatifs, ceux-ci doivent passer par un certain nombre d'étapes qui consiste en :

-Suppression des mots vides (stop words) : elle consiste à supprimer les mots considérés comme inutiles tel que les prépositions, les déterminants et les pronoms.

- Clusterisation par rapport à la variation d'écriture : généralement pour faire un regroupement (*clustering*) de termes sur la base de la variation d'écriture par exemple (email, e-mail, mail…), on applique une mesure de calcul de distance. La plus utilisée dans ce cas est celle appelée *damerau-levenshtein*. *C'*est une mesure de distance appliquée entre deux chaines de caractères, son principe est de calculer le nombre minimal d'opérations nécessaire pour transformer une chaine de caractères en une autre, ces opérations sont : l'insertion, la suppression, la substitution d'un caractère et la transposition de deux caractères (Brill & Moore, 2000). Donc au lieu de traiter les tags « email » et « e-mail » comme deux tags différents, cette mesure nous permet de les considérer comme un même tag.

- *Stemmatisation* ou racinisation : est une technique qui vise à transformer les mots en leur radical, elle cherche selon le mot et la langue et définit le radical le plus probable pour ce mot, elle fonctionne uniquement avec une base de connaissances des règles syntaxiques et grammaticales de la langue. Le stemme (racine) est la partie restante d'un mot une fois que son suffixe et préfixe sont supprimés ; il peut ne pas correspondre à un mot réel (Lovins, 1968). Nous avons appliqué l'algorithme Porter Stemmer proposé dans (Porter, 1980) qui est le plus utilisé dans les cas de normalisation des termes.

5.2.2. Construction de la dimension sociale

Comme illustré dans la Figure 11, la dimension sociale est un vecteur de tags pondérés. Le poids de chaque tag représente la fréquence de son utilisation par l'utilisateur en co-occurrence avec d'autres tags.

Notre objectif à travers cette dimension est de trouver les tags utilisateur qui décrivent le mieux son expertise. En d'autres termes, extraire les tags pouvant faire partie de ses intérêts mais aussi ses compétences. Les approches de la littérature les plus connues pour l'extraction du profil à base des tags, sont l'approche naïve et l'approche par co-occurrence. L'approche naïve consiste à compter les occurrences et à les classer en fonction de leur popularité et l'approche par co-occurrence (Cayzer & Michlmayr, 2009) consiste à prendre des tags utilisés en combinaison (co-occur). Alors que l'approche naïve est permissive pour les tags génériques et que l'approche par co-occurrence ne permet aucune pondération des tags, nous adoptons une approche hybride (Kichou et al., 2011) basée sur les deux approches précédentes. L'approche hybride permet à la fois de récupérer uniquement des tags spécifiques et avec

pondération. Le résultat de la combinaison de l'approche naïve et l'approche par co-occurrence est un graphe de nœuds et d'arcs pondérés. Les nœuds (tags) pondérés appartenant aux arcs ayant les plus grands poids composent le vecteur de la dimension sociale de l'utilisateur.

L'algorithme Add-A-Tag proposé dans (Cayzer & Michlmayr, 2009) pour mettre en œuvre une nouvelle approche de construction du profil, *Adaptive approach*[8] est adopté dans nos travaux afin de réaliser l'approche hybride.

Soit u un utilisateur taggant un nombre de ressources avec l'ensemble de tags : $T=\{t_1,t_2,...t_n\}$. Le graphe du profil de l'utilisateur u G_u (V, E) où $V=\{v_1,v_2,.....v_n\}$ est l'ensemble des nœuds (Vertices), et $E=\{e_1,e_2,...e_n\}$ est l'ensemble des arcs (Edges).

Etape 1 : Mise à jour du graphe

Les n nouveaux tags introduits par l'utilisateur *u* pour un produit donné sont ajoutés au graphe. Pour toute combinaison $t_i\ t_j$ où $i, j \in$ à $1,2...n$ et $i<j$ la procédure suivante est exécutée :

1. Pour chaque tag t_x avec $x \in i, j$ ajouter au graphe le nœud correspondant v_x si celui-ci n'existe pas ;
2. Si le nœud n'existe pas, créer un arc de poids égal à *1* entre le nœud v_i et le nœud v_j ;
3. Si le nœud existe déjà, incrémenter de 1 le poids de l'arc entre v_i et v_j ;
4. Affecter pour chaque nœud du graphe son poids (sa popularité).

Etape 2 : Extraction du profil

1. Créer un sous ensemble E_s de *E*, ordonné avec un ordre décroissant des poids des arcs ;
2. Choisir le top *k* des éléments de E_s avec k un entier non nul ($k>0$) ;
3. Ajouter au profil les tags correspondants aux arcs élus et leurs poids (popularités).

[8] L'approche adaptative est une extension de l'approche par co-occurrence, dans laquelle les auteurs ont introduit la notion d'âge du tag pour favoriser les nouveaux tags associés par l'utilisateur.

La taille du profil est déterminée par la valeur du paramètre *k*. c'est un vecteur de termes (tags) pondérés.

À partir de tous les tags de l'utilisateur, un graphe est construit. Les nœuds représentent les tags et les arcs les relations de co-occurrence entre ceux-ci. Les nœuds sont pondérés avec le nombre de fois que l'utilisateur a utilisé le tag (popularité), et les arcs avec le nombre de co-occurrence de chaque paire de tags.

Il faut noter aussi que l'approche par co-occurrence permet de prendre les tags liés au domaine de la ressource du moment qu'ils sont cités fréquemment ensemble. Le risque d'avoir des tags en dehors du sujet et le problème d'ambiguïté sont minimes. Un exemple de graphe construit avec l'approche hybride est illustré dans la *Figure 12*.

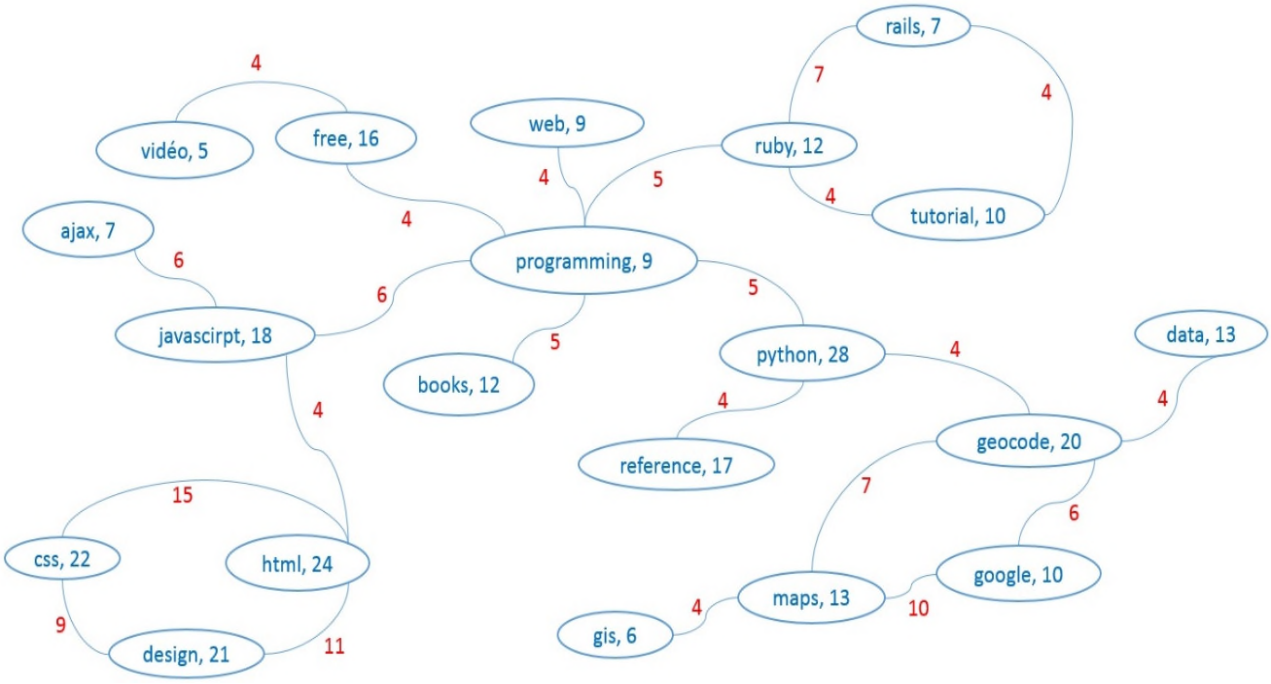


Figure 12: Exemple de graphe construit avec l'approche hybride, (Kichou et al., 2011)

Pour construire le vecteur $\overrightarrow{Soc}$, seuls les arcs ayant les poids les plus élevés sont considérés et les nœuds les formant avec leurs popularités sont intégrés dans le vecteur $\overrightarrow{Soc}$. Afin d'éviter de prendre des tags insignifiants, nous limitons au préalable la taille de $\overrightarrow{Soc}$. Par exemple, le vecteur de dimension sociale $\overrightarrow{Soc}$ correspondant à l'exemple de la Figure 12 ne considérant que les arcs de poids supérieur ou égal à 7 est le suivant :

$\overrightarrow{Soc}$= {(html,24), (css,22), (design,21), (maps,13), (google,10), (geocode,20), (ruby,12), (rails,7)}.

5.2.3. Construction de la dimension topicale (thématique)

La dimension topicale se construit en deux étapes principales : la déduction des thèmes de la dimension sociale et l'estimation de l'expertise de l'utilisateur sur chaque thème. Cette section se concentre uniquement sur la présentation des sujets et la section suivante est entièrement consacrée à l'estimation de l'expertise des utilisateurs en raison de son importance pour notre travail.

Comme nous l'avons déjà mentionné, l'approche la plus adoptée dans la littérature pour la modélisation thématique est l'algorithme d'allocation latente de Dirichlet (LDA) (Blei, Ng, et al., 2003). LDA est surtout utilisé dans le cas des documents, une adaptation pour notre cas s'impose.

Notre choix pour LDA se justifie par sa réputation et sa large utilisation dans la littérature dans le domaine du *topic modeling* et la recherche d'expertise. La quasi-totalité des travaux ayant recours au *topic modeling* utilisent LDA. Ses versions ultérieures ont chacune des contraintes spécifiques (par exemple : LLDA= documents étiquetés, AMT= documents à plusieurs auteurs…etc.).

LDA est adapté à notre contexte (*Figure 13*) en considérant les dimensions sociales des utilisateurs comme des entrées. Par analogie, le terme w dans LDA correspond au tag t. Le document d correspond au vecteur social de l'utilisateur u et le corpus correspond à l'ensemble des ressources disponibles dans le média social (l'ensemble R). Pour faire une utilisation crédible de LDA, nous considérons un grand nombre de tags pour chaque utilisateur. Le tag t_i est le tag d'index i dans la dimension sociale de l'utilisateur u. t_i appartient à un cluster de tags (le cluster contient toutes les variantes du tag liées à l'écriture, exemple : web2.0, web2-0, etc.). Nous visons à faire une distribution de tags décrivant les ressources sur des sujets et en déduire la répartition des candidats sur les sujets.

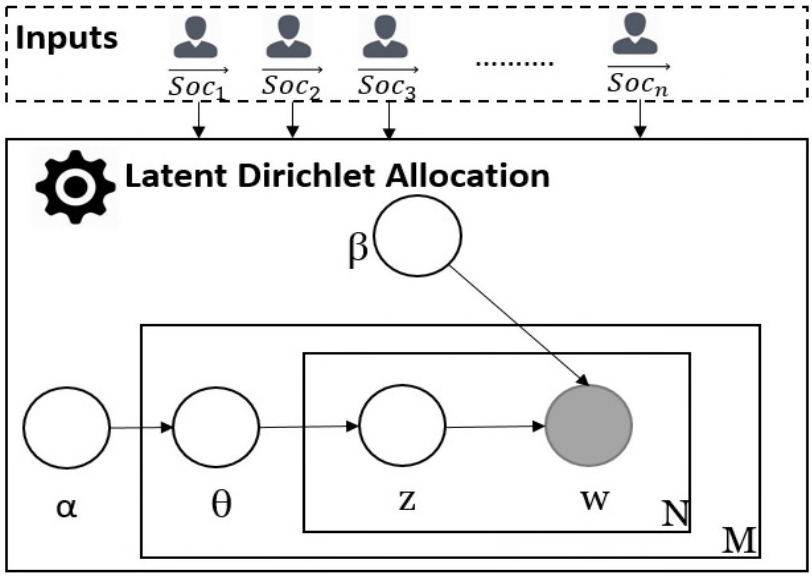


Figure 13: Description du processus de LDA

La partie inférieure de la *Figure 13* montre un modèle de représentation graphique de LDA (Bietti, 2012) adapté au contexte du tagging social, où les boîtes sont des plaques représentant des répliques. La plaque *M* représente les documents (pour notre cas c'est un candidat représenté par son ensemble de tags qui ont servi à décrire des ressources), tandis que la plaque intérieure *N* représente le choix répété de sujets et de tags pour un même document. α et β sont les hyperparamètres, α est le paramètre de Dirichlet. θ est la proportion des sujets pour chacun des candidats, z et w sont au niveau tag.

La modélisation thématique est réalisée en appliquant LDA comme l'illustre la *Figure 14* ci-dessous :

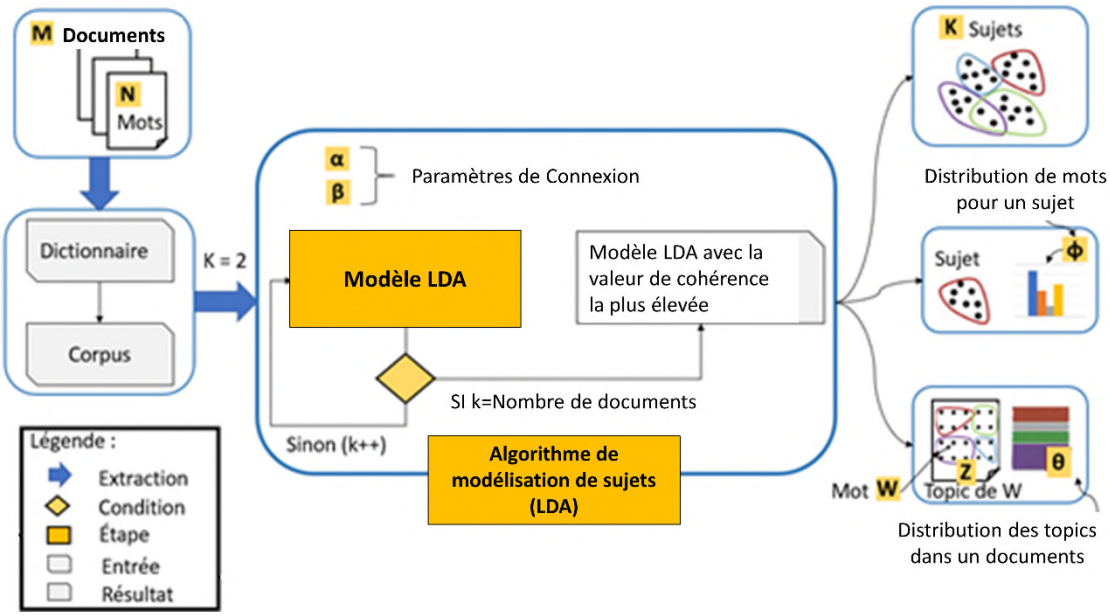


Figure 14: Processus de modélisation thématique avec LDA.

En effet, afin de trouver le nombre optimal de sujets dans les documents (rappelons que pour nous, chaque document est l'ensemble des tags du candidat) notre solution repose sur la construction de nombreux modèles LDA avec de différents nombres de sujets (*K*), et de choisir celui qui donne la meilleure valeur de cohérence et qui marque la fin de croissance rapide de cette cohérence. En effet, le calcul du score de cohérence est l'une des techniques largement utilisées pour trouver le nombre optimal de sujets. Le score de cohérence est basé sur l'information mutuelle ponctuelle (*pointwise mutual information, PMI*), il est étudié dans plusieurs travaux (pour plus de détails voir (Röder et al., 2015; Syed & Spruit, 2017).

Une fois le processus exécuté, deux résultats sont obtenus : (1) La liste des sujets avec leurs vecteurs descriptifs $\overrightarrow{s_i}$ contenant pour chaque tag appartenant au sujet considéré son poids (probabilité), (2) la liste des sujets pondérés pour chaque candidat (*Figure 15*).

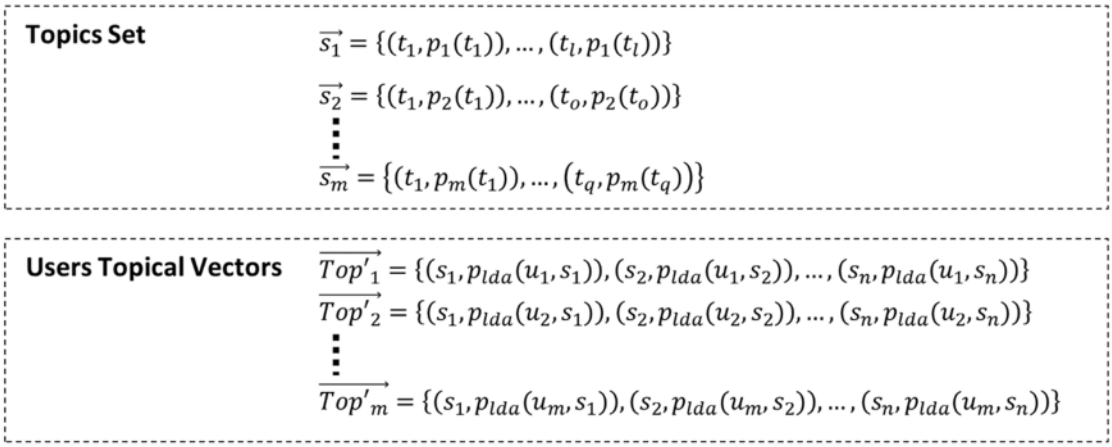


Figure 15: Les résultats de LDA (Outputs).

$P_{lda}(u_i, s_j)$ est la probabilité que le sujet s_j appartienne au vecteur thématique de l'utilisateur u_i. Dans un modèle LDA à K thèmes, $P_{lda}(u_i, s_j)$est la probabilité que le tag (terme) w d'un vecteur social utilisateur $\overrightarrow{Soc}$ (longueur Nm) soit instancié par un tag t du vocabulaire en utilisant la formule :

$$p(w_{m,n} = t) = \sum_{k=1}^{K} p(w_{m,n} = t | z_{m,n} = k) p(z_{m,n} = k) \qquad (11)$$

La plupart des implémentations supposent que les distributions de Dirichlet sont symétriques, ce qui signifie que l'on ne veut pas discriminer a priori les termes et les thèmes du modèle. À notre avis, cette probabilité est insuffisante pour évaluer l'expertise d'un utilisateur sur un sujet donné. La distribution obtenue par LDA, nous renseigne uniquement sur les éventuels sujets d'expertise d'un utilisateur. En effet, la dimension thématique que nous avons obtenue, contient les topics dont l'utilisateur est susceptible d'être expert. Nous allons dans ce qui suit raffiner l'estimation des probabilités calculées par LDA liées à chaque topic, et ce en introduisant la notion de profondeur de tag.

6. Estimation de l'expertise de l'utilisateur

La probabilité $P_{lda}(u_i, s_j)$ calculée par l'algorithme LDA lors de la modélisation des vecteurs thématiques considère les tags comme une distribution statistique, dans laquelle les tags sont sémantiquement équivalents. Certes, LDA exploite la sémantique latente d'un mot lorsqu'il apparait relativement fréquemment dans un contexte. Cependant, nous voulons introduire la relation taxonomique mettant en évidence la généralisation/spécialisation du terme. Un candidat utilisant des tags génériques ne

doit pas être considéré de la même manière qu'un utilisateur utilisant des tags plus précis et plus significatifs. C'est pourquoi l'introduction de l'aspect hiérarchique des tags dans l'évaluation de l'expertise d'un utilisateur sur un sujet donné améliorera considérablement les performances de recherche des experts. Cela passe par une meilleure différenciation des utilisateurs en fonction des tags qu'ils utilisent. Citons l'exemple d'un utilisateur qui associe le tag *medication* et un autre qui associe exactement le nom de la molécule du médicament. Le deuxième est certainement plus connaisseur que le premier.

Nous avons donc besoin d'utiliser une ontologie qui rassemble un ensemble de concepts décrivant complètement un domaine, ou une base de données lexicale telle que Wordnet[9]. Ces concepts sont liés les uns aux autres par des relations taxonomiques (concepts hiérarchiques) d'une part, et sémantiques d'une autre part. La relation hiérarchique entre les concepts peut refléter le niveau de précision d'un concept par rapport à un autre. Les concepts les plus génériques se trouvent au sommet de la structure, tandis que les concepts les plus spécifiques se situent au niveau feuille. Par conséquent, plus un tag est profond dans la structure hiérarchique de l'ontologie, plus il est précis. Ainsi, un utilisateur qui utilise un tel tag est susceptible de connaitre mieux le domaine qu'un utilisateur qui utilise des tags plus génériques. Cette hypothèse a été déjà considérée dans (Kichou et al., 2011) d'une manière différente.

Définition 8. La profondeur d'un tag *t* notée *Depth(t)* est la distance entre celui-ci et la racine du graphe de l'ontologie utilisée.

Un terme peut avoir plusieurs chemins vers la racine, ce qui engendre plusieurs valeurs de profondeurs, nous considérons la minimale.

La *Figure 16* illustre un exemple de profondeurs pour le terme *Autocar* dans wordnet.

