# Recherche sociale et personnalisée d'Information

Nawal Ould Amer

# THÈSE

Pour obtenir le grade de

**DOCTEUR DE L'UNIVERSITE GRENOBLE ALPES**

Spécialité : **Informatique**

Arrêté ministériel : 25 mai 2016

Présentée par

**Nawal OULD AMER**

Thèse dirigée par **Philippe MULHEM,** CR au CNRS, Université Grenoble Alpes, codirigée par **Mathias GÉRY,** MCF, Université Jean Monnet de Saint-Etienne.

Préparée au sein du **Laboratoire Informatique de Grenoble,** dans **l'École Doctorale Mathématiques, Sciences et technologies de l'information, Informatique**

# Recherche Sociale et Personnalisée d'Information

# Social and Personalized Information Retrieval

Thèse soutenue publiquement le **16 Décembre 2020**, devant le jury composé de :

**Monsieur Philippe MULHEM**, IMAG, Université Grenoble Alpes, Directeur de thèse

**Madame Sylvie CALABRETTO**
Professeur**,** LIRIS-INSA Lyon, Rapportrice

**Monsieur Eric SANJUAN**
Maître de conférences, Université d'Avignon, Rapporteur

**Madame Sihem AMER YAHIA**
Directeur de recherche, CNRS, Univ. Grenoble Alpes, Président

**Monsieur Mohand BOUGHANEM**
Professeur, Université Paul Sabatier Toulouse, Examinateur

*Recherche Sociale et Personnalisée d'Information*

# Résumé

Une large gamme de services et de plateformes rendent l'utilisateur de plus en plus interactif avec le web. De nombreuses informations qui concernent à la fois les utilisateurs et les ressources (documents, images, vidéos, commentaires, tweets, tags, etc.) sont constamment générées. Ces informations peuvent être très utiles dans les tâches de recherche d'information, pour la modélisation des utilisateurs et des ressources. Cependant, les modèles classiques de recherche d'information n'intègrent pas le contexte social de l'utilisateur et des ressources. Par conséquent, de nombreuses recherches se sont intéressées à combiner ces deux domaines qui sont la recherche d'information et les réseaux sociaux, ce qui a donné lieu à des modèles de recherche d'information sociale.

L'extraction, l'analyse et la représentation d'information sur les activités sociales des utilisateurs jouent un rôle important pour les systèmes de recherche d'information personnalisée. Il est important de créer des modèles d'utilisateurs précis et inférer leurs centres d'intérêts à partir de toutes les informations.

Dans cette thèse, on s'intéresse à la problématique de modélisation des profils des utilisateurs dans les folksonomies. Nous étudions la problématique de pondération des termes du profil de l'utilisateur. Plus précisément, comment estimer parmi toutes les informations des utilisateurs, les données qui peuvent représenter ses centres d'intérêts.

Dans la première partie de la thèse, nous présentons un état de l'art des travaux de la recherche d'information et la recherche d'information sociale personnalisée.

Ensuite, nous décrivons les deux principales contributions. La première contribution de cette thèse réside dans la définition d'un modèle utilisateur représenté par les tags, tel que ces tags couvrent les sujets des documents auxquels ils ont été attribués. Notre approche se distingue par l'intégration du document dans l'estimation des poids des tags de l'utilisateur.

La seconde contribution de cette thèse concerne la définition d'une nouvelle approche de modélisation de l'utilisateur basée sur les documents.La particularité de ce modèle est de faire dépendre les termes du document non seulement du contenu textuel du document mais également des tags attribués par l'utilisateur à ce document. Le but est de déterminer les termes importants du document qui reflètent les centres d'intérêts de l'utilisateur.

La dernière partie de la thèse est consacré à l'évaluation de nos propositions. Les résultats obtenus sont très encouragement et améliorent les performances des systèmes de recherche d'information.

*Social and Personalized Information Retrieval*

# Abstract

A wide range of services and platforms make the user more and more interactive with the web. A lot of information that concerns both users and resources (documents, images, videos, comments, tweets, tags, etc.) is constantly generated. This information can be very useful in information retrieval tasks, for user modeling. However, classical information retrieval models do not integrate the social context of the user.