[9] https://wordnet.princeton.edu/

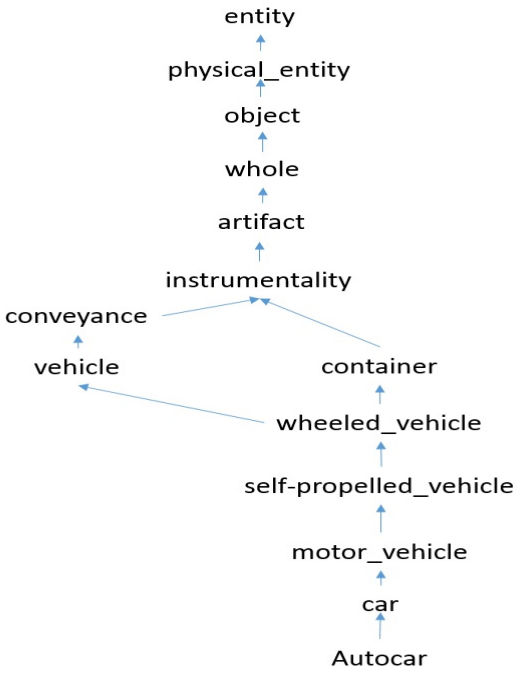


Figure 16: Exemple de profondeurs du terme 'Autocar'.

L'expertise d'un utilisateur u_i sur un sujet s_j noté *Exp* (u_i, s_j) est calculée comme suit :

$$Exp(u_i, s_j) = P_{lda}(u_i, s_j) * (1 + \sum_{k}^{|\overrightarrow{Soc_i}|} f_j(t_k)^{depth(t_k)}) \qquad (12)$$

Avec *depth* (t_k) la profondeur du tag t_k dans l'ontologie considérée et f_j une fonction qui renvoie le poids w_k du tag t_k dans le vecteur social de l'utilisateur u_i si t_k appartient au vecteur descriptif du sujet s_j et 0 sinon. f_j est définie comme suit :

$$f_j(t_k) = \begin{cases} W_k \; if \; t_k \in \overrightarrow{s_j} \\ 0 \; sinon \end{cases} \qquad (13)$$

Cette formule permet aux utilisateurs qui utilisent des tags spécifiques dans la description de la ressource d'avoir la préférence sur ceux qui utilisent des tags génériques. Plus un utilisateur utilise des tags profonds, plus son expertise exprimée par la deuxième partie de la formule est grande et plus son expertise globale est grande.

Lorsqu'un utilisateur ne taggue aucune ressource du sujet considéré, la somme dans la formule est égale à zéro, pour éviter d'ignorer ces utilisateurs, nous avons appliqué un lissage de Laplace (Additive smoothing) et ceci en rajoutant 1 à cette somme. Cela veut dire qu'un utilisateur qui n'a pas encore de tags (personomie vide pour un sujet donné), la valeur de son expertise pour ce sujet sera la probabilité calculée par LDA.

7. Recherche de l'expert (Finding)

Nous définissons un modèle de recherche d'expert basé sur une adaptation du modèle vectoriel classique de la recherche d'information (Nie, 1988). La définition d'un modèle de repérage ou de recherche d'experts se compose de deux éléments principaux : la définition de la fonction de pondération du tag dans un profil utilisateur en intégrant son expertise et la définition de la fonction de mise en correspondance.

7.1. Présentation du modèle vectoriel

Les modèles vectoriels ont été proposés pour compenser la limitation de la pondération binaire des modèles booléens. Ils sont basés sur l'attribution de poids non binaires aux termes indexés dans les documents et aux termes des requêtes.

À la fin des années 1950, Luhn a d'abord suggéré que les systèmes automatiques de recherche de texte devraient être conçus sur une comparaison entre le contenu textuel des documents et celui de la requête utilisateur. Les documents seraient représentés par des vecteurs de terme de la forme $\vec{D}$= (t_i, t_j, …, t_p), où chaque t_k identifie un terme assigné au document D. De façon analogue, un vecteur de requête pourrait être formulé $\vec{Q}$ = (q_a, q_b, …, q_r) (Dominich, 2008).

Dans les premiers modèles vectoriels proposés par *Salton* (Salton et al., 1975) chaque document et chaque requête est codé par un vecteur dans un espace vectoriel où chaque dimension représente une caractéristique du document (Figure 17). Le document et la requête ont tous les deux une représentation vectorielle, respectivement : $\vec{d} = \{w_{1,j}, w_{2,j}, ..., w_{t,j}\}$ pour le document et $\vec{q} = \{w_{1,q}, w_{2,q}, ..., w_{t,q}\}$ pour la requête. Où $w_{i,j}$ et $w_{i,q}$ représentent le poids du terme *i*, respectivement, dans le document et la requête et les deux vecteurs doivent être dans le même espace.

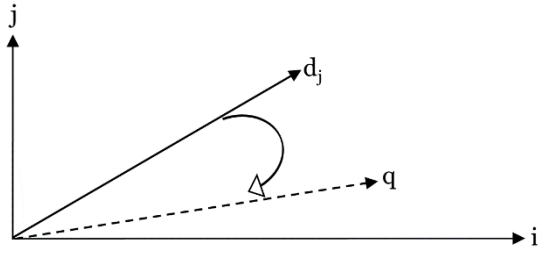


Figure 17 : Représentation des documents dans le modèle vectoriel, (Salton et al., 1975)

Les modèles vectoriels sont des modèles basés sur des statistiques qui permettent d'une part de quantifier les termes et d'autre part, de mesurer le degré de pertinence d'un document vis-à-vis d'une requête.

La pertinence est estimée grâce à des mesures de similarité dont la plus simple est le produit scalaire et la plus populaire est le cosinus.

$$RSV(\vec{d},\vec{q}) = \sum_{i=1}^{t} w_{i,j} * w_{i,q} \qquad (14)$$

$$RSV(\vec{d},\vec{q}) = \frac{\vec{d}_j * \vec{q}}{\left|\vec{d}_j\right| * \left|\vec{q}\right|} = \frac{\sum_{i=1}^{t} w_{i,j} * w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} * \sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \qquad (15)$$

D'autres fonctions de similarité ont été proposées dans la littérature (Piwowarski, 2003) telles que le pseudo-cosinus et les mesures de Jaccard et Dice.

Dans notre modèle vectoriel étendu à la recherche d'experts, le document représente l'expert dont la dimension sociale est considérée comme vecteur descriptif et la requête est formée d'un vecteur de tags. Les autres composantes du profil expert sont exploitées dans le calcul du poids du terme tel qu'expliqué dans la section suivante.

7.2. Calcul du poids d'un tag

Dans le modèle de recherche d'expert proposé, le poids d'un tag *t* lié au profil utilisateur *u* noté w_t^u est le poids du tag *t* dans le vecteur social de *u* multiplié par la moyenne de l'expertise de l'utilisateur dans les sujets où le tag *t* apparait. w_t^u est donc défini comme suit :

$$w_t^u = w_t^{\overrightarrow{soc_u}} * \sum_j^{\left|\overrightarrow{Top_u}\right|} \exp(u,\overrightarrow{s_j}) / \left|\overrightarrow{Top_u}\right| \qquad (16)$$

Où $w_t^{\overrightarrow{soc_u}}$ est le poids du tag *t* dans le vecteur social de l'utilisateur *u*, $\exp(u,\overrightarrow{s_j})$est l'expertise de l'utilisateur *u* dans le sujet s_j.

La première partie de cette formule $w_t^{\overrightarrow{soc_u}}$ représente la popularité du tag t pour l'utilisateur u calculé à base de sa dimension sociale qui est, à son tour, extraite en utilisant la méthode hybride garantissant ainsi l'apparition uniquement de tags significatifs pour l'utilisateur en question. Plus le tag est utilisé par un utilisateur plus la popularité de ce tag est grande pour l'utilisateur en question. La popularité d'un tag pour utilisateur donné traduit le niveau de compréhension de ce dernier du sens du tag

en question les tags qui sont accidentellement utilisés par l'utilisateur ou qui apparaissent rarement se verront affecté une popularité réduite par rapport aux tags fréquemment utilisés. Cependant, afin de s'assurer que le tag est utilisé par l'utilisateur à bon escient, nous intégrons la seconde partie $\sum_{j}^{|\overrightarrow{Top_u}|} \exp(u, \vec{s_j}) / |\overrightarrow{Top_u}|$ qui calcule la moyenne des expertises de l'utilisateur dans les topics dans lesquels le tag t apparait. Cette partie permet de contextualiser l'utilisation du tag t. Ainsi, si un utilisateur utilise uniquement le tag t dans les topics dans lesquels il est expert (i.e. son expertise est grande) cette partie sera importante et le poids w_t^u le sera également. En revanche, si l'utilisateur utilise le tag t dans des topics dans lesquels son expertise n'est pas importante, cette seconde partie de la formule réduira ainsi le poids w_t^u.

Soit l'exemple suivant. Dans le premier tableau, nous présentons deux tags avec leurs popularités respectives pour les utilisateurs u_1 et u_2 ainsi que les topics dans lesquels ils apparaissent. Le second tableau résume les expertises des deux utilisateurs u_1 et u_2 dans les mêmes topics estimés avec la formule 12.

Tag	Popularité		C 1	C2	C3	C4
	U1	U2				
Java	13	25	✓			✓
C#	32	10		✓	✓	

Tableau 2 : Exemple de tags, leurs popularités et les topics correspondants

		C 1	C2	C3	C4
Expertise	U1	0.26	0.12	0.02	0.45
	U2	0.12	0.25	0.68	0.01

Tableau 3 : Les expertises résultantes

Le calcul des w_t^{u1} pour l'utilisateur u_1 et les deux tags « java » et « C# » selon la formule 16 est comme suit :

$$w_{java}^{u1} = 13 * \frac{0.26 + 0.45}{2} = 4.62 \qquad w_{C\#}^{u1} = 32 * \frac{0.12 + 0.02}{2} = 2.24$$

Le tag java apparait dans les topics C_1 et C_4 dans lesquelles l'utilisateur est le plus expert et n'apparait pas dans les deux autres. A l'inverse de ce dernier, le tag C# apparait dans les topics C_2 et C_3 dans lesquelles l'utilisateur est le moins expert. Remarquons que même si la popularité du tag C# est plus grande que celle du tag Java,

le fait que java soit utilisé par l'utilisateur à partir des topics dans lesquels il est le plus expert, le poids de ce dernier est plus élevé que celui du tag C# même avec sa grande popularité.

7.3. La Fonction d'appariement

La formule 16 permet de comparer deux utilisateurs par rapports à un tag donné. Plus le poids w_t^u est grand, plus l'utilisateur u est considéré comme étant expert dans l'utilisation de ce tag. Pour un ensemble de tags représentant un topic donné, cette formule nous permet ainsi de classer les utilisateurs par rapport à leur expertise dans le topic en question. La comparaison entre la requête et les profils experts est réalisée avec la *Retrieval Statut Value (RSV)*.

Dans le cas du modèle vectoriel que nous adoptons, le cosinus est la fonction d'appariement la plus utilisée dans le domaine de la recherche d'information et c'est celle que nous adoptons dans nos travaux comme *Retrieval Statut Value (RSV)* de la manière suivante :

$$RSV(q,u) = \cos(\vec{q},\vec{u}) = \frac{\sum_i^{|V|} w_i^q w_i^u}{\sqrt{\sum_i^{|V|}(w_i^q)^2}\sqrt{\sum_i^{|V|}(w_i^u)^2}} \qquad (17)$$

Où $\vec{q}$ est le vecteur de la requête composée par un ensemble de tags et $\vec{u}$ représente l'utilisateur pour lequel la RSV est calculée. w_i^q est le poids du tag i dans la requête q qui est généralement égale à 1 et w_i^u est le poids du tag i par rapport à l'utilisateur u calculé à l'aide de la formule 16.

En considérant l'exemple précédent, la RSV calculée pour les deux utilisateurs u_1 et u_2 pour la requête formée des tags « java » et « C# » est comme suit :

Premièrement, calculons le poids de chaque tag par rapports aux deux utilisateurs en fonction de la formule 16.

U	**Java**	**C#**
$\mathcal{U}1$	$w_{java}^{u1} = 13 * \frac{0.26 + 0.45}{2} = 4.62$	$w_{C\#}^{u1} = 32 * \frac{0.12 + 0.02}{2} = 2.24$
$\mathcal{U}2$	$w_{java}^{u2} = 25 * \frac{0.12 + 0.01}{2} = 1.63$	$w_{C\#}^{u2} = 10 * \frac{0.25 + 0.68}{2} = 4.65$

Tableau 4 : Calcul des poids des tags

Calculons à présent, la RSV pour chaque utilisateur :

$RSV(q,u1) = \frac{4.62 + 2.24}{2\sqrt{4.62^2 + 2.24^2}} = 0.67$	$RSV(q,u2) = \frac{1.63 + 4.65}{2\sqrt{1.63^2 + 4.65^2}} = 0.64$

Tableau 5 : Calcul du RSV

L'utilisateur u_1 est plus expert par rapport à la requête q={java,C#} que l'utilisateur u_2.

Le modèle de recherche que nous proposons et qui est basé sur l'expertise utilisateur permet ainsi d'ordonner un ensemble d'experts par ordre de pertinence, exprimée ici par leur niveau d'expertise, par rapport à une requête exprimée sous forme d'une suite de tags. Quand la requête contient les tags représentatifs d'un topic donné, le modèle permet de rechercher les experts par rapport au topic en question.

8. Conclusion

Partant de l'idée que les tags d'un utilisateur expriment sa perception et son avis sur la ressource taguée et prenant en compte le fait que cette source d'information n'a pas été suffisamment explorée dans les travaux de recherche d'expert sur le Web, nous avons bâti un nouveau modèle permettant d'un côté le calcul de l'expertise d'un utilisateur par rapport à un domaine donné et d'extraire ainsi les domaines d'expertise de ce dernier, et d'effectuer une recherche d'experts par rapport à une requête exprimée sous forme d'une liste de Tags, d'un autre côté.

Nous avons présenté dans ce chapitre les fondements théoriques des deux modèles, à savoir le modèle de calcul de l'expertise et le modèle de recherche d'experts qui représentent notre principale contribution dans ces travaux de thèse. En résumé, cette contribution est principalement composée de deux phases clé : celle du *profiling* qui définit et construit le profil de l'utilisateur et la phase de recherche (*finding*) qui renvoie les experts choisis pour un besoin donné. Nous avons également introduit une nouvelle notion qu'est la profondeur du tag pour évaluer l'expertise d'un utilisateur en estimant que plus un tag est profond dans la hiérarchie des concepts plus ce dernier est précis et plus l'utilisateur l'ayant utilisé est expert dans le domaine de la ressource taguée. Nous avons extrait les intérêts des utilisateurs à partir des données du tagging social en utilisant un algorithme de modélisation de sujets (LDA) pour la découverte des topics traités, et la distribution des tags sur ces topics. L'algorithme LDA est affiné

par l'utilisation des profondeurs des tags afin d'avoir un sujet plus spécifique et plus proche de l'expertise du candidat.

Dans les chapitres suivants, nous allons expérimenter les deux modèles proposés. Nous allons dans un premier temps effectuer une série d'expérimentations sur des Datasets reconnus dans le domaine avec des métriques permettant la comparaison des résultats avec ceux des meilleurs modèles du domaine. Puis, dans un second temps, appliquer le modèle dans le cas d'un projet réel consistant en un modèle de recommandation des femmes artisans.

Chapitre V
Expérimentations et Résultats

1. Introduction

Dans le but d'évaluer le modèle de calcul de l'expertise utilisateur et le modèle de recherche d'expert proposés dans le chapitre précédent, nous avons réalisé un ensemble d'expérimentations sur différents datasets. Ces expérimentations ont été conçues pour répondre aux questions de recherche suivantes :

1. L'exploitation de l'activité de tagging d'un utilisateur est-elle efficace pour créer le profil d'un expert ?
2. Quelle est la contribution de l'exploitation des liens hiérarchiques entre les tags dans l'estimation de l'expertise d'un utilisateur ?
3. La combinaison de l'estimation de l'expertise et de la modélisation thématique améliore-t-elle les performances de recherche d'experts ?

Dans la suite de ce chapitre, nous présentons d'abord le Dataset utilisé pour cette première partie des tests, et la configuration des paramètres. Puis nous présentons nos résultats expérimentaux pour répondre aux questions de recherche mentionnées ci-dessus. Nous allons également pousser nos expérimentations en utilisant deux autres Datasets, à savoir twitter et Delicious afin d'étudier les performances des modèles proposés.

2. Collection et Métriques

Nous utilisons la collection CQA *Stack Overflow* qui est un ensemble de données extraites à partir du site *Stack Overflow* contenant toutes les publications faites entre août 2008 et mars 2015, soit 24 120 523 publications, avec des informations sur environ 4 millions d'utilisateurs.

Nous utilisons trois requêtes liées aux tags « Java », «C # » et « Android » pour réaliser les expériences. Chacune des requêtes est décrite avec les 200 tags les plus

populaires qui se sont produits avec le tag associé. Le *Tableau 6* ci-dessous présente les requêtes avec un extrait des tags associés :

Requête	**Tags**
Java	Java, android, multithread, xml, swing, eclipse, spring, hibernate, jsp, servlet
C#	C#, .net, asp .net, winform, multithread, xml, linq, wpf, asp .net_mvc, c++
Android	Android, java, xml, eclipse, layout, listview, sqlite, android_act, android_emul, android_int

Tableau 6: Les requêtes utilisées pour l'expérimentation.

L'ensemble des experts pertinents pour chaque requête (résultats idéaux) est construit en fonction du score attribué aux réponses des utilisateurs par leurs pairs ainsi que de la popularité de l'utilisateur considéré. La somme des scores de réponses données par un utilisateur en réponse aux messages relatifs à une requête représente le RSV de cet utilisateur dans le vecteur idéal de la requête considérée.

Pour comparer le modèle proposé aux modèles de base, plusieurs mesures de la recherche d'informations populaires (Büttcher et al., 2016) sont utilisées, notamment la précision au rang n (P@n), la moyenne de la précision moyenne (*MAP, Mean Average Precision*), le gain cumulatif actualisé normalisé au rang n (*NDCG@n, Normalized Discounted Cumulative Gain at rank n*).

(P@n) mesure le pourcentage d'experts pertinents dans le top n de la liste des candidats récupérés (Gui et al., 2018). Elle est calculée comme suit :

$$P@n = \sum_{i=1}^{n} R(C_i)/n \qquad (18)$$

Où $R(c_i) = 1$ si le $i^{ème}$ candidat récupéré est pertinent pour la requête donnée et 0 dans le cas contraire.

- La précision moyenne est définie comme suit :

$$AP = \sum_{i=1}^{R_n} (P@i * R(C_i))/R_n \qquad (19)$$

Où R_n est le candidat pertinent (expert).

- MAP est la moyenne des précisions moyennes de toutes les requêtes.

- NDCG est utilisé pour mesurer l'efficacité d'ordonnancement des candidats pertinents, elle repose sur des jugements graduels des candidats retournés. NDCG est définie comme suit :

$$NDCG@n = \sum_{i=1}^{n} R(C_i)/Log(i+1)/\sum_{i=1}^{n}[1/Log(i+1)] \quad (20)$$

3. Modèles de base (Baseline)

Afin d'évaluer les performances du modèle proposé noté Exp, nous le comparons d'abord aux deux modèles que nous avons définis. Le premier modèle BM est basé sur une pondération de tags avec le TF-IDF de la RI classique. Le deuxième modèle noté LDA, utilise la probabilité calculée avec l'algorithme LDA comme poids du tag dans le profil d'un expert. Ces deux modèles permettent de montrer l'apport de la profondeur du tag que nous avons introduite pour le calcul de l'expertise et l'exploitation des liens hiérarchiques des tags dans la performance du modèle de recherche.

Cinq modèles des plus connus dans la littérature de la recherche d'experts sont utilisés comme base de référence dans nos expérimentations. Le premier appelé DBA est une approche basée documents (documents-based) (Balog et al., 2009). Les deuxième et troisième modèles de base sont l'approche basée sur l'entropie et l'approche basée sur l'entropie étendue, respectivement appelées EBA et XEBA (Gharebagh et al., 2018). Le quatrième modèle proposé dans (Dehghan et al., 2019) exploitant le réseau de neurones LSTM pour trouver la forme d'expertise avec ses deux variantes selon les approches de traversée: DFS et BFS.

4. Définition des paramètres

Les paramètres de l'algorithme LDA utilisé dans la construction de sujets sont les suivants : le nombre d'itérations est de 1000, 20 sujets sont considérés avec un vecteur descriptif de 20 tags pour chaque sujet (après un ensemble de tests pour le calcul de la valeur de cohérence).

Le calcul de la profondeur des tags est basé sur Wordnet (Fellbaum & Vossen, 2012). Trec_eval (Balog et al., 2011) est utilisé pour calculer MAP, RPrec et les mesures Recall & Precision.

5. Résultats expérimentaux et analyse

5.1. Exp Vs LDA et le modèle de base BM

Le *Tableau 7* ci-dessous montre les résultats obtenus en comparant Exp avec le modèle de base BM et le modèle LDA, sur la base des mesures Map et Rprec.

Models	Java		Android		C#		ALL	
	Map	**Rprec**	**Map**	**Rprec**	**Map**	**Rprec**	**Map**	**Rprec**
BM	0,7829	0,7484	0,9077	0,9067	**0,2827**	**0,3064**	0,6578	0,6539
LDA	0,741	0,7262	0,9415	**0,9422**	0,2667	0,2409	0,6497	0,6364
Exp	**0,8884**	**0,8579**	**0,9634**	0,9408	0,2029	0,2177	**0,6849**	**0,6722**

Tableau 7: Résultats de Map et Rprec pour Exp, LDA et BM.

Les résultats ci-dessus montrent une amélioration de la mesure Map par notre modèle Exp de 4,12% par rapport au modèle BM et de 5,42% par rapport au modèle LDA. La combinaison de la probabilité calculée par le LDA avec l'expertise utilisateur proposée basée sur la profondeur du tag dans le modèle Exp démontre clairement son efficacité en améliorant considérablement les mesures MAP (qui est la moyenne des précisions moyennes de toutes les requêtes) et Rpec. Cette amélioration peut être ressentie davantage sur un large éventail de requêtes. D'autre part, selon la mesure Rprec (*Figure 18*), le premier expert pertinent est retourné plus tôt dans le modèle que nous proposons Exp par rapport aux modèles BM et LDA. Ce qui est un résultat satisfaisant et en faveur de notre modèle.

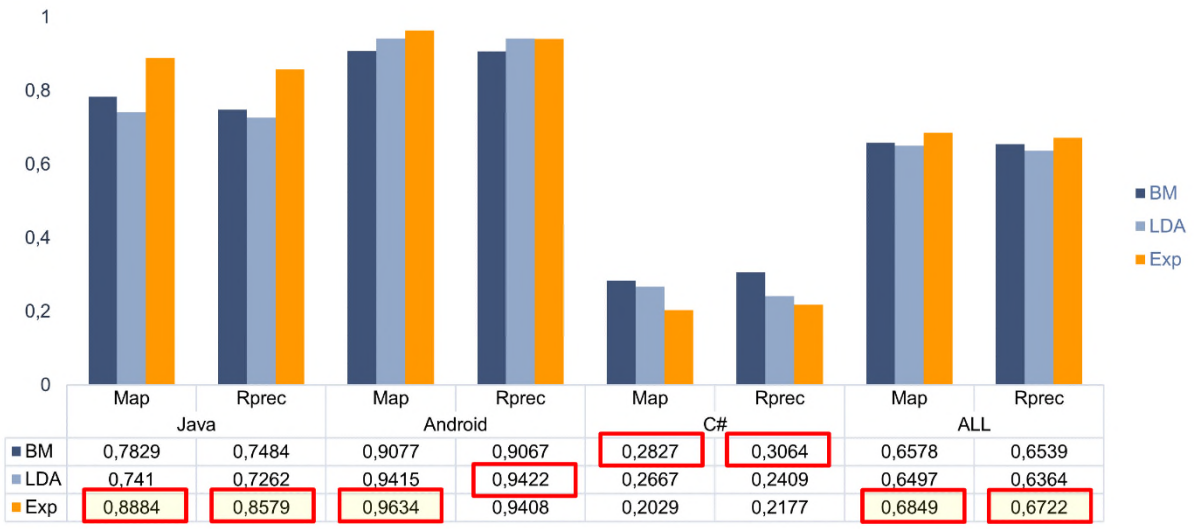
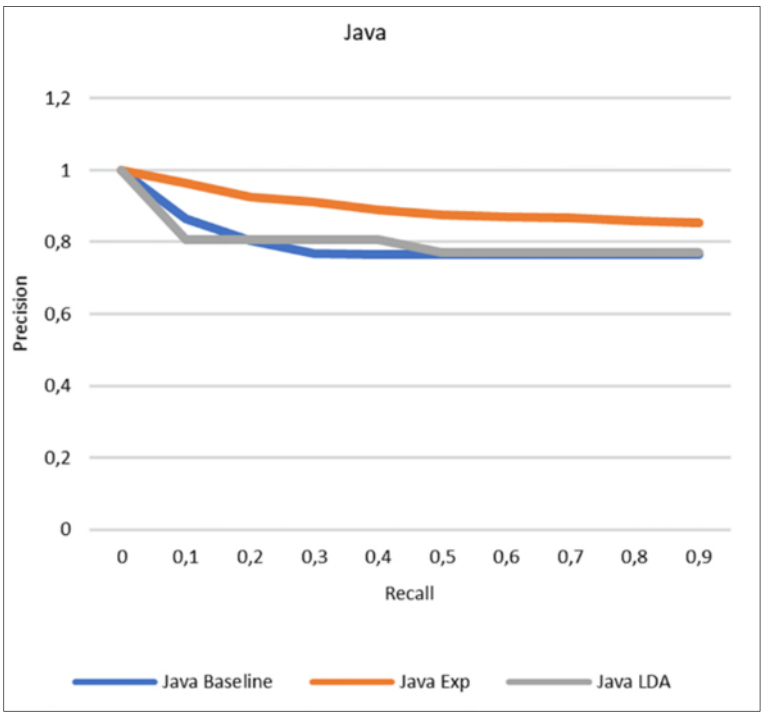


Figure 18 : Exp Vs BM et LDA (Map et Rprec)

La *Figure 19* et la *Figure 20* ci-dessous montrent également la comparaison des trois modèles en fonction de la mesure de rappel/précision interpolés. Ces résultats confirment notre observation et montrent l'amélioration apportée par notre modèle Exp dans les deux requêtes Java et C #. La troisième requête « Android » montre des résultats différents, qui sont dus aux types d'utilisateurs qui répondent à ses messages. Ces utilisateurs sont pour la plupart des experts en Java et utilisent donc des tags liés au développement et à Java en particulier. Ainsi, lors de la construction des sujets, leur expertise dans le sujet Java est plus importante que celle de la requête Android, chose qui diminue leur expertise pour cette requête.



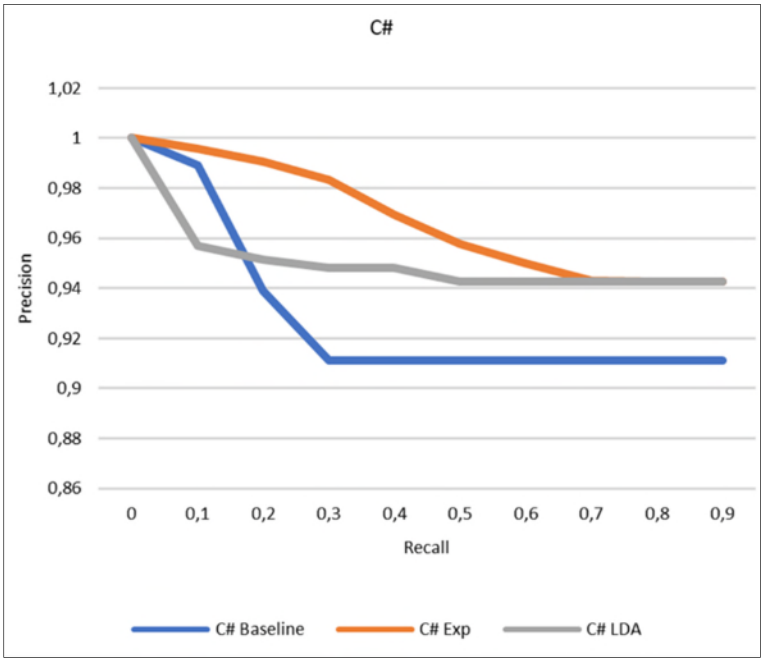
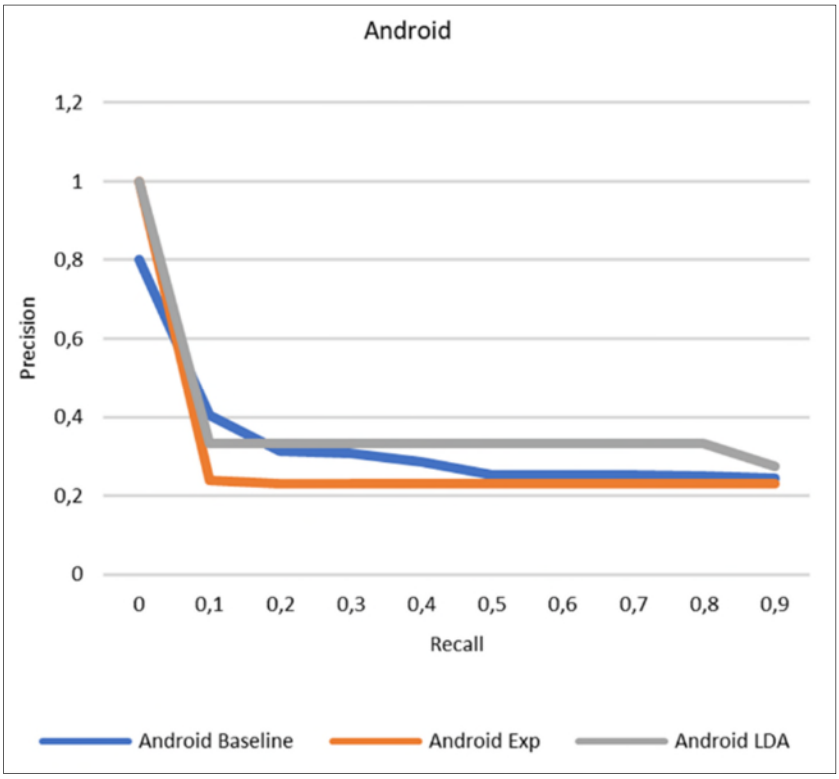


Figure 19 : Exp Vs Modèles de base (Rappel-Précision interpolés) pour la requête Java et C#



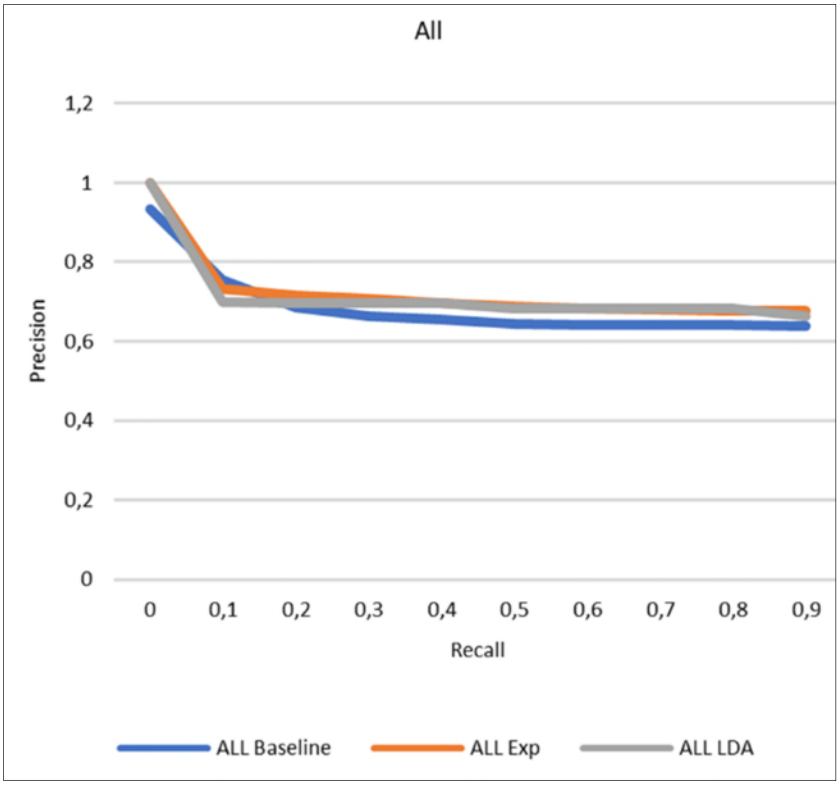


Figure 20: Exp Vs Modèles de base (Rappel/Précision interpolés) pour la requête Android et pour toutes les requêtes

5.2. Exp Vs LDA et les autres modèles

Les résultats de la comparaison de notre modèle Exp avec les autres modèles de la littérature sont présentés dans le *Tableau 8* et le *Tableau 9* ci-dessous, sur la base de la mesure NDCG aux seuils 5, 10, 20, 30, 50, 100, 150, 300 et 500. Le NDCG moyen pour toutes les requêtes est également présenté dans le *Tableau 9*.

Query	**JAVA**						**Android**					
NDCG	DBA	EBA	XEBA	BFS	DFS	EXP	DBA	EBA	XEBA	BFS	DFS	EXP
@F5	36,90	35,20	42,00	65,20	68,40	**72,64**	12,10	23,40	27,70	48,70	55,90	**59,83**
@F10	35,60	33,20	39,70	65,20	68,40	**73,23**	13,20	23,20	26,20	49,10	56,10	**59,89**
@F20	32,70	31,70	36,20	62,60	67,00	**72,94**	14,70	21,70	27,20	49,20	54,90	**40,95**
@F30	31,10	30,10	35,70	61,50	64,90	**73,04**	14,70	20,90	26,40	49,30	**62,50**	41,06
@F50	30,40	29,10	34,90	59,70	63,10	**71,89**	15,30	20,70	26,00	51,20	53,90	**41,27**
@F100	31,30	29,20	35,70	59,60	62,70	**72,00**	17,50	23,20	30,40	52,60	55,60	22,81
@F150	31,50	29,80	36,30	62,00	64,70	**70,17**	20,80	25,50	34,00	55,70	57,40	23,86
@F300	35,30	33,80	41,40	67,10	**70,60**	69,88	27,60	31,10	41,00	61,20	62,30	25,13

@F500	40,30	39,20	47,90	68,40	**71,80**	69,89	32,90	37,10	48,00	62,20	63,40	27,49

Tableau 8: Comparaison du modèle Exp avec les autres modèles de base (requête Java et Android)

Query	**C#**						**Mean nDCG**					
NDCG	DBA	EBA	XEBA	BFS	DFS	EXP	DBA	EBA	XEBA	BFS	DFS	EXP
@F5	26,40	30,60	34,00	61,30	64,90	**79,89**	25,13	29,73	34,57	58,40	63,07	**70,79**
@F10	26,40	28,50	32,90	61,50	63,40	**79,89**	25,07	28,30	32,93	58,60	62,63	**71,00**
@F20	25,90	28,40	33,20	59,40	61,10	**77,07**	24,43	27,27	32,20	57,07	61,00	**63,65**
@F30	25,90	28,20	32,50	58,70	59,30	**77,06**	23,90	26,40	31,53	56,50	62,23	**63,72**
@F50	24,40	26,40	30,80	57,50	58,10	**76,99**	23,37	25,40	30,57	56,13	58,37	**63,38**
@F100	23,60	26,20	30,10	57,80	58,60	**76,56**	24,13	26,20	32,07	56,67	**58,97**	57,13
@F150	24,00	25,80	29,80	59,30	60,30	**76,37**	25,43	27,03	33,37	59,00	**60,80**	56,80
@F300	26,80	29,00	34,10	63,20	64,50	**75,46**	29,90	31,30	38,83	63,83	**65,80**	56,83
@F500	30,40	33,00	39,20	67,60	68,50	**74,95**	34,53	36,43	45,03	66,07	**67,90**	57,44

Tableau 9: Comparaison du modèle Exp avec les autres modèles de base (requête C# et moyenne nDCG)

Selon la moyenne NDCG qui permet de calculer le gain cumulé à un rang donné, notre modèle Exp surpasse les autres modèles et cela permet de cumuler une grande quantité de gain à partir des premiers résultats. Une valeur élevée de NDCG dans les premiers seuils signifie que les premiers éléments renvoyés sont des experts pertinents. Néanmoins, nous constatons que le modèle DFS se comporte mieux que Exp pour la requête Android. Un autre point que nous pouvons faire sur la base de ces résultats (la *Figure 21* et la *Figure 22*) est que les performances du score sur les modèles de langues (*Language Model Based Scoring, LMS*) sont nettement inférieures à celles d'autres approches qui exploitent les caractéristiques des utilisateurs et les activités sociales.

Figure 21: Comparaison avec les autres modèles (requêtes Java et C#)

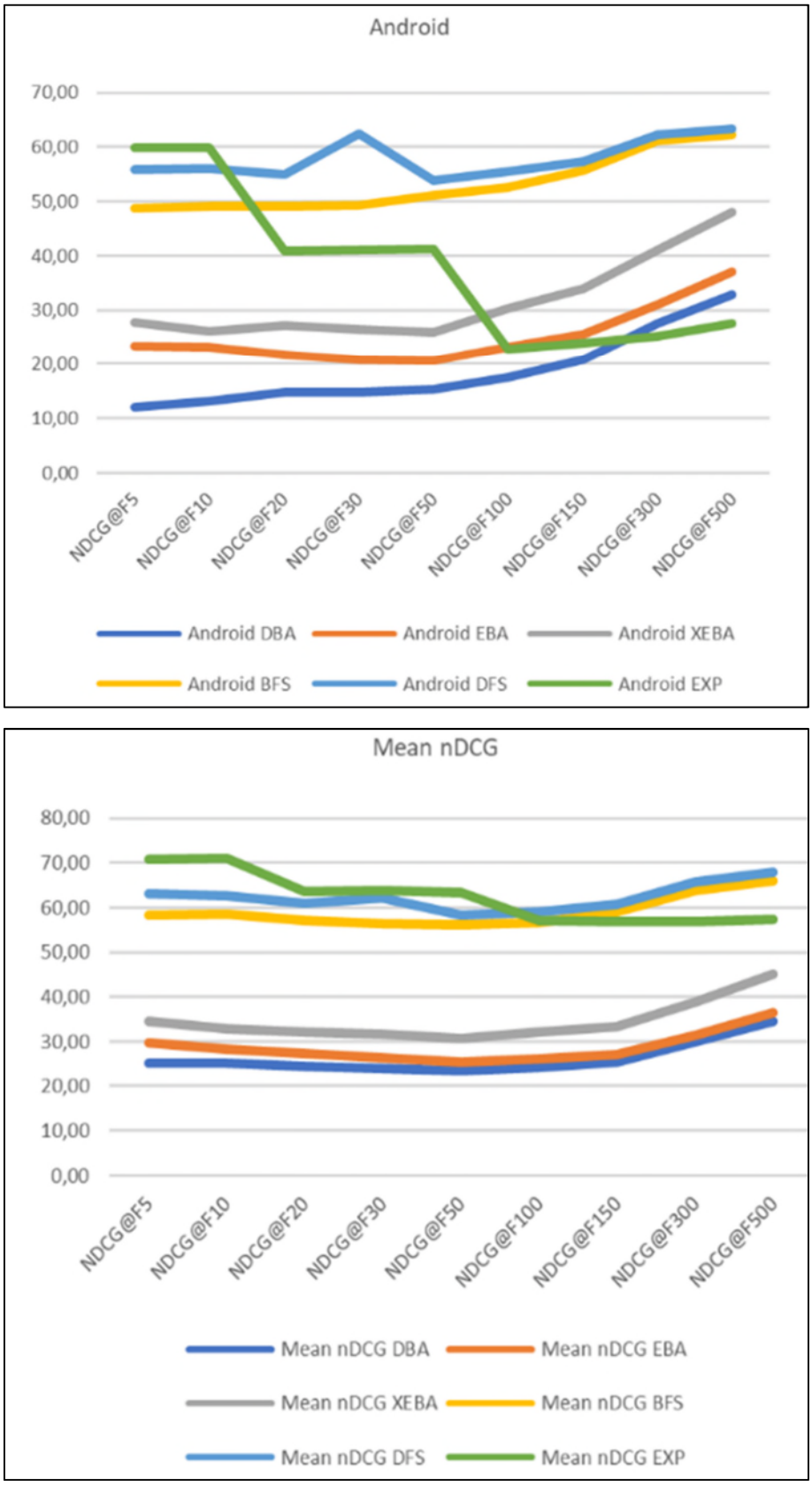


Figure 22 : Comparaison avec les autres modèles (requête Android et moyenne ndcg)

6. Expérimentation sur le dataset Delicious

Delicious est l'un des services de bookmarking social les plus populaires, nous exploitons ce datset dans l'étape du profiling que nous avons proposé. Nous catégorisons les tags des utilisateurs en utilisant LDA, puis nous appliquons le calcul de la profondeur et nous comparons les résultats.

La collection utilisée contient plus de 69 226 URL, tagguées par 1 861 utilisateurs avec plus de 53 000 tags, en réalisant environ 437 000 opérations de tagging.

Après application de LDA sur un ensemble de tags de 100 utilisateurs. Les tags sont catégorisés en topics, c'est le premier output du modèle LDA. Après un certain nombre d'itérations et modification de paramètres, un extrait des résultats est illustré dans la *Figure 23*. Le *topic* est illustré comme une distribution multinomiale de termes, où la probabilité de chaque terme à appartenir au *topic* est calculée. Certains termes tels que les cas du topic 2 et 12 présentent la probabilité 0. Dans ce cas, le modèle n'est pas encore stable, il faut réitérer pour pouvoir corriger l'appartenance des termes aux *topics*.

```
[(0, '0.037*"design" + 0.035*"webdesign" + 0.025*"tool" + 0.020*"web" + 0.019*"inspir" + 0.014*"googl" +
0.014*"tutori" + 0.013*"blog" + 0.011*"css" + 0.010*"refer"'), (1, '0.031*"ux" + 0.031*"ui" +
0.028*"webdesign" + 0.026*"design" + 0.025*"blog" + 0.022*"usabl" + 0.018*"news" + 0.016*"googl" +
0.014*"wikipedia" + 0.012*"mobil"'), (2, '0.000*"viapackratiu" + 0.000*"art" + 0.000*"journal" +
0.000*"educ" + 0.000*"book" + 0.000*"twitter" + 0.000*"tool" + 0.000*"softwar" + 0.000*"polit" +
0.000*"viatwitt"'), (3, '0.036*"architect" + 0.035*"tech" + 0.031*"minfin" + 0.029*"susi" +
0.023*"ilustra" + 0.020*"krista_janicki" + 0.019*"govern" + 0.019*"socialmedia" + 0.017*"photographi" +
0.014*"lio"'), (4, '0.039*"design" + 0.032*"webdesign" + 0.032*"web" + 0.029*"xblog" + 0.023*"tutori" +
0.020*"rubi" + 0.020*"css" + 0.019*"develop" + 0.019*"program" + 0.012*"video"'), (5, '0.027*"educ" +
0.019*"tool" + 0.017*"mobil" + 0.015*"resourc" + 0.014*"siwd" + 0.012*"twitter" + 0.012*"viatwitt" +
0.012*"viapackratiu" + 0.011*"iphon" + 0.011*"socialmedia"'), (6, '0.025*"art" + 0.021*"inform" +
0.017*"data" + 0.013*"tool" + 0.013*"media" + 0.012*"verybetastil" + 0.012*"societi" + 0.012*"design" +
0.012*"music" + 0.011*"human"'), (7, '0.025*"mobil" + 0.021*"ux" + 0.020*"android" +
0.016*"twitter_automatisch" + 0.013*"jqueri" + 0.010*"facebook" + 0.010*"video" + 0.010*"tool" +
0.009*"rubi" + 0.009*"socialmedia"'), (8, '0.123*"inspir" + 0.042*"webdesign" + 0.042*"design" +
0.033*"portfolio" + 0.028*"blog" + 0.022*"photographi" + 0.017*"art" + 0.016*"illustr" +
0.014*"interact" + 0.013*"motion"'), (9, '0.023*"design" + 0.023*"art" + 0.020*"video" + 0.019*"644" +
0.013*"music" + 0.011*"photographi" + 0.011*"refer" + 0.011*"newmedia" + 0.011*"visual" +
0.010*"theori"'), (10, '0.028*"tool" + 0.021*"scienc" + 0.019*"video" + 0.017*"physic" + 0.015*"educ" +
0.011*"softwar" + 0.011*"game" + 0.011*"resourc" + 0.011*"teach" + 0.011*"window"'), (11,
'0.023*"instructifi" + 0.011*"softwar" + 0.008*"game" + 0.008*"ireland" + 0.007*"filter" +
0.007*"privaci" + 0.007*"funni" + 0.006*"write" + 0.006*"linux" + 0.006*"trade"'), (12, '0.000*"design"
+ 0.000*"inspir" + 0.000*"program" + 0.000*"webdesign" + 0.000*"peopl" + 0.000*"languag" + 0.000*"blog"
```

Figure 23 : Exemple de catégorisation de tags en topics

Dans la *Figure 24* sont illustrés les tags du topic 1 avec leur fréquence globale en bleu et la fréquence au sein du topic en rouge. De la même manière, la *Figure 25* illustre le topic 9.