Therefore, a lot of research has been interested in combining these two areas of information retrieval and social networks, which has given rise to models of social information retrieval and personalised social information retrieval.

The extraction, analysis and representation of information about the social activities of users play an important role in the personalized information retrieval systems. Hence, it is crucial to create accurate user models and infer their interests from all this information.

In this thesis, we investigate how to create a user profile using folksonomies. We study the problem of terms weighting. Specifically, how to estimate among all the user data, the useful information that can be used to represent his interests.

In the first part of this thesis, we present a review of state-of-the-art research on information retrieval and personalized social information retrieval work.

In the second part, we describe our two main contributions. The first contribution of this thesis lies in the definition of a user tag-based model, where these tags cover the topics of the documents to which they are associated. Our approach is distinguished by the integration of the document content into the estimation of user tag weights.

The second contribution of this thesis concerns the definition of a new approach of user modeling based on documents. The particularity of this model is to use user tags to estimate the relevant document terms. The goal is to select only the terms that describe the document topics, which interest the user.

The last part of the thesis is dedicated to the evaluation of our proposals. The results obtained are very encouraging and our approaches improve the performance of the IR systems.

# Remerciements

D'abord et avant tout, je voudrais exprimer ma sincère gratitude à mes directeurs de thèse Philippe Mulhem et Mathias Gèry. Sans vous ce travail n'aurait jamais pu être fini !

Je tiens à vous remercier d'avoir accepté de diriger mes travaux de thèse, de m'avoir conseillé tout au long de ces années et de m'avoir fait confiance.

J'ai passé de très belles années à travailler avec vous dans une bonne ambiance et une très bonne entente. Vous avez toujours été là pour moi, vous m'avez vraiment soutenu et vous avez été vraiment très patient.

Pour être honnête, je n'aurais jamais pu terminer cette thèse sans vous et je ne vous remercierai jamais assez.

Je remercie les rapporteurs de cette thèse : les Professeurs Sylvie Calibretto et Eric San-Juan d'avoir consacré du temps à lire et à rapporter mon travail de thèse. J'aimerais également profondément remercier les examinateurs : les Professeurs Sihem Amer-Yahia et Mohand Boughanem pour avoir accepté d'examiner cette thèse et de participer à ma soutenance.

Je remercie grandement tous les membres de l'équipe MRIM, Lorraine Goeriot, Marie-Christine Fauvet, Georges Quénot, Catherine Berrut, Jean-Pierre Chevallet, Nathalie Denos, pour tous les moments qu'on a passé ensemble, votre écoute, et vos encouragements.

Les doctorants de l'équipe MRIM, Mohannad Almasri, Karam Abdulahhad, Bahjat Safadi, Maxime Portaz, Seydou Doumbia, Anuvabh Dutt, et Jibril Frej, ont été présent tout au long de cette thèse, pour des discussion, des échanges, et partages de super moments. Je n'oublierai jamais !

Enfin, je remercie Isma et Sabira et toute ma famille pour leur amour et soutien.

*A la mémoire de ceux qui me sont chers . . .*

# Table des matières

# Table des figures

# Liste des tableaux

# Première partie

# Introduction

# Chapitre 1

# Introduction

## 1.1 Contexte

Actuellement, le web compte plus de 60 milliards de documents [1] et plus de 4.5 milliards de personnes sur le internet et spécifiquement sur les réseaux sociaux [2] comme présenté dans la figure 1.1 [3].

Les plateformes sociales tels comme Twitter [4], Facebook [5], YouTube [6], Digg [7] , ou Delicious [8], permettent aux gens de publier, de partager des publications, des commentaires, des opinions, des vidéos, et des images.