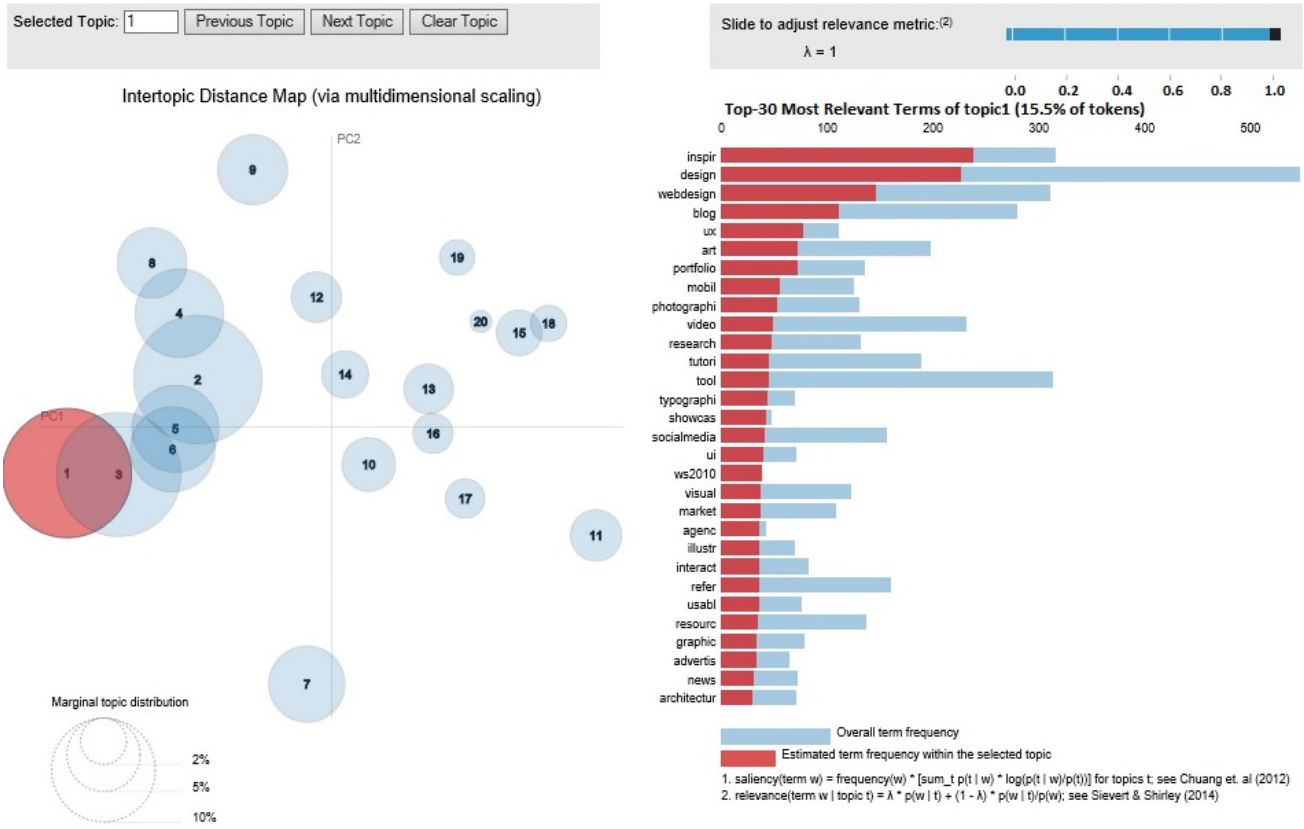


Figure 24: Top-30 des tags pertinents composant le topic 1

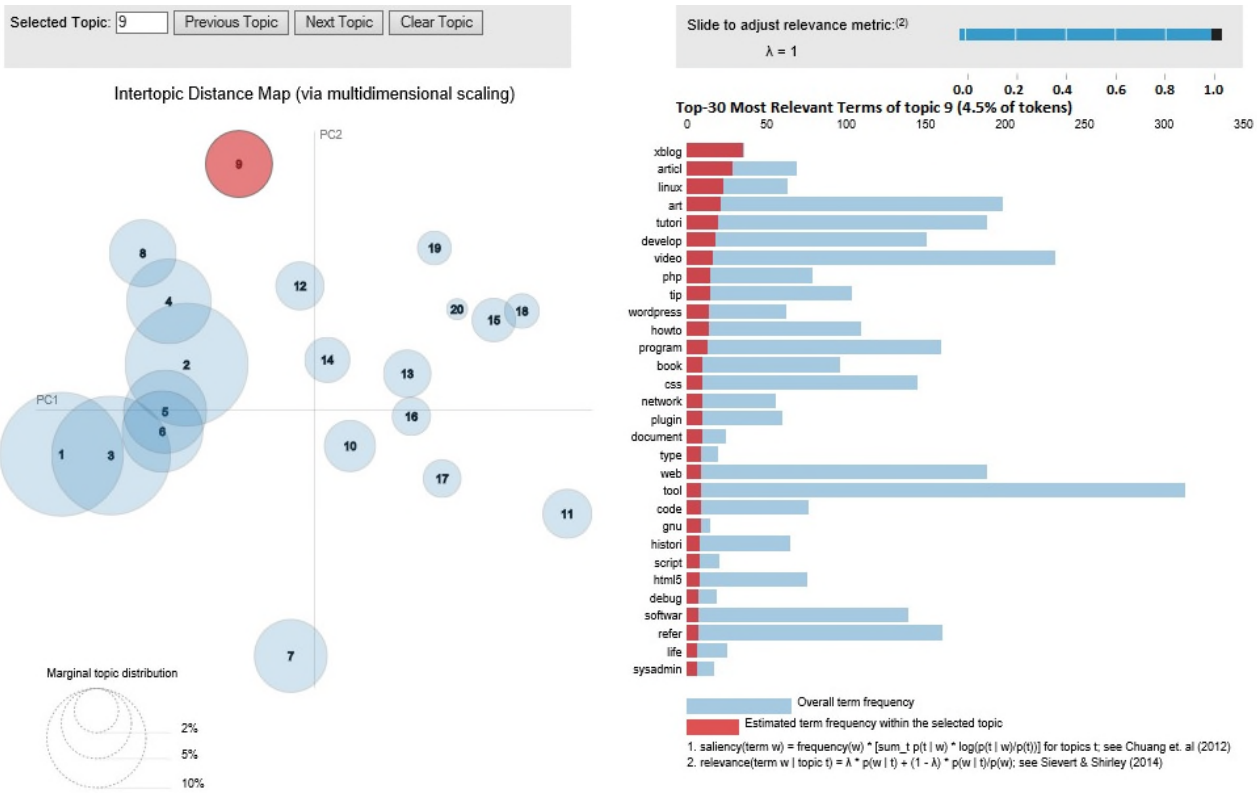


Figure 25 : Top-30 des tags pertinents composant le topic 9

Le deuxième résultat souhaité est la répartition des candidats sur les topics. Dans le *Tableau 10*, nous résumons la répartition des sujets pour certains candidats. Notez que nous conservons 7 à 9 nombres décimaux des fréquences de sujets en raison des légères différences entre les valeurs. Le fait de prendre moins de décimales ne donnera pas de discrimination entre les fréquences des topics.

User	(Topic, fréquence)	User	(Topic, fréquence)
1	(17,0.99394906)	16	(9, 0.9966071)
2	(9,0.9634615)	17	(0, 0.996415)
3	(0, 0.075509354), (14, 0.9223115)	18	(10, 0.9945403)
4	(1,0.9833333)	19	(3, 0.99201685)
5	(9, 0.66755074), (14, 0.3298705)	20	(3, 0.93214285)
….	…….	……	……
30	(4, 0.9959916)	46	(17, 0.994837
31	(4, 0.03315658), (9, 0.815993), (19, 0.14566748)	47	(0, 0.9975388)
32	(18, 0.9952971)	48	(3, 0.013806164), (12, 0.9817603)
33	(12, 0.9981589)	49	(5, 0.99659497)
.	……...	……	……
50	(0, 0.9866198)	73	(9, 0.45957693), (14, 0.06782649),
51	(5, 0.9974666)	74	(10, 0.9919129)
52	(13, 0.98827165)	75	(2, 0.98782057)
53	(5, 0.1732477), (12, 0.12707555), (14, 0.019809633), (17, 0.08017887), (19, 0.59709305)	76	(1, 0.9958515)
……	…….	..	……
92	(9, 0.8871206), (19, 0.1119874)	98	(14, 0.69429475), (17, 0.29185903)
93	(16, 0.99819046)	99	(5, 0.28329068), (7, 0.62867606), (9, 0.084837765)
94	(11, 0.9931655)	100	(1, 0.98602945)

Tableau 10: Un exemple de topics et leurs fréquences attribués à un ensemble d'utilisateurs

Les résultats montrent que les sujets liés à un candidat sont au nombre de 1 à 2. Bien que, pour un candidat, ce nombre dépasse 3 (candidat 53). Ce résultat confirme que les candidats peuvent avoir plusieurs expertises dans différents sujets.

Nous avons utilisé *NLTK (Natural Language ToolKit)*, une plateforme leader pour la création de programmes Python pour travailler avec des données en langage humain. De plus, nous utilisons le package *wordnet* de python, pour exploiter la profondeur du tag. La *Figure 26* montre la variation des profondeurs des tags composant le sujet 2. Les résultats montrent que la profondeur, dans ce cas, varie entre 0 et 8 (les tags recip, visual, ipad ont une profondeur égale à 0). Dans d'autres cas, elle peut atteindre la valeur 14.

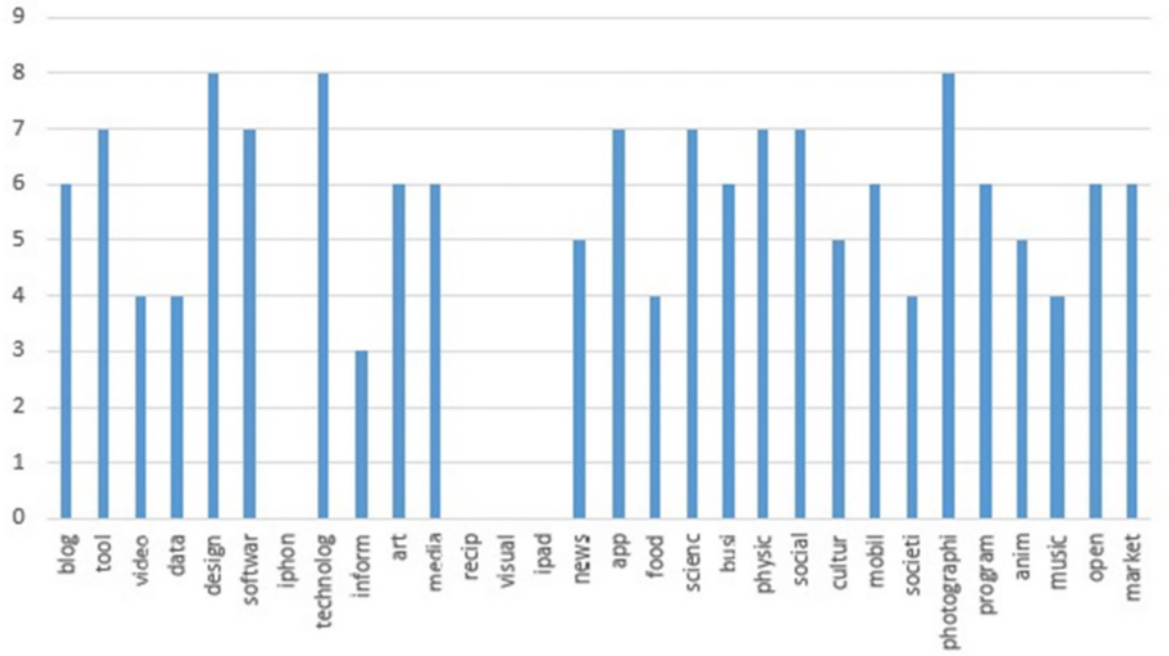


Figure 26: Variation des profondeurs des tags du topic 2

7. Expérimentations sur un dataset de twitter

Nous nous intéressons dans ce cas d'étude aux tweets, en effet *Twitter* est devenu un moyen important de socialisation par lequel les gens communiquent avec le monde et décrivent leurs activités actuelles et leurs opinions dans des textes courts. Les tweets peuvent être analysés automatiquement afin de tirer de nombreuses informations potentielles telles que les intérêts de l'utilisateur, son influence sociale, les communautés auxquelles il appartient et même ses compétences et expertises.

Nos expérimentations consistent à appliquer LDA pour les tweets d'un utilisateur, pour pouvoir détecter ses intérêts thématiques. Nous précisons que LDA et d'autres techniques de modélisation de sujet ont été appliquées pour le cas des tweets dans différents travaux, mais d'une manière à classer les tweets de tous les utilisateurs (tous les tweets de la communauté sans distinction entre les utilisateurs). Ceci est pour des objectifs de veille sociale, politique ou technologique. Nous proposons, dans notre cas, d'enrichir les tweets par les commentaires qui leurs sont liés. Chose qui n'a pas été réalisée dans les différents travaux de la littérature. Ceci est pour ajouter plus d'informations à chaque tweet qui exprime un contenu d'un sujet. Nous avons également proposé d'intégrer une analyse de sentiments sur les tweets afin de capter les avis de l'utilisateur envers les sujets des tweets qu'il publie pour pouvoir par la suite

utiliser ses différentes mentions comme indicateurs utiles pour trouver les intérêts de cet utilisateur en ignorant les tweets ayant des sentiments négatifs.

7.1. Acquisition des données

Afin de pouvoir collecter les données de *Twitter* d'un utilisateur spécifique en temps réel, l'API[10] *Twitter (Streaming*) est utilisée directement pour collecter implicitement les tweets ainsi que leurs commentaires appropriés en utilisant la bibliothèque python *Tweepy*[11]. La Figure 27, présente le processus de collecte de données Twitter.

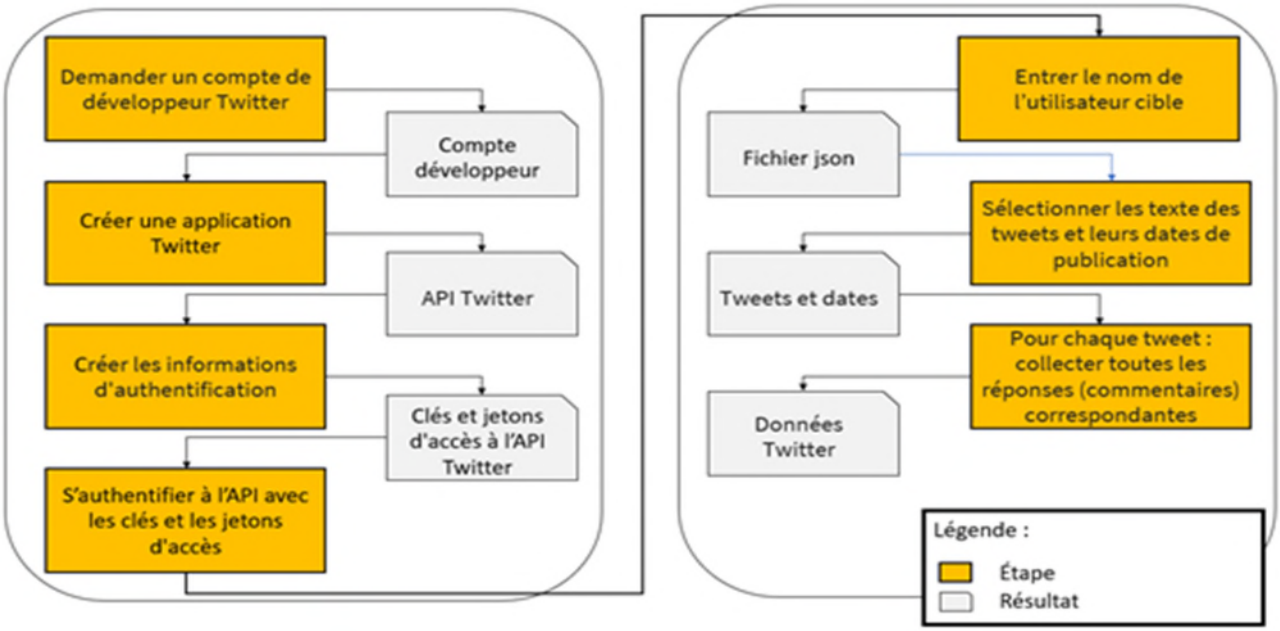


Figure 27: Processus de collecte de données Twitter

Comme le montre ci-dessus, afin de pouvoir collecter les tweets et leurs commentaires, il faut tout d'abord accéder au site des développeurs *Twitter*[12] pour demander un compte développeur. *Twitter* accorde des informations d'authentification aux applications pas aux comptes, donc il faut créer une application pour pouvoir passer des appels API, après avoir généré et copié les clés.

[10] Application Programming Interface

[11] https://www.tweepy.org/

[12] https://developer.twitter.com/en

7.2. Prétraitement des données

Le prétraitement effectué sur l'ensemble des données collectées est le suivant (illustré dans la *Figure 28*) :

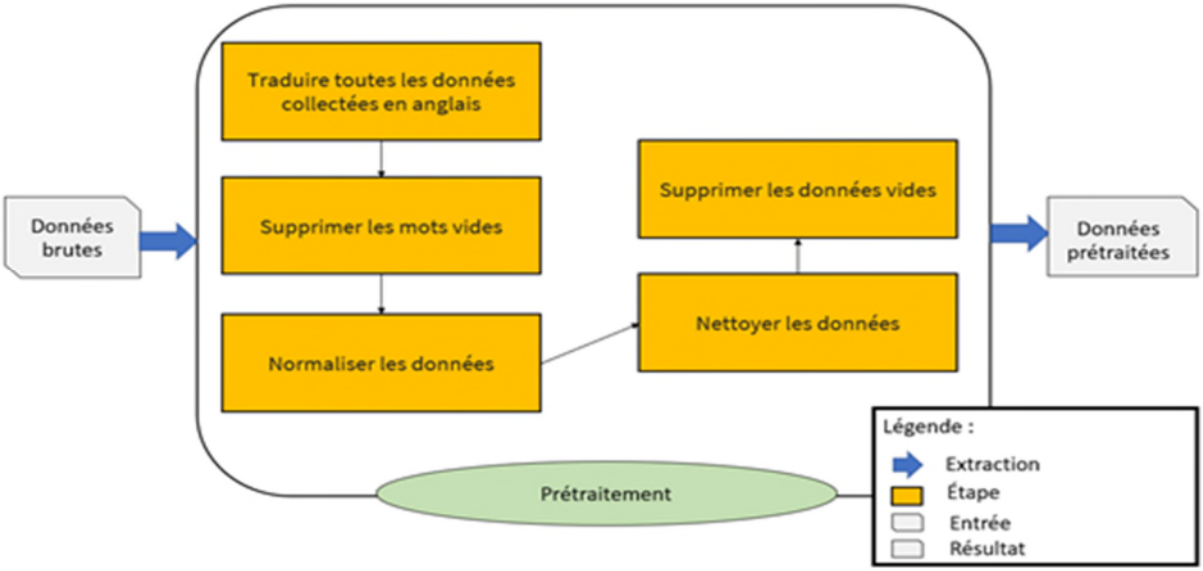


Figure 28: Processus du prétraitement des tweets

La traduction des tweets qui ne sont pas en langue anglaise : la traduction des termes non anglais est effectuée avec la bibliothèque de Python *Googletrans*.

La suppression des mots vides : les mots vides sont supprimés en utilisant la liste des mots vide anglais de *NLTK*, nous avons enrichi cette liste par une autre collection de listes de mots vides trouvée dans *RANKS NL*[13].

La normalisation des termes : Le dictionnaire en ligne *NoSlang*[14] a été utilisé pour normaliser les mots d'argot et les abréviations Internet.

Le nettoyage de données : pour le nettoyage de données de *Twitter*, nous avons conçu un ensemble d'expressions régulières.

L'analyse des sentiments : la bibliothèque *TextBlob* a été utilisée pour calculer la polarité et la subjectivité de chaque tweet.

[13] https://www.ranks.nl/stopwords

[14] https://www.noslang.com/

Les deux principales tâches qui nous intéressent dans l'analyse des sentiments sont la classification de la polarité et la classification de la subjectivité. La classification de la polarité vise à détecter la polarité du tweet parmi trois valeurs possibles : *positif (+)*, *neutre (=)* ou *négatif (-)*. La polarité prend les valeurs -1, 0, 1 respectivement pour la négative, la neutre et la positive. La classification de la subjectivité vise à classer le tweet en '*objectif*' ou '*subjectif*'. La subjectivité varie dans l'intervalle [0,1], plus elle est grande, plus le tweet est considéré '*subjectif*', (Ahuja & Dubey, 2017).

Les intérêts de l'utilisateur sont exprimés par des tweets subjectifs de polarité positive. Par exemple le tweet : « *Je n'aime pas le football* » est considéré subjectif de polarité négative, '*le football*' ne doit pas être considéré parmi les intérêts de l'utilisateur. Le tweet sera négligé. Pour le tweet : « *Le football est un sport collectif* » est un tweet objectif de polarité positive, il peut exprimer l'intérêt de l'utilisateur de manière indirecte.

Les tweets subjectifs ayant une polarité négative ne jouent aucun rôle dans la recherche d'intérêt, leur suppression est donc plus que souhaitable (*Figure 29*).

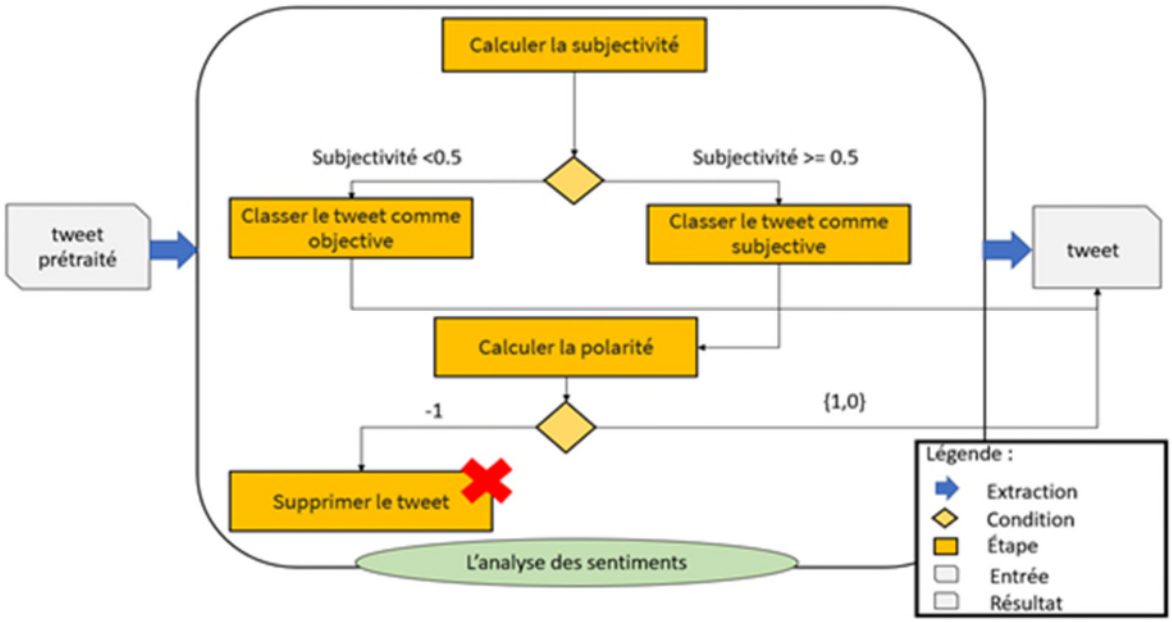


Figure 29: Analyse des sentiments dans les tweets

7.3. Tests et résultats

Nous avons effectué une succession d'expériences sur les étapes de construction du profil de l'utilisateur et discuté les résultats obtenus.

Nous avons choisi de faire les tests sur un échantillon de deux utilisateurs de *Twitter*, avec des tweets couvrant un large éventail de sujets :

« Cuisine et mets »[15] : un compte *Twitter* qui tweete des liens vers des recettes de cuisine uniquement en français.

« Computer Science »[16] : un compte *Twitter* qui tweete quotidiennement sur l'informatique et autres sujets connexes, uniquement en anglais.

7.4. Modélisation thématique

Les étapes de la phase de la modélisation thématique sont les mêmes pour le cas de *Stack Overflow*. La bibliothèque *Gensim* est utilisée pour appliquer la tokenisation et la création du dictionnaire, en outre *Spacy*[17] est utilisée pour la lemmatisation.

Tokenisation : La tokenisation sert à décomposer les données textuelles de chaque document (tweets et commentaires) en une liste de mots atomiques (voir la *Figure 30*).

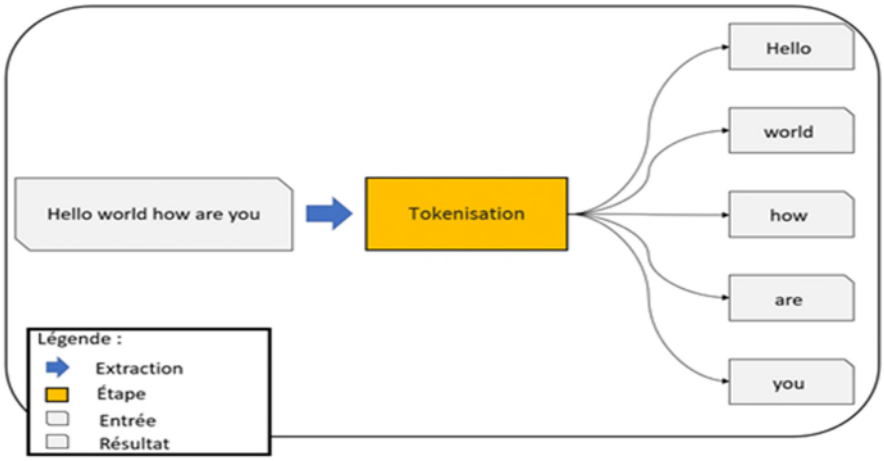


Figure 30: La tokenisation

[15] https://twitter.com/cuisineetmets

[16] https://twitter.com/CompSciFact

[17] https://spacy.io/api/annotation

Lemmatisation : La lemmatisation désigne un traitement lexical apporté à un texte en vue de son analyse. Ce traitement consiste à convertir un mot en sa forme de base canonique (par exemple *est, sois, sut, étais, fussions* … en *être*). Le modèle des bi-grammes construit est utilisé dans la lemmatisation uniquement pour les noms, les adjectifs, les verbes et les adverbes en utilisant le tagger *POS*[18] (Toutanova & Manning, 2000), qui fournit un contexte grammatical pour la lemmatisation, (voir la *Figure 31*).

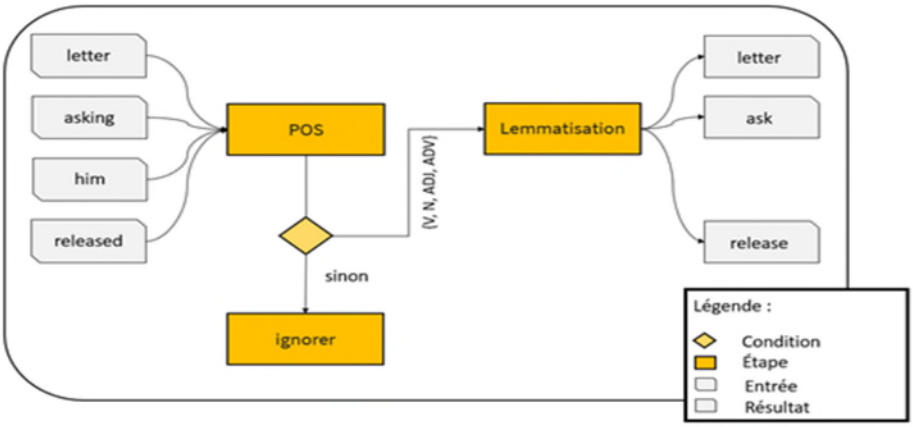


Figure 31: La lemmatisation

Création du dictionnaire : Le dictionnaire permet d'attribuer un identifiant unique à chaque bi-gramme (jeton) de chaque document, comme l'illustre *Figure 32* .

Figure 32: La création du dictionnaire

Afin de générer les distributions thématiques pour chaque utilisateur, nous avons utilisé l'implémentation LDA avec le package *Gensim*. La mesure de cohérence C_V est la technique utilisée pour estimer le nombre de sujets pour chaque modèle.

[18] *Part-Of-Speech*: est la classification de mots d'une langue en categories: nom, verbe, adjectif, etc. La langue anglaise comporte neuf (9) catégories dont 4 principales : nouns, verbs, adjectives, adverbs.

Les rubriques produites et les mots-clés associés, sont examinés visuellement via leurs nuages des mots à l'aide de la bibliothèque *Wordcloud*. Les mots s'affichent dans des tailles et des polices différentes où la taille des mots représente leurs poids. Les figures ci-dessous présentent les distributions thématiques de chaque utilisateur.

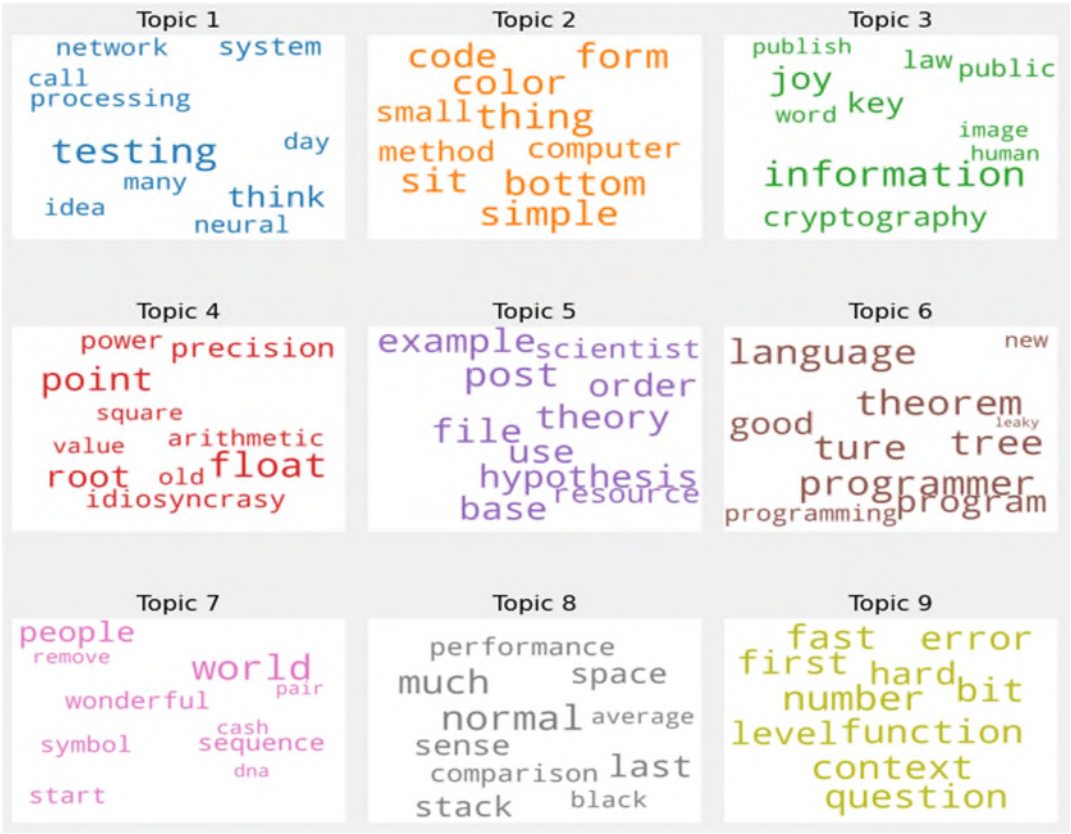


Figure 33: Topics de l'utilisateur Computer Science

Le domaine de l'informatique est clairement mis en évidence à travers les mots-clés des neuf topics extraits du compte « computer science », comme le montre la Figure 33.

Figure 34: Topics de l'utilisateur Cuisine et mets

Le domaine de la cuisine et de la confiserie domine les six sujets d'intérêt de l'utilisateur « Cuisine et mets » comme le montre la Figure 34.

7.5. Évaluation et discussion des résultats

LDA est une méthode non supervisée qui représente chaque document sous forme d'une distribution de sujets, et chaque sujet en une distribution de mots. Le Tableau 11 présente les résultats de nos expérimentations liées à la combinaison ou non de TF-IDF, et à l'intégration ou non des commentaires liés aux tweets considérés :

	Commentaires	Topic Modeling	Cohérence
Cuisine et mets	OUI	LDA	0.9442
		LDA+TF-IDF	**0.9551**
	NON	LDA	0.5832
		LDA+TF-IDF	0.5901
Computer Science	OUI	LDA	0.6633

		LDA+TF-IDF	**0.7421**
	NON	LDA	0.6436
		LDA+TF-IDF	0.6552

Tableau 11: Comparaison des résultats

Les résultats de nos expérimentations pour les deux utilisateurs, montrent que l'enrichissement des tweets avec leurs commentaires donne les meilleures performances par rapport aux approches classiques basées sur les tweets uniquement. En outre, les valeurs de cohérence pour LDA seule et LDA avec TF-IDF sont relativement proches, LDA avec TF-IDF étant supérieure. Ces résultats prouvent que l'application de LDA et TF-IDF sur des tweets enrichis par des commentaires, permet d'obtenir de meilleurs résultats en termes de valeur de cohérence, et cela est dû au fait que les commentaires renforcent le sens des tweets, surtout si ces commentaires sont complètement liés aux tweets. Cependant, l'utilisation directe des commentaires des tweets, sans les analyser et les filtrer, peut affecter les centres d'intérêt de l'utilisateur négativement. Un commentaire spam est un commentaire ayant une signification non liée à celle de tweet. Les spams peuvent être des publicités, des commentaires négatifs qui visent à attaquer l'utilisateur, etc. Des études moyennant l'apprentissage automatique supervisé pour la classification du texte (tweets, commentaires…) en spam et non spam existent dans la littérature.

8. Conclusion

Dans ce travail, nous avons évalué nos propositions de recherche et profilage d'experts, qui a de nouveau émergé dans les communautés en ligne. Nous avons proposé une nouvelle approche dans laquelle des indicateurs sociaux (tags) et leurs profondeurs sont introduits pour juger les expertises des candidats, et une description simple et précise de leurs profils est construite. Nous avons extrait l'intérêt des utilisateurs des données du tagging social et utilisé un algorithme de modélisation de sujets (LDA) pour la distribution des tags sur les sujets. L'algorithme LDA est affiné par l'utilisation de profondeurs des tags afin d'avoir un sujet plus spécifique et plus proche de l'expertise du candidat.

L'approche proposée est appliquée à la collection *Stack Overflow*. Les tests montrent des résultats significatifs, avec des sujets spécifiques et une valeur d'expertise quantifiée pour les candidats. Notre approche filtre les sujets d'intérêt du candidat, et ne donne que ceux dans lesquels le candidat est probablement expert, en fonction de la profondeur du tag. Les valeurs d'expertise, dans notre proposition sont plus significatives et dépassent les modèles de base de l'état de l'art dans plusieurs cas.

D'autres expérimentations ont également été appliquées sur la collection *Delicious*, où il a été question de montrer que les résultats obtenus en appliquant LDA pour la distribution de tags en sujets confirment nos hypothèses.

Nous avons également étudié le cas de *Twitter*, où nous avons pris un échantillon de deux utilisateurs pour lesquels nous avons tenté de construire des profils thématiques en menant un ensemble de tests liés à l'application de LDA. Nous avons d'un côté combiné TF-IDF à LDA et d'un autre côté testé la prise en considération ou non des commentaires liés aux tweets. Dans ce cas également les résultats obtenus sont en faveur de nos propositions.

Le modèle de profiling que nous proposons dans ces travaux a été évalué de manière académique dans ce chapitre. Il est à présent intéressant de voir un cas d'application et ceci en l'instanciant et en l'adaptant au modèle de recommandation des femmes artisans qui est l'un des objectifs du projet de coopération Algéro-Tunisien mené par le CERIST et ses partenaires Tunisiens MIRACL et SOIE. Cette étude fera l'objet du prochain chapitre.

Chapitre VI
Cas d'Étude : Application au projet Algéro-Tunisien « Recommandation des Femmes Artisans »

1. Introduction

La présente étude s'inscrit dans le cadre du projet de coopération Algéro-Tunisien dont l'un des objectifs est de repérer les compétences liées à une activité artisanale particulière et recommander des femmes artisans aux clients susceptibles d'être intéressés par leurs productions. Nous allons présenter un ensemble de propositions ayant pour objectif principal le repérage de ces femmes via l'intégration du tagging social et une meilleure visibilité par la proposition de recommandations.

Un système de recommandation est utilisé pour identifier des ensembles d'éléments susceptibles d'intéresser un utilisateur, en exploitant diverses sources d'informations liées à la fois à l'utilisateur et aux éléments de contenu (Bogers, 2009). Les systèmes de recommandation sociale supposent que les utilisateurs sont corrélés lorsqu'ils établissent des relations sociales (Tang et al., 2013). La recommandation sociale est basée sur des informations sociales ; ces dernières peuvent être collectées par des opérations du tagging social.

Les tags collectés par l'utilisateur font partie de ses préférences ou intérêts (Huang & Lin, 2010), et plus un tag est utilisé par un utilisateur, plus celui-ci est important pour lui, (Cantador et al., 2010). Des recherches ont montré que le tagging social peut être utilisé pour améliorer les systèmes de recommandation (Huang & Lin, 2010).

Lorsque nous évoquons la notion de recommandation dans les réseaux sociaux, on ne peut que penser au réseau Facebook, Viadeo ou encore plus au réseau professionnel LinkedIn. Pour mieux comprendre cette notion, nous citons les différents types que propose LinkedIn :

La recommandation de compétences : L'utilisateur est amené à citer un certain nombre de ses compétences, l'ensemble de ses relations (ses amis et connaissances) auront la possibilité de confirmer. Ils le recommandent donc pour une ou plusieurs compétences.

La recommandation libre : Les amis de l'utilisateur ont la possibilité d'écrire un texte libre pour reconnaitre et témoigner certaines compétences, c'est donc une recommandation libre. Nous avons cité ce type de recommandation dans la section 3.2 du chapitre1, il s'agit d'un type de source d'évidence pour l'expertise appelé *referral web* ou *referral networks*.

La recommandation du système : Dans ce cas c'est le système qui recommande ou suggère à l'utilisateur des personnes qui ont des profils similaires, des compétences et intérêts communs.

Ce qui nous intéresse dans notre cas, c'est cette dernière. Pour ce cas d'étude, l'utilisateur principal est *la femme artisan*. En effet, comme cité dans l'introduction générale, nous avons eu l'occasion de travailler sur le projet portant sur l'amélioration des conditions de travail des femmes artisans en Algérie et en Tunisie. Notre contribution consiste d'abord à intégrer le tagging social, construire les profils à base des tags et proposer une approche de recommandation.

Nous envisageons un système pouvant recommander à la femme artisan d'autres femmes artisans ayant les mêmes intérêts, des clients s'intéressant à sa production et aussi des fournisseurs pouvant lui être utiles. La recommandation pourra aussi se faire dans les autres sens, c'est-à-dire recommander au client et au fournisseur des femmes artisans proches de leurs intérêts. Le système recommande des femmes artisans pour les clients, mais recommande aussi des fournisseurs pour les femmes artisans et des utilisateurs différents les uns pour les autres en se basant sur leurs profils.

Le tagging social permet à des acteurs tels que les utilisateurs finaux, les femmes artisans, les clients et les fournisseurs de tagguer le contenu. Nous prévoyons d'exploiter la dimension sociale du profil des utilisateurs qui en résulte pour recommander principalement des femmes artisans aux clients (Kichou & Meziane, 2015) et (Kichou et al., 2016).

2. Recommandation basée sur le tagging

Dans le domaine des systèmes de recommandation, on distingue trois approches principales de la recommandation de ressources (Puglisi et al., 2015) : (1) le filtrage basé sur le contenu, (2) le filtrage basé sur l'utilisateur et (3) le filtrage collaboratif. En (1), les ressources sont comparées sur la base d'une mesure de similarité. L'hypothèse est que l'individu en question considérerait des ressources similaires à celles qu'un utilisateur a déjà tagguées dans le passé comme plus pertinentes. Si en fait, un utilisateur a taggué des ressources dans certaines catégories avec plus de fréquence, il est plus probable qu'il annoterait également des éléments appartenant aux mêmes catégories. Dans (2), les utilisateurs sont comparés à d'autres utilisateurs sur la base d'une mesure de similarité définie. On suppose, dans ce cas, que si deux ou plusieurs utilisateurs ont des intérêts similaires, c'est-à-dire qu'ils ont exprimé leur préférence pour des ressources de catégories similaires ; les ressources utiles pour l'un d'entre eux peuvent également être importantes pour les autres. (3) Le filtrage collaboratif utilise à la fois une combinaison des techniques décrites précédemment ainsi que les actions collectives d'un groupe ou réseau d'utilisateurs et leurs relations sociales.

Dans (BellogíN et al., 2013), les auteurs citent des recommandations sociales, appelées souvent approches hybrides du moment qu'elles combinent des approches de filtrage collaboratif et des informations sociales.

Afin d'améliorer les systèmes de recommandation, plusieurs travaux ont été proposés dans la littérature. Dans (Diederich & Iofciu, 2006), les auteurs utilisent la "Personomie" (voir section 4 définition3 du chapitre 4) de l'utilisateur, qui représente les tags utilisateur, et les ressources taguées pour recommander les utilisateurs qui ont ajouté des tags et des ressources similaires. Tout d'abord, un profil utilisateur est créé. Ensuite, à partir de ce profil, le système est en mesure de recommander des utilisateurs en utilisant une mesure de similarité entre eux.

Par ailleurs, dans (Hu, 2012), les auteurs se sont basés à la fois sur l'historique du tagging des utilisateurs (tags et ressources) et sur leurs contacts sociaux. Sa principale limitation est qu'il nécessite des contacts sociaux existants de l'utilisateur afin d'avoir des recommandations sociales.

(Hotho et al., 2006) ont proposé l'algorithme *FolkRank* qui effectue un classement personnalisé des résultats basés sur la folksonomie et une recommandation d'utilisateurs, de tags et de ressources.