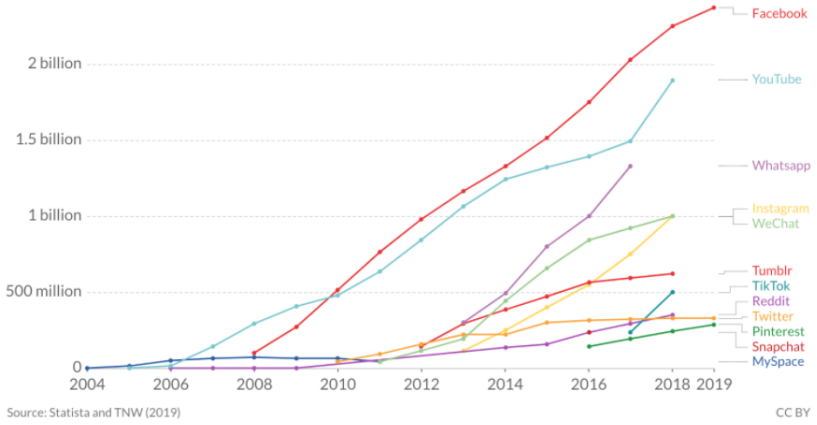



FIGURE 1.1 – Nombre d'utilisateurs sur les réseaux sociaux

La personnalisation est devenue clairement un élément essentiel pour les médias sociaux et les moteurs de recherche. Les exemples de recommandation de produits sur Amazon [9] ou la publicité ciblée de Google [10], en est la preuve.

Un défi important pour les fournisseurs d'applications sociales est d'offrir une personnalisation pertinente et robuste sans avoir à demander des informations explicites aux utilisateurs.

1. https ://www.worldwidewebsize.com/
2. https ://ourworldindata.org/internet
3. https ://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/
4. https ://twitter.com/?lang=en
5. https ://facebook.com/
6. https ://YouTube.com/
7. https ://Digg.com/
8. http ://del.icio.us//
9. https ://www.amazon.com/
10. https ://www.google.com/

L'extraction, l'analyse et la représentation d'information sur les activités sociales des utilisateurs sur le web jouent un rôle important pour les systèmes de recherche d'information personnalisée et pour les systèmes de recommandation [1]. Ainsi, il est important de créer des modèles d'utilisateurs précis et inférer leurs centres d'intérêts à partir de toutes les informations disponibles sur ces plateformes [2].

Pour être réaliste, il faut être capable de faire la distinction entre les différents types d'activités des utilisateurs, reconnaître les thématiques importantes à partir de données structurées et non structurées générées par l'utilisateur et sélectionner les plus appropriés en fonction de leur usage ultérieurs, comme pour une tâche de recommandation ou de recherche d'information.

Les informations fournies par les utilisateurs sur le web comprennent également des "Méta-données" fournies explicitement par les utilisateurs. L'une des fonctionnalités les plus importantes du web social et de ses applications est le concept d'annotation [3]. L'annotation consiste à attribuer manuellement aux ressources web (documents web, images et vidéos) des méta-données sous forme de mots-clés appelés "tags", créés à la volée par des utilisateurs [4].

Simplement, l'annotation permet aux utilisateurs d'exprimer leurs opinions sur une ressource [2]. En tant que tel, cela peut être considéré comme une forme d'un nouveau type d'interaction, qui implique les trois entités : les utilisateurs, les tags et les ressources. Ce triplet est communément appelé "Folkonomie", qui est une combinaison entre "folk" (utilisateur) et "taxonomy" [5].

Certains auteurs ont mené des études et analyses pour mieux comprendre ces structures [3], et concluent que les folksonomies ou les systèmes de marquage social comme Delicious ou Digg, peuvent fournir des résultats non fournis par des systèmes classiques de recherche d'information. Ainsi, plusieurs travaux se sont intéressés à l'exploitation des folksonomies pour la recherche d'information sociale personnalisée [2, 6, 7], ou pour la recommandation [1, 8].

## 1.2 Problématique de modélisation de l'utilisateur pour la recherche d'information sociale personnalisée

Dans le cadre de la recherche d'information sociale personnalisée (RISP), les folksonomies sont utilisées pour construire les profils des utilisateurs pour les intégrer dans les systèmes. Principalement, les tags et les documents annotés sont utilisés pour inférer les centres d'intérêts des utilisateurs. Nous distinguons deux types de profils utilisateurs. Des profils qui se basent sur les tags et des profils qui se basent sur les contenu des documents annotés.

### 1.2.1 Profil basé sur les tags

Les tags des utilisateurs employés pour annoter les documents sont souvent utilisé pour déterminer leurs centres d'intérêts [2, 6, 7, 9-16]. Plusieurs travaux ont été proposés pour estimer le poids de chaque tag, où ce poids reflète son importance pour l'utilisateur. Certaines approches proposent une représentation vectorielle des tags de l'utilisateur. Le poids de chaque tag peut suivre un schéma de pondération basé sur la fréquence d'utilisation de ce tag par l'utilisateur (TF) [2], ou par un TF-IDF [10], ou une combinaison du modèle TF-IUF et du modèle BM25 [14].

D'autres travaux se sont intéressés à inclure le réseau social de l'utilisateur [6, 12, 17]. Par exemple, une version de TF-IUF [12, 18] en tenant compte du réseau social des utilisateurs est proposée. L'idée globale est que si plusieurs utilisateurs

attribuent un même tag à un même document, alors potentiellement ce tag est un tag non erroné et donc éventuellement, ce tag traite du sujet du document.

Cependant, le problème des folksonomies est que la sémantique des attributions de tags aux documents n'est pas bien définie et les tags représentent une information courte et donc qui peut être ambigu. De plus, les modèles de folksonomie font abstraction du contexte d'utilisation dans lequel les activités d'annotation ont été effectuées.

Plusieurs études ont été menées pour comprendre les structures des systèmes de folksonomies [19]. La conclusion était que les tags des utilisateurs sont corrélés avec leurs motivations d'annotation. Les auteurs ont identifié que les utilisateurs ont différentes motivations dans leur activité d'annotation. Pour tenir compte des variétés des types de tags, des travaux proposent de faire une classification des tags d'un utilisateur en plusieurs catégories : tags libres, tags généraux, tags spécifiques, tags synonymes, tags contextuels, tags subjectifs [20].

Par conséquent, à partir d'une attribution de tag pour une ressource (un document par exemple), il est difficile de déduire l'intention réelle de l'utilisateur, le tag a-t-il été attribué pour faciliter la récupération future ou exprime-t-il plutôt une opinion ?

Le contexte de notre thèse est la recherche d'information sociale personnalisée. Notre objectif est de construire des profils utilisateurs afin de les intégrer dans les systèmes de RISP. Le type d'information que nous souhaitons mettre en avant est de type contenu afin de retourner des documents qui répondent aux requêtes de type contenu, aux utilisateurs. Les profils que nous souhaitons construire sont de types contenu et sont basés sur les tags de l'utilisateur. Alors, les tags que nous devons prendre en compte doivent être de type contenu.

Cependant, les tags de l'utilisateur peuvent être associés à n'importe quelle catégorie : tags libres, généraux, spécifiques, ou tags synonymes, suivant ses motivations d'annotation. Pour résoudre le problème d'identification des tags qui sont de type contenu, nous proposons d'employer le document comme référence, permettant ainsi de contrôler les tags à choisir.

### 1.2.2 Profil basé sur les documents annotés

Les documents annotés par les utilisateurs sont une source riche d'information. Ces documents offrent un contenu plus large que les tags et ce contenu pourrait encore mieux aider à identifier les centres d'intérêts de l'utilisateur. De plus, un utilisateur pourrait sauvegarder un document sans forcement l'avoir annoté ou pourrait utiliser très peu de tags. Ceci indique que potentiellement que le contenu des documents est particulièrement intéressant et peut fournir plus d'information liée aux centres d'intérêts de l'utilisateur.

A notre connaissance, très peu de travaux ont été menés pour cette catégorie de cette construction de profils, où les documents sont exploités pour construire les profils des utilisateurs dans les folksonomies. Deux grandes principales approches ont été proposées.

La première approche [7] propose de prendre en compte tous les documents d'un utilisateur pour le représenter et proposent ainsi un profil basé sur les termes des documents. La pondération des termes du profil est estimée par modèle de langue standard.

D'autres approches proposent de modéliser l'utilisateur par des thèmes latents en utilisant les modèles thématiques comme le modèle LDA [21]. Des approches similaires [22, 23] proposent d'intégrer les tags des utilisateurs aux documents. Ainsi, un apprentissage des thèmes latents est réalisé sur ces documents.

Un utilisateur peut avoir des intérêts différents pour un même document. C'est-à-dire, pour un document couvrant 3 thèmes par exemple, l'utilisateur peut n'être intéressé que par une seule thématique et donc potentiellement n'annoterait le document qu'avec des tags couvrant cette thématique. Donc, les autres thématiques du document n'intéressent pas énormément l'utilisateur. Donc, les termes de ces thématiques ne devraient donc pas être pris en compte ou du moins avoir des poids moins importants par rapport aux autres termes décrivant les thématiques importantes pour l'utilisateur.