Les auteurs de (Niwa et al., 2006; Shepitsen et al., 2008) ont proposé une recommandation avec le regroupement de tags pour minimiser la redondance des informations et contextualiser les recommandations de ressources.

(Jelassi et al., 2013) a considéré une nouvelle dimension dans les folksonomies comme une information supplémentaire pour fournir aux utilisateurs une recommandation plus ciblée et mieux adaptée à leurs besoins, cette quatrième dimension peut couvrir différents aspects : par exemple, le profil (sexe, âge, profession, idoles… etc…). Abel dans (Abel et al., 2011) représente le profil basé sur les tags comme un ensemble de tags pondérées pour le profil utilisateur inter-systèmes, ainsi que (Firan et al., 2007) crée un profil utilisateur basé sur des tags utilisés pour une meilleure recommandation musicale sur *Last.fm*, basée sur un fonction logarithme.

Le système de recommandation aide les utilisateurs à prendre les décisions appropriées au cours de leurs activités professionnelles. De même, il leur apporte une assistance automatique et des opportunités personnalisées tout en économisant le temps et les efforts.

3. Apport du tagging dans le cas des femmes artisans

Permettre aux utilisateurs finaux, femmes artisans, clients, et fournisseurs de tagguer des contenus ou des acteurs pourra d'une part enrichir d'avantages les descriptions initiales de ceux-ci, d'autre part cet ensemble de tags manipulé par un utilisateur donné jouera le rôle d'un indicateur très important sur les préférences de celui-ci.

Dans notre cas nous définissons deux types de tagging :

- Le Tagging (*User- User*) consiste à associer un ou plusieurs mots clés (tags) aux différents acteurs (F-artisan, fournisseur, client);
- Le Tagging (*User- Product*) consiste à associer un ou plusieurs mots clés pour décrire les produits existants.

En utilisant le Tagging (*User- User*), la femme artisan, le fournisseur et le client seront décrits par un ensemble de tags classés par ordre décroissant de leurs poids, c-à-d un vecteur pondéré de mots clés (qu'on appellera Réputation), voir la Figure 35. Exemple : Femme artisan *i* est décrite par le vecteur {(excellente, 16), (bon-travail, 12), (bonne-finition, 10)}. Précisons ici que dans ce type de tagging et dans un but d'éviter toute sorte d'impolitesse ou autres choses qui peuvent nuire aux utilisateurs (puisqu'il s'agit de juger le travail ou le comportement d'une personne), ceux-ci sont guidés par un ensemble de tags bien défini (tagging automatique).

Ces ensembles de tags nous permettront de connaitre ce que pensent les autres utilisateurs de celui-ci.

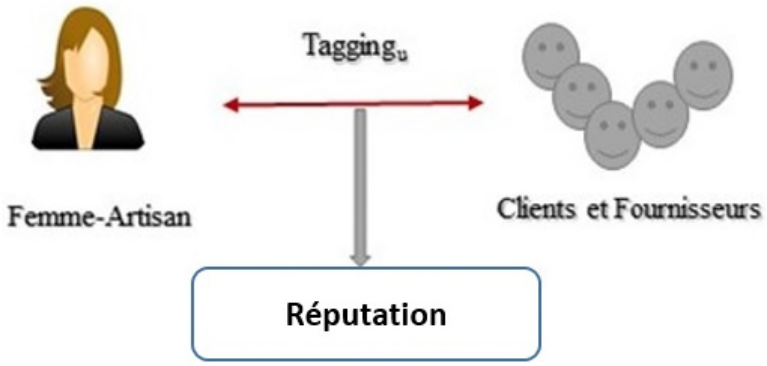


Figure 35: Le Tagging User-User

L'application du deuxième type de tagging (Tagging (*User- Product*)), permet d'une part d'avoir un descripteur (vecteur pondéré de tags) du produit (qu'on appellera *Descripteur-produit*), donc enrichir sa description pour entre autres faciliter sa recherche et le rendre plus visible à la communauté. D'une autre part, ce type de tagging nous permettra de savoir à quels produits s'intéressent les utilisateurs, et ceci en repérant leurs intérêts qu'on appellera Intérêt-Social, voir la Figure 36.

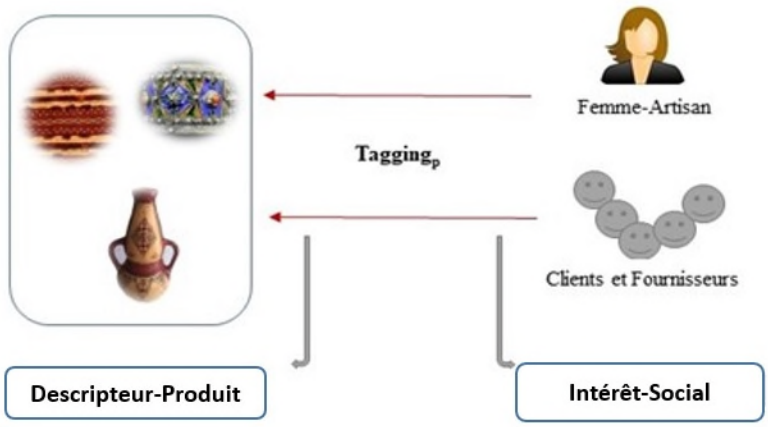


Figure 36: Le Tagging User-Product

Une première étape primordiale est la définition d'un modèle du profil utilisateur. Dans ce cas des femmes artisans, le modèle du profil de celle-ci a été défini par les équipes de recherche algérienne et tunisienne sous forme ontologique. Au moment de cette définition, il n'était pas question d'intégrer le tagging social et proposer une approche de recommandation, choses qui sont venues bien après l'avancement du projet et le souhait d'exploiter les avantages du tagging est bien compréhensible. Nous avons donc mis en place quelques extensions pour pouvoir mettre en œuvre nos propositions.

4. Modélisation du profil

4.1. Représentation du profil de la femme artisan

La définition du profil utilisateur pour une application donnée revient à sélectionner les dimensions jugées utiles, le tableau suivant montre l'ensemble des concepts définis pour la description du profil (structure ontologique proposée dans le projet pour pouvoir effectuer d'éventuels raisonnements liés à l'adaptation dynamique de la solution) pour représenter les informations de la femme artisan, auxquels nous rajoutons les concepts : Intérêt-Social, Tag, et Expertise dans le domaine (illustré avec une couleur différente), que nous allons expliquer plus loin.

Concepts	**Nom**	**Description**
Artisane	Craftswoman	Artisane
Activité	Activity	Métier de l'artisane

Diplôme	Diploma	Diplômes obtenus
Académique	Academic	Diplômes universitaires
Professionnel	Professional	Diplômes professionnels
Parent	Parent	Personnes ayant un lien parental avec l'artisane
Capacité et Compétence	AbilityAndProficiency	Capacités et compétences de l'artisane
Capacité	Ability	Capacités de l'artisane
Capacité-cognitive	Cognitive-Ability	Niveau cognitif de l'artisane
Capacité-Physique	Physical-Ability	Capacités physique
Capacité-Sensorielle	Sensory-Ability	Capacités sensorielles
Capacité-Auditive	AbilityToHear	Capacité auditive
Capacité- Visuelle	AbilityToSee	Capacité visuelle
Capacité-Vocale	AbilityToTalk	Capacité vocale
Capacité-de-Toucher	AbilityToTouch	Capacité de toucher
Habileté-Motrice	Motor-Ability	Habileté motrice
Compétence	Proficiency	Connaissances et les compétences de l'artisane
Compétence-Informatique	Computer-Literacy	Compétences et le niveau de connaissances en informatique (utilisation d'un ordinateur)
Langue-Maitrisée	Language	Les langages maitrisés en lecture, écriture, et les Langues parlées (dialectes inclus)
Préférence-liée-Interface	Preference-Related-Interface	Préférences d'interaction de l'artisane avec l'interface du système
Mode-Entrée	Input-Mode	Mode d'interaction en entrée (manipulation directe ou vocale)
Mode-Sortie	Output-Mode	Mode d'interaction en sortie (image, voix, texte…)
Personnalité	Personality	Personnalité et caractère de l'artisane (accepte de vendre sur Internet, accepte de participer à travailler en groupe….)
Intérêts	Interest	Les aspects à améliorer afin de faciliter l'activité productive de l'artisane
Intérêt-Outillage	Tool	Son intérêt à améliorer tout ce qui est lié aux outillages
Intérêt-Processus-Production	Production-Process	intérêt à améliorer le processus de production

Intérêt-Relation-Fournisseur	Provider-Relationship	Intérêt à améliorer la relation avec les fournisseurs.
Intérêt-Relation-Client	Customer- Relationship	Intérêt à améliorer la relation avec les clients.
Intérêt-Communication	Communication	Intérêt à améliorer les outils de communication, de l'informatique et de l'Internet
Mode d'Organisation et de Coordination	Coordination-And-Organization-Mode	Moyens de coordination et d'organisation utilisés par l'artisane dans le processus de production et de vente
Mode d'Organisation	Coordination-Mode	Moyens pour organiser son activité
Mode de Coordination	Organization-Mode	Moyens pour interagir avec les différents acteurs
Intérêt Social	*Centres d'intérêt issus du tagging*	*Les centres d'intérêt extraits du tagging (user-produit).*
Réputation	*Ce que pensent les autres de lui*	*L'ensemble des tags issus du tagging (user-user)*
Tag	*Tag, mot-clé*	*Ce sont les différents tags qui représentent les centres d'intérêt*
Expertise	*Expertise*	*Expertise de la femme artisan dans son domaine, traduite par un nombre.*

Tableau 12: Dictionnaire des concepts du profil de la femme artisan complété par les 4 nouveaux concepts.

4.1.1. Le Concept Intérêts-Social

Le concept Intérêts-Social englobe les centres d'intérêt *Int* (u_i) qui nous renseignent sur les intérêts et préférences de l'utilisateur à base des tags qu'il a associés aux différents produits via le Tagging *User-Product*. Cet ensemble de tags pondérés, est construit en utilisant l'approche hybride (Naïve-par co-occurrence que nous allons détailler dans la section 5.1.1).

$$\boldsymbol{int(u_i) = \{(t_1, w_1), (t_2, w_2), \ldots . (t_j, w_j)\}}$$

avec t_i le tag d'indice i et w_i le poids de ce tag. *Exemple : Int(ui) = {(céramique,5), (tapisserie, 3), (poterie, 2)}.*

4.1.2. Le Concept Réputation

Le Concept Réputation nous renseigne sur ce que pense l'ensemble des utilisateurs d'un utilisateur donné, résultat du Tagging *User-User*. C'est un ensemble de tags pondérés par leurs fréquences d'apparition.

4.1.3. Le Concept Expertise

Comme déjà présenté, les utilisateurs experts dans un domaine, ont tendance à utiliser des termes spécifiques pour tagguer vu qu'ils ont une parfaite maîtrise des concepts de ce domaine. Ce concept nous renseigne sur le degré de maitrise de l'utilisateur dans le domaine des produits taggués. Dans ce cas précis, l'expertise est calculée en fonction du niveau des tags cités par l'utilisateur dans l'ontologie du domaine de l'artisanat. Par exemple les tags *marne*, *silice* et *argile* sont considérés plus profonds (donc plus précis) que le tag *terre* dans le domaine de la poterie. Il faut également noter que dans ce cas d'étude l'expertise est unique puisqu'il s'agit d'un seul domaine particulier (une activité de l'artisanat).

4.1.4. Les Attributs

Les attributs des nouveaux concepts sont illustrés dans le Tableau 13 :

Concepts	**Synonymes**	**Propriétés (attributs)**
Intérêt-Social	Intérêt	Name
Réputation	Réputation	Name
Tag	Mot-clé	Label du tag Poids du tag
Expertise	Expertise	Valeur de l'Expertise

Tableau 13: Les attributs des nouveaux concepts

4.1.5. Les Relations

Les relations entre les concepts du profil de la femme Artisan sont représentées dans le Tableau 14 ci-dessous.

Relation	**Concept source**	**Concept cible**	**Cardinalité**
hasIntérêt-Social	Artisane	Intérêt-Social	(0.1)
hasRéputation	Artisane	Réputation	(0.1)
hasExpertise	Artisane	Expertise	(1.1)

hasTag	Intérêt-Social	Tag	(0.n)
hasTag	Réputation	Tag	(0.n)
hasPoids	Tag	Poids	(1.1)

Tableau 14: Les relations binaires des trois concepts du profil de la femme artisan

4.2. Représentation du profil utilisateur (Fournisseur, Client)

Jusque-là, les autres types d'utilisateurs (fournisseurs et clients) ne sont pas pris en considération dans la structure précédente, nous allons donc proposer une autre structure ontologique pour contenir leurs informations personnelles. Dans le Tableau 15 ci-dessous la liste des concepts du profil User.

Concepts	**Nom**	**Description**
User	User	Fournisseur, Client
Activity	Job, occupation	Travail de l'utilisateur
Intérêt-Général	Intérêt	Les centres d'intérêt introduits par l'utilisateur
Intérêt-Social	Centres d'intérêt	Les centres d'intérêt (tags) issus via le tagging(user-produit)
Réputation	Ce que pensent les autres de lui.	L'ensemble des tags issus du tagging (user-user)

Tableau 15: Les concepts du profil User

Les attributs associés aux concepts du profil *user* sont illustrés dans le Tableau 16 suivant :

Concepts	**Synonymes**	**Propriétés (attributs)**
User	User	Name, Birthday, Birthplace, email, Address, Country, Mobile-Number
Activity	Job ; Occupation	Name, Description, Location
Interest		Name, Description

Intérêt-Social	Intérêt	Name
Réputation	Réputation	Name
Tag	Tag	Label du tag Valeur poids

Tableau 16 : Les attributs des concepts du profil User

Les relations entre les classes du profil User sont représentées dans le tableau suivant.

Relation (***Object Properties***)	**Concept source**	**Concept cible**	**Cardinalité**
hasActivity	User	Activity	(0.n)
hasInterest	User	Interest	(0.n)
hasIntérêt-Social	User	Intérêt-Social	(0.1)
hasRéputation	User	Réputation	(0.1)
hasTag	Intérêt-Social	Tag	(0.n)
hasTag	Réputation	Tag	(0.n)
hasPoids	Tag	Poids	(1.1)

Tableau 17: Les relations binaires des concepts du profil User

5. Approche de repérage du profil des femmes artisans

À la suite d'un besoin de recherche exprimé par une femme artisan ou un autre utilisateur, nous aurons à repérer un profil bien précis. Exemple : un client qui cherche des femmes artisans expérimentées dans un domaine donné ; une femme artisan qui cherche des clients intéressés par son type de production, etc. Notre but est donc de faciliter ces recherches en donnant une sorte de miroir reflétant les intérêts des utilisateurs, les expertises des femmes artisans et les différentes descriptions données par leur entourage, et ceci de manière dynamique vu que tout peut changer en fonction du temps.

5.1. Acquisition des centres d'intérêt

À partir de tous les tags de l'utilisateur, un graphe est construit, nous adoptons l'approche hybride présentée dans le chapitre 4. Les nœuds représentent les tags et les arcs les relations de co-occurrence entre ceux-ci. Les nœuds sont pondérés avec le nombre de fois que l'utilisateur a utilisé le tag (popularité), et les arcs avec le nombre de co-occurrence de chaque paire de tags, comme le montre la Figure 37.

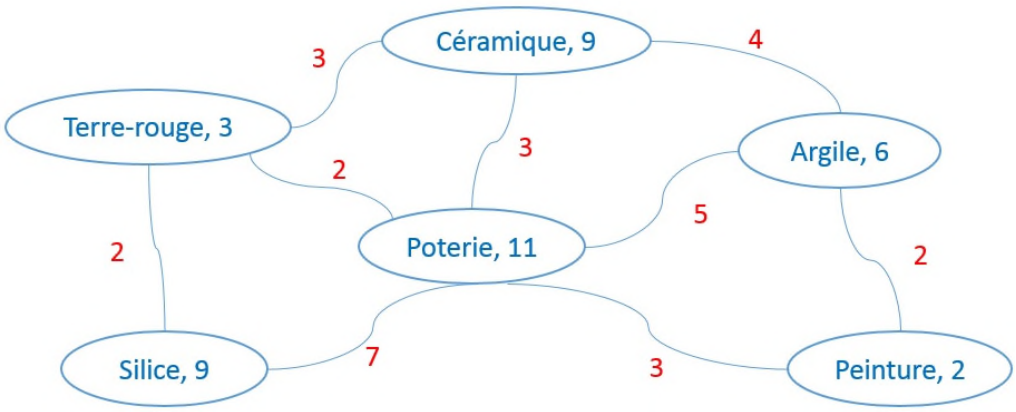


Figure 37: Graphe Construit avec l'approche hybride

Dans l'exemple illustré dans le graphe précédent, le vecteur sera : {(poterie,11), (Silice,9), (Argile,6), (Céramique,9)}.

5.2. Acquisition de l'expertise

Nous mettons l'hypothèse que le calcul de l'expertise concerne uniquement les femmes artisans. Celle-ci est supposée être candidate à l'expertise dans son domaine d'activité, la valeur de l'expertise est donc un scalaire qui concerne un seul domaine (Kichou et al., 2011).

Lorsque nous parlons d'une expertise, nous évoquons un domaine donné, si un tag n'appartient pas au domaine de la femme artisan, il sera ignoré dans le calcul de l'expertise. Afin de localiser les niveaux des tags utilisés pour le calcul de l'expertise, nous allons utiliser l'ontologie du domaine de l'artisanat conçu pour ce projet ainsi que l'ontologie métier.

La profondeur du tag (considéré comme le nœud feuille) est calculée par rapport à la profondeur de la racine, *Root* qui représente le concept le plus général englobant tous les concepts.

Le calcul de la profondeur d'un tag revient donc à calculer la profondeur du concept c correspondant : *Depth(c)*, sachant que *Depth(root)*, la profondeur du *root* est égale à *0*. Cette notion est donc exprimée par la formule suivante :

$$Expertise(U_i) = \frac{\sum Depth(t_j)}{|T_u|} \qquad (21)$$

Où *Depth* (t_j), la profondeur du tag t_j, est le nombre de nœuds le séparant de la racine (*root*) ; T_u est un sous-ensemble de la personomie[19] de l'utilisateur u_i contenant les tags que celui-ci a associé aux produits. Pour le calcul de l'expertise on s'intéresse au Tagging User-Produit qui permet de tagguer les produits.

Reprenons l'exemple du concept 'terre' cité précédemment : si on considère que le concept 'matière première' est la racine, alors *Depth (matière_première)=0*, le concept 'terre' vient juste en dessous, donc Depth (terre)=1, Depth (argile)=2 , Depth *(argile_verte)=3, Depth (argile_rouge)=3* (car argile verte et argile rouge sont au même niveau).

Un utilisateur u_1 citant les tags suivants pour différents produits : argile et terre, possède une expertise notée :

expertise(u1)= (depth(argile)+depth(terre))/2= ***1,5***.

Un utilisateur u_2 citant les tags suivant pour différents produits : argile verte et argile rouge, a une expertise notée :

*expertise(u2) = (depth(argile_verte)+depth(argile_rouge))/2=****3***.

5.2. Acquisition de la réputation

Les tags composant la Réputation sont issus des différents autres utilisateurs, dans ce cas l'approche naïve sera très adéquate car on veut juste savoir ce que pensent les autres sans différentiation entre tag générique ou spécifique. D'autant plus, que le tagging dans ce cas est automatique (liste prédéfinie de tags).

Le résultat sera donc une liste de tags classés par ordre décroissant de leurs fréquences d'apparition appelée popularité. Un exemple est illustré dans la Figure 38.

[19] Rappelons que la Personomie concerne un utilisateur donné u, défini comme $P_u = (R_u, T_u, A_u)$, où A_u est une projection des tuples de A sur T et R pour l'utilisateur u.

Figure 38: Exemple d'une Réputation d'un fournisseur

5.3. Enrichissement de la description du produit

Comme résultat du Tagging user-produit, on retrouve des descripteurs des produits enrichissant les descriptions déjà données par les auteurs (femmes artisans). Le descripteur du produit est un vecteur de tags pondérés classés par ordre décroissant de popularités.

Pour la mise en œuvre de notre approche et la représentation du descripteur du produit, il est nécessaire de compléter l'ontologie métier par les concepts suivants, comme c'était le cas avec les autres ontologies du profil. Le concept ajouté est présenté dans le Tableau 18 ci-dessous :

Concepts	**Nom**	**Description**
Descripteur-Produit	Descripteur-Produit	Un ensemble de tags pondérés décrivant le produit.

Tableau 18: Extension de l'ontologie métier

Les attributs du nouveau concept sont présentés dans le Tableau 19 ci-dessous :

Concepts	**Synonymes**	**Propriétés (attributs)**
Descripteur-Produit	Descripteur-Produit	Name
Tag	Tag, mot-clé	Label du tag Poids du tag

Tableau 19: Attributs du concept Descripteur-Produit

Les relations entre les concepts sont présentées dans le Tableau 20 ci-dessous :

Relation (*Object Properties)*	**Concept source**	**Concept cible**	**Cardinalité**
hasDescripteur-Produit	Produit	Descripteur-Produit	(0.1)
hasTag	Descripteur-Produit	Tag	(0.n)
hasPoids	Tag	Poids	(1.1)

Tableau 20: Les relations ajoutées pour l'ontologie Métier

6. Recommandation basée sur les intérêts de l'utilisateur

Les opérations du tagging social aident les femmes artisans à se mettre en contact avec d'autres communautés de femmes artisans, de clients et / ou de fournisseurs. En plus, elles améliorent leurs activités en les rendant ainsi que leurs produits visibles pour les autres.

Pour recommander une femme artisan au client, la distance entre le vecteur d'intérêt client *int (u)* et le vecteur d'intérêt de la femme artisan *int (a)* est calculée, résultant du tagging (*user-product*) et reflétant les intérêts du client ou de la femme artisan. Nous avons choisi la mesure de similarité cosinus : cosinus (*Int (u), Int (a)*). Cette dernière est la plus utilisée dans les modèles de l'espace vectoriel. L'originalité n'est pas l'utilisation de cette métrique, mais l'utilisation de vecteurs d'intérêt basés sur les tags.