Notre intuition est que les utilisateurs peuvent être intéressés par des sujets différents pour un même document, et que leur intérêt pour un document est reflété par les tags qu'ils emploient pour l'annoter. Ainsi, un terme du document est un terme important si ce terme est en relation avec les tags de l'utilisateur attribués au même document.

Cependant, comme nous l'avons mentionné dans le point précédent, les tags des utilisateurs sont corrélés avec leurs motivations d'annotation. Donc, tous les tags ne doivent pas être pris en compte pour estimer les termes importants du document.

Alors, comment peut-on ne prendre en compte que les termes du document qui sont en liens avec les tags de l'utilisateur, tel que ces tags représentent le sujet du document ?

## 1.3 Contributions

Dans cette thèse, nous souhaitons étudier comment exploiter les tags et les documents d'un utilisateur pour créer un profil précis et reflétant les centres d'intérêts de cet utilisateur. Principalement :

1. **Comment estimer les tags pertinents pour modéliser l'utilisateur ?**
2. **Comment estimer les termes pertinents du document pour modéliser l'utilisateur ?**

Tout en nous appuyant sur les systèmes de folksonomies comme une source d'information, notre objectif dans cette thèse est d'exploiter les informations sociales pour fournir des approches de modélisation de l'utilisateur dans le but d'améliorer les performances des systèmes de recherche d'information sociale personnalisée.

Les contributions de cette thèse sont résumées dans les points suivants :

1. **Modèle de l'utilisateur basé sur les tags - Exploitation des documents pour la pondération des tags**
   La première contribution de cette thèse réside dans la définition d'un modèle utilisateur représenté par les tags, tel que ces tags couvrent les sujets des documents auxquels ils ont été attribués.
   Notre approche se distingue par l'intégration du document dans l'estimation des poids des tags de l'utilisateur.
   Nous proposons de donner une définition au lien entre le document et les tags et qui est porté par notre hypothèse principale qui est :

*"**Seuls les tags qui décrivent les sujets des documents doivent être pris en compte**"*.

Partant de cette hypothèse globale, nous proposons trois sous hypothèses qui sont :

(a) **H1** : *Seuls les tags de l'utilisateur qui sont des termes du document sont des tags pertinents et décrivent le contenu du document d'après l'utilisateur.*
Pour répondre à cette hypothèse, nous proposons un Modèle de Tag Standard, où le poids des tags est estimé par un modèle de langue standard.

(b) **H2** : *Seuls les tags de l'utilisateur qui sont dans le même espace latent que les thèmes du document sont des tags pertinents et décrivent le contenu du document d'après l'utilisateur.*
Nous proposons un Modèle de Tag Thématique pour l'estimation des tags qui sont dans le même espace latent que les thèmes du document. Le calcul des poids des tags est estimé par le modèle thématique LDA [24].

(c) **H3** : *Seuls les tags de l'utilisateur qui sont dans le même espace sémantique que les termes du document sont des tags pertinents et décrivent le contenu du document d'après l'utilisateur.*
Pour estimer les tags qui sont dans le même espace sémantique que les termes du document, nous proposons un Modèle de Tag Sémantique, qui emploie les plongements de mots, estimés avec le modèle word2vec [25].

2. **Modèle de l'utilisateur basé sur les documents - Exploitation des tags pour la pondération des termes du document**
La seconde contribution de cette thèse concerne la définition d'une nouvelle approche de modélisation de l'utilisateur basée sur les documents.
La particularité de ce modèle est de faire dépendre les termes du document non seulement du contenu textuel du document mais également des tags attribués par l'utilisateur à ce document. Le but est de déterminer les termes importants du document qui reflètent les centres d'intérêts de l'utilisateur. Notre hypothèse principale est :

*"**Seuls les termes du document qui sont liés aux tags doivent être pris en compte**"*.

De plus, les tags que nous prenons en compte doivent être représentatifs des centres d'intérêts de l'utilisateur. Ainsi, nous intégrons notre contribution précédente dans l'estimation du modèle du document.
Donc, nous présentons 4 modèles utilisateurs. Chacun de ces modèles exploitent un modèle de Tags, et sont :

(a) Un modèle de Contenu Basique qui prend en compte tous les tags d'un utilisateur.

(b) Un modèle de Contenu Standard qui intègre le Modèle de Tags Standard dans l'estimation des termes importants du document.

(c) Un modèle de Contenu Thématique qui intègre le Modèle de Tags Thématique dans l'estimation des termes importants du document.

(d) Un modèle de Contenu Sémantique qui intègre le Modèle de Tags Sémantique dans l'estimation des termes importants du document.

Nous évaluons nos modèles et les comparons aux modèles de l'état de l'art en utilisant la collection de test Delicious que nous avons construit.

## 1.4 Plan du manuscrit

Le travail mené dans cette thèse est décrit par le plan suivant :

1. **Le chapitre 2 : Concepts Généraux** présente les principaux concepts abordés dans cette thèse. Plus en détails, nous présentons les connaissances de base de la recherche d'information, les modèles de la recherche d'information, les modèles de représentation de l'information (les modèles thématiques des documents et les modèles de représentation vectorielle continue des termes).
2. **Le chapitre 3 : État de l'art - Modélisation de l'utilisateur dans la RISP** présente les différents travaux réalisés dans le domaine de la recherche d'information sociale personnalisée. Plus spécifiquement, nous présentons les données exploitées pour modéliser l'utilisateur, les différentes représentations possibles des profils utilisateurs et les approches d'évaluation des systèmes de recherche d'information sociale personnalisée. Enfin, nous présentons notre positionnement par rapport aux travaux de l'état de l'art.
3. **Le chapitre 4 : Modèle de l'utilisateur basé sur les tags - Exploitation des documents pour la pondération des tags** détaille notre notre première contribution de modèle utilisateur dans lequel nous nous intéressons à la représentation des centres d'intérêts de l'utilisateur au travers de leurs tags. Nous proposons une nouvelle approche de construction de profil utilisateur en intégrant les documents dans l'estimation des poids des tags de l'utilisateur.
4. **Le chapitre 5 : Modèle de l'utilisateur basé sur les documents - Exploitation des tags pour la pondération des termes du document** présente notre deuxième contribution de modèle utilisateur, dans lequel nous nous intéressons à la représentation des centres d'intérêts de l'utilisateur extraits des documents qu'il a annoté en exploitant les tags associés à ces documents. Nous définissons un nouveau modèle qui consiste à estimer les termes les plus importants du document annoté par l'utilisateur.
5. **Le chapitre 6 : Cadre Expérimental** présente les modèles de références avec lesquelles nous comparons nos modèles, les contraintes expérimentales, les collections de tests et enfin les paramètres des différents systèmes.
6. **Le chapitre 7 : Évaluation du profil utilisateur basé sur les tags** présente les résultats des expérimentations menées pour évaluer la qualité et la robustesse de notre première contribution présentée dans le chapitre 4.
7. **Le chapitre 8 : Évaluation du profil utilisateur basé sur les documents** présente les résultats des expérimentations menées pour évaluer la qualité et la robustesse de notre seconde contribution présentée dans le chapitre 5.
8. **Le chapitre 9 : Conclusion et Perspectives** présente notre conclusion sur les différentes contributions de cette thèse, ainsi que sur les travaux futurs et complémentaires.

## 1.5 Publications

[16] **Ould-Amer Nawal**, Philippe Mulhem, and Mathias Géry. *Personalized Parsimonious Language Models for User Modeling in Social Bookmarking Systems European Conference on Information Retrieval, ECIR 2017*

[26] **Ould-Amer Nawal**, Philippe Mulhem, and Mathias Géry. *Modèles de Document Parcimonieux basés sur les annotations et les words embeddings - Application à la personnalisation. Conférence en Recherche d'Informations et Applications, CORIA 2017*

[27] **Ould-Amer Nawal**, Philippe Mulhem, and Mathias Géry. *"Toward Word Embeddingfor Personalized. Information Retrieval, Neu-IR SIGIR Workshop on Neural Information Retrieval, SIGIR 2016*

[28] **Ould-Amer Nawal**. *"Enhancing Personalized Document Ranking using Social In-formation". International Conference on User Modeling, Adaptation and Personalization, UMAP 2016*