Les résultats sont classés par ordre décroissant de cosinus, et les 5 premiers sont les plus proches de la femme artisan, ils sont donc recommandés. (5 est un choix qui peut être modifié).

La recommandation dans l'autre sens, c'est-à-dire recommander des clients à des femmes artisans (dans le but de les contacter pour faire de la publicité par exemple) est également prévue.

Une autre direction de recommandation importante est de recommander des fournisseurs aux femmes artisans. Dans ce cas, le vecteur d'intérêt ne suffit pas, un tri sur le descripteur de fournisseur (à partir du tagging *user-user*) est ajouté, qui reflète ce que les autres pensent de cet utilisateur, tel que *{(bon, 10), (sérieux, 4) …}* afin que le fournisseur soit dans le domaine artisanal et ait une bonne «réputation» dans le

réseau. A cet effet, un vecteur idéal « Réputation-idéale» est proposé, contenant par exemple au moins les termes suivants: sérieux, honnête et bon. Pour être appliqué, le tagging (*user-user*) sera utilisé avec les tags suggérés (tagging automatique).

6.1. Calcul de similarité

Nous adoptons la mesure de similarité cosinus donnée par la formule suivante :

$$\boldsymbol{cos}\left(\overrightarrow{\boldsymbol{Interet(u1)}}, \overrightarrow{\boldsymbol{Interet(u2)}}\right) = \sum_{i=1,k} \frac{wi,I1}{\sqrt{\sum_{i=1,k} wi,I1^2}} \frac{wi,I2}{\sqrt{\sum_{i=1,k} wi,I2^2}} \qquad (22)$$

Avec $\overrightarrow{Interest(u1)}$ et $\overrightarrow{Interest(u2)}$ vecteur d'intérêts respectivement des utilisateurs u_1 et u_2. $W_{i,I1}$ est le poids du tag d'index i du vecteur $\overrightarrow{Interest(u1)}$ et $W_{i,I2}$ le poids du tag d'index i du vecteur $\overrightarrow{Interest(u2)}$.

Plus la valeur cosinus est grande, plus l'angle entre les deux vecteurs est petit et donc les deux utilisateurs sont proches.

Recommandation de la femme artisan pour les clients

Dans ce cas, la similarité entre les vecteurs d'intérêt pour les clients et les femmes artisans est calculée, le résultat est classé par ordre décroissant du cosinus pour un client donné.

Recommandation des différents utilisateurs

Dans ce cas, le cosinus entre les vecteurs d'intérêts des utilisateurs est calculé (deux à deux). Les tests montrent que les utilisateurs sont considérés comme similaires lorsque le cosinus varie entre 0,42 et 0,81, ces utilisateurs sont recommandés les uns aux autres. Les valeurs au-delà de 0,85 sont rares.

Recommandation du fournisseur à la femme artisan

La différence entre la recommandation client et celle du fournisseur est que le vecteur d'intérêt du fournisseur ne suffit pas pour prendre une décision, sur sa recommandation, pour la femme artisan. Le fournisseur peut être intéressé par des produits sans les fournir, ou même sans être un bon fournisseur. Pour cela, les étapes suivantes sont proposées pour chaque femme artisan :

1. Trouver des fournisseurs dont l'activité est proche de celle de la femme artisan (fournir ce dont la femme artisan a besoin), ceci avec une simple question sur leurs informations personnelles, et élimination de ceux dont l'activité est très différente.
2. Vérifier la « Réputation » de ceux-ci au sein de la communauté, et ce en comparant sa Réputation (du tagging (*user-user*)) à la Réputation initiale définie, en calculant le degré de similarité entre eux (cosinus).
3. Tri des fournisseurs par ordre décroissant des valeurs du cosinus.
4. Recommander les 5 meilleurs fournisseurs.

Il faut noter qu'en plus de ces critères, le critère localisation doit aussi être pris en considération lors de la recommandation (favoriser les utilisateurs proches du point de vue localisation géographique).

7.Expérimentation

Afin de tester notre approche de recommandation des femmes artisans, nous avons choisi un jeu de donnée créé en interne au sein du projet. Des tests préliminaires sont menés sur un ensemble de femmes artisans, des clients et des fournisseurs, qui tagguent un ensemble de produits et s'identifient mutuellement à l'aide de mots libres.

7.1 Acquisition des données

L'objectif principal du projet des femmes artisans, est de mettre à leur disposition une plateforme qui puisse leur permettre de mettre en ligne leurs produits, être en contact entre elles, entre elles et les clients et les fournisseurs. Comme une première étape et avant la mise en ligne de la plateforme, nous avons réalisé un jeu de tests sur des données fictives introduites à cet effet. La Figure 39 illustre la page d'accueil de la plateforme réalisée dans le cadre du projet.

Figure 39: Accueil de la plateforme

7.2. Tests et résultats

L'approche hybride est appliquée pour extraire les vecteurs d'intérêt des utilisateurs et choisir les cinq premiers tags classés en fonction de leur poids. Il faut noter que ce vecteur des cinq valeurs est dynamique et peut changer dans le temps à travers l'utilisation du système. Les tests sont effectués avec un prototype réalisé avec Delphi.

Les similarités sont calculées entre le client U_0: *{(ceramic, 10), (pottery,6), (cray,4), (polishing,4), (earth,3)},* et les femmes artisans (Tableau 21), puis entre différents utilisateurs (Tableau 22).

Dans ce qui suit, les intérêts qui en résultent sont les vecteurs de certaines femmes artisans. Nous citons juste quelques femmes artisans, les tableaux montrent les résultats pour 10 femmes.

-Interest (u_1) = {(painting-silk, 12), (silk,8), (embroidery,4), (ceramic,3), (pottery,2)} ;

- Interest (u_2) = {(ceramic,8),(painting,7),(jar,5),(pottery,4),(polishing,3)};

- Interest (u_3) = {(sewing,10),(embroidery,5),(painting,3),(silk,3),(pottery,1)};

-Interest(u_4)={(embroidery,11),(sewing-thread,6),(sew,5),(dress,2), (sewing,2)};

-Interest(u_5)={(pottery,14),(polishing,11),(ceramic,10),(clay,3),(baked-earth,1)} ;

- Interest (u_6)={(carpet,12),(wool,8),(dye,4),(loom,3), (polyster,3)} ;

-Interest (u_7)={(carpentry,9),(cabinetmaker,7),(wood,5), (smooth, 4), (furniture,2)} ;

Le calcul de similarité entre le client *U0= {(ceramic, 10), (pottery,6), (cray,4), (polishing,4), (earth,3)}* et les femmes artisans citées donnent les résultats suivants :

User	**Cos (U_0,U_i)**
U_1	**0.20**
U_2	**0.682**
U_3	**0.037**
U_4	**0**
U_5	**0.87**
U_6	**0**
U_7	**0**
U_8	**0**
U_9	**0.689**
U_{10}	**0.78**

Tableau 21: Résultats du calcul de similarité pour la recommandation des femmes artisans

Dans une première analyse des résultats obtenus et en comparaison avec des utilisateurs réels et en s'appuyant sur un ensemble de tests plus importants, nous en déduisons ce qui suit :

- Les tests montrent que les utilisateurs avec un score cosinus inférieur à 0,4 ne sont pas recommandés, car ils ne sont pas considérés comme suffisamment proches au client.

- Les vecteurs avec plusieurs tags en commun ayant des poids proches sont considérés comme les plus similaires. Dans ce cas, la femme artisan U_5 présente 4 des 5 tags en commun avec le client ayant des poids approximativement proches, est considérée comme la femme artisan la plus similaire avec un score de 0,87. La femme artisan U_{10} est la deuxième plus proche du client avec un score de 0,78. U_9 arrive en 3ème position avec 0,689 ayant les mêmes tags que le client mais avec un poids très différents.

La liste des femmes artisans recommandées pour le client *U0* est : *U5, U10, U9, U2*.

Le calcul de similarité entre les différents utilisateurs donne les résultats suivants :

User	**U_1**	**U_2**	**U_3**	**U_4**	**U_5**	**U_6**	**U_7**	**U_8**	**U_9**	**U_{10}**
U_1	1	0.16	0.24	0.20	0.18	0	0	0	0.08	**0.13**
U_2	0.16	1	0.16	0	**0.64**	0	0	0	**0.42**	**0.36**
U_3	0.24	0.16	1	**0.45**	0.05	0	0	0.3	0.02	**0.05**
U_4	0.20	0	**0.45**	1	0	0	0	**0.58**	0	**0**
U_5	0.18	**0.64**	0.05	0	1	0	0	0	**0.81**	**0.68**
U_6	0	0	0	0	0	1	0	0	0	**0**
U_7	0	0	0	0	0	0	1	0	0	**0**
U_8	0	0	0.3	**0.58**	0	0	0	1	0	**0**
U_9	0.08	**0.42**	0.02	0	**0.81**	0	0	0	1	**0.51**
U_{10}	**0.13**	**0.36**	**0.05**	**0**	**0.68**	**0**	**0**	**0**	**0.51**	**1**

Tableau 22: Résultats du calcul de similarité entre différents utilisateurs

Les valeurs en gras montrent des similarités assez importantes entre les utilisateurs en fonction de la valeur obtenue, d'où la recommandation de ces utilisateurs entre eux (recommander par exemple *U_5* à *U_9, U_5 à U_{10}* et vice versa, etc).

7.3. Synthèse

L'analyse économique montre que les entreprises visent à garder leurs clients et de promouvoir les relations avec eux pour augmenter leurs ventes et leurs affaires. Les réseaux sociaux sont rationnellement utilisés pas uniquement pour promouvoir leurs produits mais aussi pour valoriser la marque, et permettre des relations solides avec les consommateurs ainsi que pour améliorer la qualité de service et des produits grâce aux feedbacks (commentaires, avis, tags…) du marché lui-même.

Le premier objectif de cette étude est d'aider les femmes artisans à communiquer entre elles et entrer en contact avec les clients et fournisseurs. Le second consiste à extraire un profil dynamique pour mieux prendre en compte les changements de préférences au fil du temps. Avec une finalité de proposer aux utilisateurs une recommandation basée sur leurs opérations de tagging, ce qui est une sorte de recommandation sociale. L'approche proposée est évaluée avec des tests préliminaires.

8. Conclusion

Dans ce travail, nous avons présenté une manière d'intégrer et d'exploiter les activités du tagging social au profit des femmes artisans. Ceci a été effectué dans le cadre du projet de coopération Algéro-Tunisien visant à promouvoir le commerce, l'échange et l'approvisionnement de ces femmes. Il vise aussi à améliorer leur repérage via la découverte de leurs intérêts et compétences.

Le premier objectif est d'aider les femmes artisans à communiquer entre elles et à entrer en contact avec les clients et les fournisseurs. Le second consiste à extraire un profil précis et dynamique pour mieux prendre en compte les changements de préférences au fil du temps. Avec une finalité de proposer des recommandations aux utilisateurs en fonction de leurs opérations de tagging, ce qui est une sorte de recommandation sociale.

Cette étude est une application du modèle de profiling que nous avons proposé dans cette thèse. En effet, l'estimation de l'expertise pour ce cas d'étude considère un seul domaine alors que dans le modèle initial plusieurs domaines peuvent être envisagés. De la même manière que dans le cas de ce modèle de recommandation, notre modèle de profiling et de recherche d'expert peut être appliqué et adapté à d'autres domaines. Notre manière d'estimer l'expertise d'un utilisateur par rapport à un domaine donné en exploitant les tags peut également être d'un apport considérable dans une multitude d'applications.

Conclusion Générale

It is possible to fly without motors, but not without knowledge and skill.

Wilbur Wright

À ces débuts, la recherche d'information s'est intéressée à la recherche de documents adéquats à un besoin particulier. Le besoin de la recherche des personnes a été senti par les systèmes commerciaux. Chercher une personne peut être d'ordre personnel ou professionnel. Dans ce travail, nous nous sommes intéressés au repérage des professionnels utilisant le critère de la compétence ou de l'expertise.

La recherche d'expertise est considérée comme une branche de la recherche d'information. Elle regroupe deux tâches principales : la recherche d'experts dont l'objectif est de répondre à la question : quels sont les experts dans une thématique particulière X ?; et le profilage d'expert dont l'objectif est de répondre à la question : quelles sont les thématiques d'expertise d'une personne Y ? Les deux tâches sont considérées comme les deux facettes d'une même pièce, en effet leur finalité est de localiser l'expert dans une thématique donnée.

L'intérêt pour ce domaine s'est accentué avec l'explosion des réseaux sociaux, ceci a permis d'explorer de nouvelles sources de données pouvant s'informer davantage sur les éventuelles compétences d'un candidat. Ces sources, qui étaient, au début, limitées aux rapports techniques, productions scientifiques du candidat, se sont élargies aux différentes activités sociales. Cela dit, même les organisations ont élargi leurs investigations sur le web communautaire utilisant plusieurs types de sources d'évidence et de méthodes.

La problématique de localisation ou de modélisation de l'expertise demeure d'actualité. En effet, les études tentent d'améliorer le processus de localisation en explorant de nouvelles sources d'évidence. Notre étude de l'état de l'art a révélé certaines limites concernant les travaux proposés. Les limites concernent en particulier les sources d'évidences utilisées qui sont généralement représentées par la production scientifique des candidats. Celle-ci est certainement une source crédible, mais peut ne

pas concerner tous les candidats notamment lorsque l'investigation dépasse le domaine scientifique et académique pour explorer le web communautaire grand public. Ce dernier peut être très fructueux du point de vue compétences, pour ceci nous nous sommes intéressés à l'étude des activités sociales des candidats, considérée comme très prometteuse. Nous nous sommes intéressés particulièrement à l'utilisation des activités du tagging et des tweets comme indicateur de l'expertise et d'intérêts.

Après exploration du domaine de la recherche d'expertise et ses deux principales tâches à savoir la recherche d'experts et le profilage d'expert, nous avons présenté une étude de l'état de l'art. Nous avons exploré les sous-domaines liés à la recherche d'expertise et présenté les notions fondamentales à savoir la notion de compétence, et ses relations avec la notion d'expertise. Sans oublier de parler de la compétence en entreprise et sur le web. Par la suite, nous avons présenté l'état de l'art lié aux travaux les plus importants du domaine. La dernière notion que nous avons évoquée est celle de la modélisation de sujets ou thématiques (*topic modeling*) vu sa forte corrélation au domaine de la recherche d'expertise. Comme nous avons utilisé les activités du tagging comme indicateur principal de l'expertise, la notion du tagging social a aussi été évoquée.

Dans ces travaux de recherche, nous avons présenté une nouvelle manière d'estimer l'expertise d'un utilisateur en se basant sur son activité de tagging et en exploitant une caractéristique sémantique des tags qu'est la profondeur. Une technique du *topic modeling* (*Latent Dirichlet Allocation (LDA)*) pour déduire les *topics* d'expertise des utilisateurs avec le niveau de chaque expertise est également exploitée. Partant du principe que plus un tag est profond plus il est précis et plus son utilisation démontre du niveau d'expertise de l'utilisateur l'ayant utilisé, nous avons construit un modèle de profiling et un modèle de recherche d'experts issu du modèle vectoriel de la recherche d'information.

Le modèle que nous avons proposé a été doublement validé : premièrement, une validation académique sous forme d'expérimentations sur plusieurs datasets avec une comparaison des résultats avec ceux des autres modèles proposés dans le domaine. Les résultats obtenus sont très satisfaisants, démontrant ainsi de l'efficacité de notre modèle et du net apport de notre approche d'estimation de l'expertise au domaine de la recherche d'experts. Deuxièmement, le modèle proposé a été instancié à un cas

pratique concernant la recommandation des femmes artisans, ce qui fait une seconde validation démontrant l'applicabilité du modèle et encore une fois son efficacité.

Finalement, nous pouvons résumer les contributions apportées dans ces travaux dans les points suivants :

- Proposition d'un nouveau modèle d'estimation de l'expertise en se basant sur les activités du candidat liées au tagging social ;
- Études des cas sur les réseaux sociaux Delicious et *Twitter* pour l'obtention d'un profil thématique de l'utilisateur ;
- Intégration des activités du tagging social pour le cas des femmes artisans ;
- Proposition d'une approche de recommandation basée sur le profil utilisateur pour le cas des femmes artisans.

Dans nos travaux futurs, nous planifions d'étendre nos propositions ainsi que les expérimentations en se focalisant sur un certain nombre de points. Nous voulons évaluer l'approche de recommandation sur des données réelles (utilisant les données de la plateforme des femmes artisans conçue dans le projet de coopération Algéro-Tunisien). En effet, ceci nous permettra d'avoir de larges échantillons de données et nous pourrions à travers celles-ci proposer une approche de génération du vecteur de réputation du fournisseur et son évaluation. Par ailleurs, et dans le modèle d'estimation de l'expertise à base des tags, nous prévoyons de tester une technique d'apprentissage automatique pour une meilleure prédiction pour le cas des candidats qui n'ont pas eu encore d'activité sociale. Car pour ces candidats, les dimensions sociales sont vides, une prédiction basée sur ces relations sociales ou son activité antérieur serait intéressante. Nous planifions également de combiner notre modèle à un modèle utilisant la production scientifique ou autres types de documents rédigés comme source d'évidence. Ceci est pour voir le comportement du modèle vis-à-vis de l'amélioration du processus d'évaluation de l'expertise.

L'un des points intéressants, est de tester d'autres versions de LDA et d'autres techniques du *topic modeling*, pour en déduire celle apportant la meilleure classification d'experts. Mais aussi, élargir l'échantillon de test sur *Twitter* pour la proposition du profil thématique à base des tweets, et la proposition d'une distinction entre les topics d'intérêts et ceux d'expertise.

Références

Abel, F., Araújo, S., Gao, Q., & Houben, G.-J. (2011). Analyzing cross-system user modeling on the social web. *International Conference on Web Engineering*, 28–43.

Ahuja, S., & Dubey, G. (2017). Clustering and sentiment analysis on Twitter data. *2017 2nd International Conference on Telecommunication and Networks (TEL-NET)*, 1–5.

Al-Barakati, A., & Daud, A. (2018). Venue-influence language models for expert finding in bibliometric networks. *International Journal on Semantic Web and Information Systems (IJSWIS)*, *14*(3), 184–201.

AlSumait, L., Barbará, D., Gentle, J., & Domeniconi, C. (2009). Topic significance ranking of LDA generative models. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 67–82.

Al-Taie, M. Z., Kadry, S., & Obasa, A. I. (2018). Understanding expert finding systems : Domains and techniques. *Social Network Analysis and Mining*, *8*(1), 57.

Amato, F., Cozzolino, G., & Sperlì, G. (2019). A hypergraph data model for expert-finding in multimedia social networks. *Information*, *10*(6), 183.

Balog, K., Azzopardi, L., & de Rijke, M. (2009). A language modeling framework for expert finding. *Information Processing & Management*, *45*(1), 1–19.

Balog, K., Azzopardi, L., & De Rijke, M. (2006). Formal models for expert finding in enterprise corpora. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 43–50.

Balog, K., & De Rijke, M. (2007). Determining Expert Profiles (With an Application to Expert Finding). *IJCAI*, *7*, 2657–2662.

Balog, K., Fang, Y., De Rijke, M., Serdyukov, P., & Si, L. (2012). Expertise retrieval. *Foundations and Trends in Information Retrieval*, *6*(2–3), 127–256.

Balog, K., Serdyukov, P., & de Vries, A. P. (2011). Overview of the TREC 2011 Entity Track. *TREC*, *2011*, 11.

BellogíN, A., Cantador, I., & Castells, P. (2013). A comparative study of heterogeneous item recommendations in social systems. *Information Sciences*, *221*, 142–169.

Berendsen, R., Rijke, M., Balog, K., Bogers, T., & Bosch, A. (2013). On the assessment of expertise profiles. *Journal of the American Society for Information Science and Technology*, *64*(10), 2024–2044.

Bietti, A. (2012). *Latent Dirichlet Allocation*. mai 2012, working paper, http://alberto. bietti. me/files/rapport-lda. pdf.

Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, *57*(2), 1–30.

Blei, D. M., Griffiths, T. L., Jordan, M. I., Tenenbaum, J. B., & others. (2003). Hierarchical topic models and the nested Chinese restaurant process. Advances in neural information systems *NIPS*, vol *16, 17-24*.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd international conference on Machine learning*, 113–120.

Blei, D. M., Lafferty, J. D., & others. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, *1*(1), 17–35.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Bogers, A. M. (2009). *Recommender systems for social bookmarking*. Doctoral dissertation, Tilburg University, Netherlands.

Boughanem, M., Kraaij, W., & Nie, J.-Y. (2004). Modeles de langue pour la recherche d'information. *Les systemes de recherche d'informations*, 163–182.

Brill, E., & Moore, R. C. (2000). An improved error model for noisy channel spelling correction. *Proceedings of the 38th annual meeting of the association for computational linguistics*, 286–293.

Broudoux, E. (2006). Folksonomies et indexation collaborative. Rôle des réseaux sociaux dans la fabrique de l'information. DocForum, vol 24.

Brusilovsky, P., Smyth, B., & Shapira, B. (2018). Social search. In *Social Information Access* (p. 213–276). Springer.

Buckley, C., & Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 25–32.

Budura, A., Bourges-Waldegg, D., & Riordan, J. (2009). Deriving expertise profiles from tags. *Computational Science and Engineering, 2009. CSE'09. International Conference on*, *4*, 34–41.

Büttcher, S., Clarke, C. L., & Cormack, G. V. (2016). *Information retrieval : Implementing and evaluating search engines*. Mit Press.

Campbell, C. S., Maglio, P. P., Cozzi, A., & Dom, B. (2003). Expertise identification using email communications. *Proceedings of the twelfth international conference on Information and knowledge management*, 528–531.

Campion, M. A., Guerrero, L., & Posthuma, R. (2011). Reasonable human resource practices for making employee downsizing decisions. *Organizational Dynamics*, *40*(3), 174–180.