[29] Philippe Mulhem, **Ould-Amer Nawal**, and Mathias Géry. *Variations axiomatiques pour la recherche d'information personnalisée. Conférence en Recherche d'Informations et Applications, CORIA 2017*

[30] Philippe Mulhem, Lorraine Goeuriot, Nayanika Dogra, and **Nawal Ould Amer** : *TimeLine Illustration Based on Microblogs : When Diversification Meets Metadata Re-ranking, CLEF 2017*

[31] Philippe Mulhem, **Ould-Amer Nawal**, and Mathias Géry. *"Axiomatic term-based Personalized Query Expansion using Bookmarking System. International Conference on Database and Expert Systems Applications, DEXA 2016*

[32] Seyyed Hadi Hashemi, **Ould-Amer Nawal** and Jaap Kamps. *"University of Amsterdam at TREC 2016 : Contextual suggestion track. The Twenty-Fifth Text REtrieval Conference Notebook, TREC 2016*

[33] **Ould-Amer Nawal**, Philippe Mulhem, Mathias Géry, and Karam Abdulahhad. *"Word Embedding for Social Book Suggestion". Conference and Labs of the Evaluation forum, CLEF 2016*

[34] Dogra, Nayanika, Philippe Mulhem, **Ould-Amer Nawal** , and Lorraine Goeuriot. *"LIG at CLEF 2016 Cultural Microblog Contextualization : TimeLine Illustration based on Microblogs". Conference and Labs of the Evaluation forum, CLEF 2016*

[35] **Ould-Amer Nawal**, Philippe Mulhem, and Mathias Géry. *"LIG at CLEF 2015 SBSLab". Conference and Labs of the Evaluation forum, Toulouse, France, CLEF 2015*

[36] **Ould-Amer Nawal**, Philippe Mulhem, and Mathias Géry. *"LIG at CLEF 2015 SBSLab". Conference and Labs of the Evaluation forum, Toulouse, France, CLEF 2015*

[37] **Ould-Amer Nawal**, Philippe Mulhem, and Mathias Géry. *Recherche de conversations dans les réseaux sociaux : modélisation et expérimentations sur Twitter. Conférence en Recherche d'Informations et Applications, CORIA 2015*

# Deuxième partie

# État de l’art

# Chapitre 2

# Concepts Généraux

## 2.1 Introduction

Dans ce chapitre, nous présentons et analysons les principaux concepts abordés dans cette thèse. Dans la section 2.2, nous présentons les connaissances de base de la recherche d'information ainsi que les modèles de recherche d'information que nous avons employés dans cette thèse. En section 2.3, nous présentons les modèles de représentation de l'information, à savoir les modèles thématiques des documents (Topic Models) et les modèles de représentation vectorielle continue des termes (Word Embedding).

## 2.2 Recherche d'information

La recherche d'information (RI) est un domaine qui a pour but de faciliter l'accès à l'information pour un utilisateur en passant par plusieurs processus comme définit par [38] :

*"An information retrieval system is an information system that is used to store items of information that need to be processed, searched, retrieved, and disseminated to various user populations"*

Les utilisateurs précisent généralement leur besoin en information sous la forme de mots-clefs (requête) que le système de RI va traiter pour déterminer et retourner les documents correspondant à leurs besoins. Étant donné une requête, le système de RI, après un ensemble de processus, tente de récupérer des documents qui doivent être pertinents pour la requête de l'utilisateur.

Dans la suite de la section, nous allons présenter les concepts de base d'un système de RI et les modèles de RI.

### 2.2.1 Concepts de base de la recherche d'information

Le système de RI réalise certaines actions pour répondre de façon pertinente et pour satisfaire le besoin en information de l'utilisateur. Ces actions sont principalement décrites dans un processus appelé "Processus en U". Les étapes principales de ce processus sont : *l'indexation*, *la récupération* et *l'ordonnancement des documents*, et sont présentées dans la Figure 2.1.