Cantador, I., Bellogín, A., & Vallet, D. (2010). Content-based recommendation in social tagging systems. *Proceedings of the fourth ACM conference on Recommender systems*, 237–240.

Carchiolo, V., Longheu, A., Malgeri, M., & Mangioni, G. (2015). Searching for experts in a context-aware recommendation network. *Computers in Human Behavior*, *51*, 1086–1091.

Cayzer, S., & Michlmayr, E. (2009). Adaptive User Profiles. *Collaborative and Social Information Retrieval and Access: Techniques for Improved User Modeling*. *IGI Global, 65-87.*

Chang, J., & Blei, D. (2009). Relational topic models for document networks. *Artificial Intelligence and Statistics*, 81–88.

Cifariello, P., Ferragina, P., & Ponza, M. (2019). Wiser : A semantic approach for expert finding in academia based on entity linking. *Information Systems*, *82*, 1–16.

Daud, A., Li, J., Zhou, L., & Muhammad, F. (2010). Temporal expert finding through generalized time topic modeling. *Knowledge-Based Systems*, *23*(6), 615–625.

De Vocht, L., Softic, S., Verborgh, R., Mannens, E., & Ebner, M. (2017). Social semantic search : A case study on web 2.0 for science. *International Journal on Semantic Web and Information Systems (IJSWIS)*, *13*(4), 155–180.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391–407.

Dehghan, M., & Abin, A. A. (2019). Translations diversification for expert finding : A novel clustering-based approach. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *13*(3), 1–20.

Dehghan, M., Biabani, M., & Abin, A. A. (2019). Temporal expert profiling : With an application to T-shaped expert finding. *Information Processing & Management*, *56*(3), 1067–1079.

Diederich, J., & Iofciu, T. (2006). Finding communities of practice from user profiles based on folksonomies. *Innovative approaches for learning and knowledge sharing, ec-tel workshop proc*, 288–297.

Dominich, S. (2008). *The Modern Algebra of Information Retrieval*. Heidelberg: Springer. (Vol. 24), 74-93.

Faggioli, G., Polato, M., & Aiolli, F. (2019). Tag-based user profiling : A game theoretic approach. *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, 267–271.

Faisal, M. S., Daud, A., Akram, A. U., Abbasi, R. A., Aljohani, N. R., & Mehmood, I. (2019). Expert ranking techniques for online rated forums. *Computers in Human Behavior*, *100*, 168–176.

Fang, H., & Zhai, C. (2007). Probabilistic models for expert finding. *European conference on information retrieval*, 418–430.

Fang, Y., & Godavarthy, A. (2014). Modeling the dynamics of personal expertise. *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 1107–1110.

Fazel-Zarandi, M. (2013). Representing and reasoning about skills and competencies over time. Doctoral dissertation, University of Toronto, Canada.

Fazel-Zarandi, M., & Fox, M. S. (2011). Constructing expert profiles over time for skills management and expert finding. *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, 1-6.

Fellbaum, C., & Vossen, P. (2012). Challenges for a multilingual wordnet. *Language Resources and Evaluation*, *46*(2), 313–326.

Ferracci, A. B. (2012). Évaluation comparative de l'expertise psychologique et psychiatrique: vers une méthodologie systématique de l'évolution. Doctoral dissertation, Université Toulouse le Mirail-Toulouse II, France.

Firan, C. S., Nejdl, W., & Paiu, R. (2007). The benefit of using tag-based profiles. *2007 Latin American Web Conference (LA-WEB 2007)*, 32–41.

Font, F., Serra, J., & Serra, X. (2013). Folksonomy-based tag recommendation for collaborative tagging systems. *International Journal on Semantic Web and Information Systems (IJSWIS)*, *9*(2), 1–30.

Foulds, J., Kumar, S., & Getoor, L. (2015). Latent topic networks : A versatile probabilistic programming framework for topic models. *International Conference on Machine Learning*, 777–786.

Fu, Y., Xiang, R., Liu, Y., Zhang, M., & Ma, S. (2007). Finding experts using social network analysis. *IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)*, 77–80.

Gharebagh, S. S., Rostami, P., & Neshati, M. (2018). T-shaped mining : A novel approach to talent finding for agile software teams. *European conference on information retrieval*, 411–423.

Golder, S., & Huberman, B. A. (2005). The structure of collaborative tagging systems. *arXiv preprint cs/0508082*.

Guy, M., & Tonkin, E. (2006). Folksonomies : Tidying up tags? *D-lib Magazine*, *12*(1).

Hertzum, M., & Pejtersen, A. M. (2000). The information-seeking practices of engineers : Searching for documents as well as for people. *Information Processing & Management*, *36*(5), 761–778.

Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 50–57.

Hotho, A., Jäschke, R., Schmitz, C., & Stumme, G. (2006). *Folkrank : A ranking algorithm for folksonomies*. Workshop on Information Retrieval of the Special Interest Group Information Retrieval, 111-114.

Hu, Y. (2012). A music recommendation system based on user behaviors and genre classification. Doctoral dissertation, University of Miami, USA.

Huang, C.-L., & Lin, C.-W. (2010). Collaborative and content-based recommender system for social bookmarking website. *International Journal of Computer and Information Engineering*, *4*(8), 1310–1315.

Jelassi, M. N., Ben Yahia, S., & Mephu Nguifo, E. (2013). A personalized recommender system based on users' information in folksonomies. *Proceedings of the 22nd International Conference on World Wide Web*, 1215–1224.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet Allocation (LDA) and Topic modeling : Models, applications, a survey. *Multimedia Tools and Applications*, *78*(11), 15169–15211.

Jiao, J., Yan, J., Zhao, H., & Fan, W. (2009). Expertrank : An expert user ranking algorithm in online communities. *New Trends in Information and Service Science, 2009. NISS'09. International Conference on*, 674–679.

Jukic, M. R., & Huljenic, D. (2007). New Approach to Competence Management in Telecom Industry. *2007 9th International Conference on Telecommunications*, 103–108.

Kardan, A., Omidvar, A., & Farahmandnia, F. (2011). Expert finding on social network with link analysis approach. *2011 19th Iranian Conference on Electrical Engineering*, 1–6.

Kasper, G., de Siqueira Braga, D., Martins, D. M. L., & Hellingrath, B. (2017). User profile acquisition : A comprehensive framework to support personal information agents. *2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, 1–6.

Kavitha, V., Manju, G., & Geetha, T. (2014). Learning to rank experts using combination of multiple features of expertise. *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on*, 1053–1058.

Kichou, S., Boussaid, O., & Meziane, A. (2020). Tag's Depth-Based Expert Profiling Using a Topic Modeling Technique. *International Journal on Semantic Web and Information Systems (IJSWIS)*, *16*(4), 81–99.

Kichou, S., Mellah, H., Amghar, Y., & Dahak, F. (2011). Tags weighting based on user profile. *International Conference on Active Media Technology*, p 206–216.

Kichou, S., Mellah, H., Boussaid, O., & Meziane, A. (2016). Handicraft women Recommendation Approach based on User's Social Tagging Operations. *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 618–621.

Kichou, S., Mellah, H., Lasbeur, I., & Abdelouahid, I. (2013). *User interest extraction based on weighted tags*. In Proceedings of the 2nd International Workshop on Web Intelligence (WEBI-2013), 22-31

Kichou, S., & Meziane, A. (2015). *User Profile Extraction Based on Social Tagging*.

Case Study: Handicrafts Women in Emerging Countries, Conférence sur les avancées des Systèmes Décisionnels, ASD'2015, 103-113.

Le Moigne, M. (2007). *Les compétences des collectivités territoriales en droit public français : Essai de compréhension d'une structure complexe*. Doctoral dissertation, university of Brest, France.

Liang, S. (2018, April). Dynamic user profiling for streams of short texts. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1), 5860-5867.

Lin, S., Hong, W., Wang, D., & Li, T. (2017). A survey on expert finding techniques. *Journal of Intelligent Information Systems*, *49*(2), 255–279.

Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, *5*(1), 1-22.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mech. Transl. Comput. Linguistics*, *11*(1- 2), 22–31.

Macdonald, C., & Ounis, I. (2006). Voting for candidates : Adapting data fusion techniques for an expert search task. *Proceedings of the 15th ACM international conference on Information and knowledge management*, 387–396.

Marrelli, A. F. (1998). An introduction to competency analysis and modeling. *Performance Improvement*, *37*(5), 8–17.

Mathes, A. (2004). Folksonomies-cooperative classification and communication through shared metadata, Computer Mediated Communication. In LIS590CMC (Doctoral Seminar), Graduate School of Library and Information Science, University of Illinois Urbana-Champaign (Dec. 2004).

Mcdonald, D. W., & Ackerman, M. S. (1998). Just Talk to Me : A Field Study of Expertise Location. *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work (CSCW '98)*, 1- 11. https://doi.org/doi: 10.1145/289444.289506

Momtazi, S., & Naumann, F. (2013). Topic modeling for expert finding using latent Dirichlet allocation. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(5), 346–353.

MOREAU, C. (2010). *La gestion des compétences: un défi pour l'avenir du volontariat en secteur associatif. Mémoire en sciences de gestion soutenu sous la direction de M. Frédéric Schoenaers, 110p*.

Moreira, C., Calado, P., & Martins, B. (2011). Learning to rank for expert search in digital libraries of academic publications. *Portuguese Conference on Artificial Intelligence*, 431–445.

Neshati, M., Fallahnejad, Z., & Beigy, H. (2017). On dynamicity of expert finding in community question answering. *Information Processing & Management*, *53*(5), 1026–1042.

Nie, J. (1988). An outline of a general model for information retrieval systems. *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, 495–506.

Niwa, S., Doi, T., & Honiden, S. (2006). Web page recommender system based on folksonomy mining for ITNG'06 submissions. *Third International Conference on Information Technology: New Generations (ITNG'06)*, 388–393.

Nobari, A. D., Neshati, M., & Gharebagh, S. S. (2020). Quality-aware skill translation models for expert finding on StackOverflow. *Information Systems*, *87*, 101413.

Olieman, A., Kamps, J., Satyukov, G., & de Valk, E. (2016). Topical Generalization for Presentation of User Profiles. *arXiv preprint arXiv:1608.07952*.

Omidvar, A., Garakani, M., & Safarpour, H. R. (2014). Context based user ranking in forums for expert finding using WordNet dictionary and social network analysis. *Information Technology and Management*, *15*(1), 51–63.

Paddeu, J. (1999). Gérard Donnadieu, Philippe Denimal, Classification-Qualification : De l'évaluation des emplois à la gestion des compétences, Paris, Éditions Liaisons, 1993. *Formation Emploi*, *67*(1), 127–128.

Paisley, J., Wang, C., Blei, D. M., & Jordan, M. I. (2014). Nested hierarchical Dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(2), 256–270.

Parlier, M. (1994). La compétence au service d'objectifs de gestion. *La compétence. Mythe, construction ou réalité*, 91–108.

Petkova, D., & Croft, W. B. (2008). Hierarchical language models for expert finding in enterprise corpora. *International Journal on Artificial Intelligence Tools*, *17*(01), 5–18.

Piwowarski, B. (2003). *Techniques d'apprentissage pour le traitement d'informations structurées : Application à la recherche d'information*. Doctoral dissertation, University Paris 6, France.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137.

Puglisi, S., Parra-Arnau, J., Forné, J., & Rebollo-Monedero, D. (2015). On content-based recommendation and user privacy in social-tagging systems. *Computer Standards & Interfaces*, *41*, 17–27.

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. *Proceedings of the 2009 conference on empirical methods in natural language processing*, 248–256.

Rampisela, T. V., & Yulianti, E. (2020). Semantic-Based Query Expansion for Academic Expert Finding. *2020 International Conference on Asian Language Processing (IALP)*, 34–39.

Riahi, F., Zolaktaf, Z., Shafiei, M., & Milios, E. (2012). Finding expert users in community question answering. *Proceedings of the 21st International Conference on World Wide Web*, 791–798.

Ribeiro, I. S., Santos, R. L., Gonçalves, M. A., & Laender, A. H. (2015). On tag recommendation for expertise profiling : A case study in the scientific domain. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 189–198.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *Proceedings of the eighth ACM international conference on Web search and data mining*, 399–408.

Rodriguez, M. A., & Bollen, J. (2008). An algorithm to determine peer-reviewers. *Proceedings of the 17th ACM conference on Information and knowledge management*, 319–328.

Roqueplo, P. (1997). *Entre savoir et décision, l'expertise scientifique. Editions Quae.*

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2012). The author-topic model for authors and documents. *arXiv preprint arXiv:1207.4169*.

Rybak, J., Balog, K., & Nørv\a ag, K. (2014). Temporal expertise profiling. *European Conference on Information Retrieval*, 540–546.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613-620.

Sarahelen Thompson, Mark L Waller, J. E. F. (1988). The Intraday Variability Of Soybean Futures Prices: Information and Trading Effects. At Urbana-Champaign. *Review of Futures Markets*, *7, 110-126*.

Serdyukov, P., Taylor, M., Vinay, V., Richardson, M., & White, R. W. (2011). Automatic people tagging for expertise profiling in the enterprise. *European Conference on Information Retrieval*, 399–410.

Shepitsen, A., Gemmell, J., Mobasher, B., & Burke, R. (2008). Personalized recommendation in social tagging systems using hierarchical clustering. *Proceedings of the 2008 ACM conference on Recommender systems*, 259–266.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*, *427*(7), 424–440.

Swanson, R. (2007). *Analysis for improving performance: tools for diagnosing organizations and documenting workplace expertise. San Francisco: Berrett-Koehler; 1994.*

Syed, S., & Spruit, M. (2017). Full-text or abstract ? Examining topic coherence scores using latent dirichlet allocation. *2017 IEEE International conference on data science and advanced analytics (DSAA)*, 165–174.

Tang, J., Hu, X., & Liu, H. (2013). Social recommendation : A review. *Social Network Analysis and Mining*, *3*(4), 1113–1133.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2005). Sharing clusters among related groups : Hierarchical Dirichlet processes. *Advances in neural information processing systems*, 1385–1392.

Titov, I., & McDonald, R. (2008). Modeling online reviews with multi-grain topic models. *Proceedings of the 17th international conference on World Wide Web*, 111–120.

Toutanova, K., & Manning, C. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. *Proceedings of the 2000 Joint SIGDAT Conference EMNLP/VLC, 63-71, 2000*.

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582.

Wang, G. A., Jiao, J., Abrahams, A. S., Fan, W., & Zhang, Z. (2013). ExpertRank : A topic-aware expert finding algorithm for online knowledge communities. *Decision Support Systems*, *54*(3), 1442–1451.

Xie, X., Li, Y., Zhang, Z., Pan, H., & Han, S. (2016). A Topic-Specific Contextual Expert Finding Method in Social Network. *Asia-Pacific Web Conference*, 292–303.

Yang, B., & Manandhar, S. (2014). Tag-based expert recommendation in community question answering. *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, 960–963.

Yimam, D., & Kobsa, A. (2000). Demoir : A hybrid architecture for expertise modeling and recommender systems. *Proceedings IEEE 9th International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WET ICE 2000)*, 67–74.

Yuan, S., Zhang, Y., Tang, J., Hall, W., & Cabotà, J. B. (2020). Expert finding in community question answering : A review. *Artificial Intelligence Review*, *53*(2), 843–874.

Zeng, J., Cheung, W. K., Li, C., & Liu, J. (2010). Coauthor network topic models with application to expert finding. *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, *1*, 366–373.

Zhang, Jing, Tang, J., Liu, L., & Li, J. (2008). A mixture model for expert finding. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 466–478.

Zhang, Jun, Ackerman, M. S., & Adamic, L. (2007). Expertise networks in online communities : Structure and algorithms. *Proceedings of the 16th international conference on World Wide Web*, 221–230.

Zhao, F., Zhu, Y., Jin, H., & Yang, L. T. (2016). A personalized hashtag recommendation approach using LDA-based topic model in microblog environment. *Future Generation Computer Systems*, *65*, 196–206.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., & Li, X. (2011). Comparing twitter and traditional media using topic models. *European conference on information retrieval*, 338–349.

Zhou, D., Lawless, S., Wu, X., Zhao, W., & Liu, J. (2016). Enhanced personalized search using social data. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 700–710.

Zhu, H., Chen, E., & Cao, H. (2011). Finding experts in tag based knowledge sharing communities. *International Conference on Knowledge Science, Engineering and Management*, 183–195.

Résumé

Les organisations s'interrogent de plus en plus sur les avantages que peuvent leur procurer les modèles du Web collaboratif grand public (wiki, réseaux sociaux, crowdsourcing, etc.). Les réseaux sociaux sont un des moyens pour diffuser la connaissance et pour innover grâce à l'utilisation des informations qu'ils génèrent.
La problématique générale de notre travail s'inscrit dans le cadre de la recherche d'expertise, un domaine issu de la recherche d'information, qui se focalise sur l'estimation, la découverte, repérage et classement des expertises des utilisateurs. Les indicateurs de l'expertise, appelés souvent sources d'évidence, sont dans la majorité du temps liés aux documents rédigés par ces utilisateurs, leurs activités électroniques liées aux emails, propriétés de pages ou de sites web, etc. Vient s'ajouter ces dernières années leurs activités sociales : commentaires, postes, réponses aux questions, etc.
Dans notre contexte, nous nous intéressons plus particulièrement aux réseaux communautaires formés de professionnels amenés à partager de l'information et de la connaissance (sous diverses formes) autour de projets et/ou dans le cadre d'activités métiers. Ces professionnels sont connus par leurs profils et préférences mais aussi par leurs expertises et compétences, des notions complexes mais nécessaires à capitaliser. Ces notions difficiles à quantifier vu leur dynamicité, complexité et diversité des éléments pouvant apporter un plus à leurs estimations.
Dans ce travail, et pour répondre à notre problématique, nous avons d'abord mené une étude sur le domaine de la recherche d'expertise, nous avons mis en exergue les principaux travaux ayant apporté des plus dans ce domaine. Puis nous avons proposé un ensemble de contributions à l'estimation de l'expertise en se basant sur l'une des activités sociales largement utilisées par les utilisateurs, mais qui n'a pas été profondément exploitée à savoir l'activité du tagging social ainsi que les tweets.

Mots-clés : Compétence, expertise, recherche d'experts, profilage d'expert, topic, topic modeling, tagging social.

Abstract

Organizations are increasingly wondering about the advantages that collaborative Web models (wiki, social networks, crowdsourcing, etc.) can provide. Social networks are one of the means to disseminate knowledge and to innovate through the use of the information they generate.
The general problematic of our work falls within the expertise search, a field resulting from the information retrieval field, which focuses on the estimation, discovery, identification and classification of users' expertise. Expertise indicators, often called sources of evidence, are usually documents written by these users, their electronic activities like emails, page or website properties, etc. In recent years, their social activities have been added: comments, posts, answers to questions, etc.
In our context, we are particularly interested in community networks made up of professionals who share information and knowledge (in various forms) around projects and / or in the context of business activities. These professionals are known by their profiles and preferences but also by their expertise and skills, concepts that are complex but necessary to capitalize. These notions are difficult to quantify given their dynamicity, complexity and diversity of elements that can enhance their estimations.
In this work, and to resolve our problem, we first conducted a study on the field of expertise research; we highlighted the main work that has contributed more in this area. Then we proposed a set of contributions to the estimation of expertise based on one of the social activities widely used by users, but which has not been deeply exploited, namely the activity of social tagging and also tweets.

Keywords: Competency, expertise, expert finding, expert profiling, topic, topic modeling, social tagging.

ملخص

تتساءل المنظمات بشكل متزايد عن الفوائد التي يمكن أن تحصل عليها من نماذج الويب التعاونية (ويكي، الشبكات الاجتماعية ، التعهيد الجماعي ، وما إلى ذلك).الشبكات الاجتماعية هي إحدى وسائل نشر المعرفة والابتكار من خلال استخدام المعلومات التي تنتجها.
تندرج الإشكالية العامة لعملنا في إطار البحث عن الكفاءة، وهو مجال ناتج عن البحث عن المعلومات، والذي يركز على تقدير، اكتشاف، تحديد وتصنيف خبرة المستخدمين.
ترتبط مؤشرات الخبرة، التي غالبًا ما تسمى بمصادر الأدلة، في الغالب بالوثائق التي كتبها هؤلاء المستخدمون، وأنشطتهم الإلكترونية المرتبطة برسائل البريد الإلكتروني أو خصائص الصفحة أو موقع الويب ، إلخ .في السنوات الأخيرة، تمت إضافة أنشطتهم الاجتماعية :تعليقات، منشورات، إجابات على الأسئلة، إلخ.
في سياقنا، نحن مهتمون بشكل خاص بالشبكات الاجتماعية المكونة من محترفين يشاركون المعلومات والمعرفة (بأشكال مختلفة) حول المشاريع و / أو في سياق الأنشطة التجارية .يُعرف هؤلاء المحترفون بملفاتهم الشخصية وتفضيلاتهم ولكن أيضًا بخبراتهم ومهاراتهم، فهي مفاهيم معقدة ولكنها ضرورية للاستفادة منها . يصعب تحديد هذه المفاهيم كمياً بالنظر إلى ديناميكيتها وتعقيدها وتنوع العناصر التي يمكن أن تزيد من تقديراتها.
في هذا العمل، وسعيا منا للمساهمة في حل الإشكالية، أجرينا أولاً دراسة في مجال أبحاث الكفاءة، سلطنا الضوء على الاعمال الرئيسية التي ساهمت بشكل أكبر في هذا المجال .ثم اقترحنا مجموعة من المساهمات لتقدير الخبرة بناءً على أحد الأنشطة الاجتماعية التي يستخدمها المستخدمون على نطاق واسع، ولكن لم يتم استغلالها بعمق، ألا وهو نشاط العلامات الاجتماعية.

الكلمات الرئيسية :الكفاءة، الخبرة، اكتشاف الخبراء، تحديد سمات الخبراء، الموضوع، نمذجة الموضوع، وضع العلامات الاجتماعية.