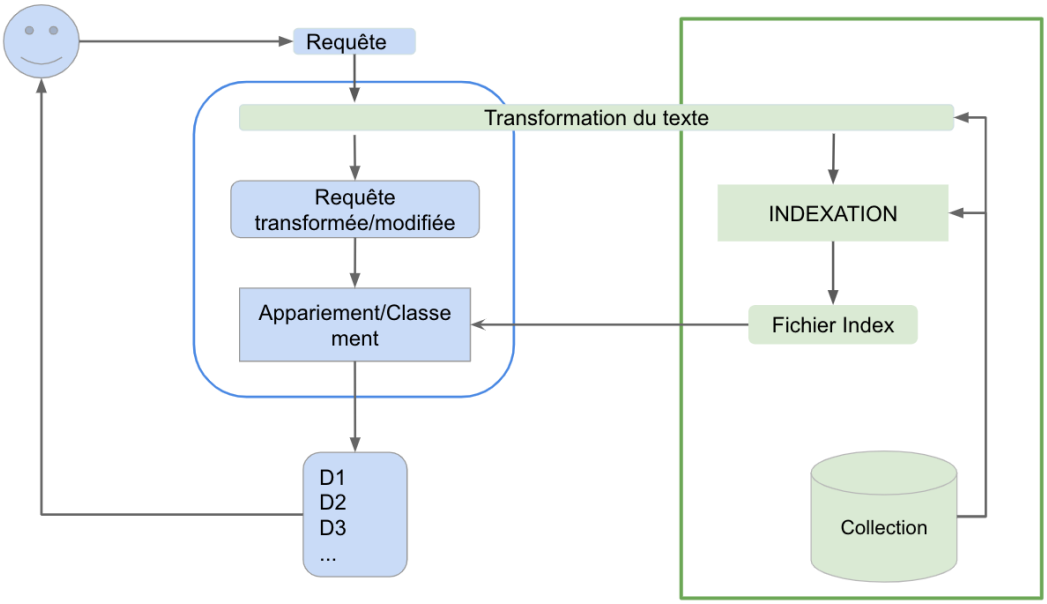



FIGURE 2.1 – Architecture d'un système de Recherche d'information.

1. Un premier processus représentant **l'indexation** (illustré à droite de la figure 2.1) consiste à réduire le texte d'un document en un ensemble de mots-clés, puis les sauvegarder sous forme d'un index avec une architecture bien spécifique dans le but de faciliter la recherche. Cette transformation suit une série de traitement qui sont :
   — **La segmentation** consiste typiquement à éliminer les espaces blancs, la ponctuation, les liaisons, etc, dans un texte donné, ainsi, produire des segments (tokens).
   — **La lemmatisation** permet de réduire le mot en sa racine et donc confondre toutes les formes d'un même mot. Alors, tous les mots de la même famille seront tous représentés par un même mot. Par exemple, les mots : scientifiquement et scientifiques seront représentés par le mot "scientifiqu".
   — **L'élimination de mots vides** a pour but d'éliminer les mots très courants qui ne contribuent pas ou seulement de manière insignifiante au contenu du document, tels que : Le, La, Les, Donc, etc. De plus, cela permet de réduire la taille de l'index sans que cela n'affecte les performances du système de RI.

   Une fois que toutes les étapes ci-dessus sont réalisées, les documents sont représentés dans des fichiers index qui stockent la cartographie des couples "terme-document" en y associant un poids. Ce poids peut par exemple être estimé par une pondération "TF-IDF" [39]. Cette formule favorise les termes qui sont à la fois fréquents dans le document et peu fréquents dans la collection.
2. Un second processus illustré à gauche de la figure 2.1 commence par appliquer à la requête de l'utilisateur les mêmes processus décrits dans l'étape précédente. Ensuite, le processus ***d'appariement*** appliquant un modèle de correspondance entre les mots de la requête et les documents, permet de retourner un ensemble de documents qui seront ensuite ordonnancés par leur

score de correspondance. Cette partie est la plus critique car l'ordre des documents dépend du modèle de correspondance. Les modèles utilisés dans cette étape sont présentés dans la section suivante 2.2.2.

### 2.2.2 Les modèles de recherche d'information

L'étape de correspondance des documents et la requête de l'utilisateur repose sur des modèles de RI qui ont pour objectif l'identification et l'ordonnancement des documents pertinents. Il existe plusieurs modèles de RI, tels que le modèle booléen [40], le modèle vectoriel [41], et le modèle probabiliste [42].

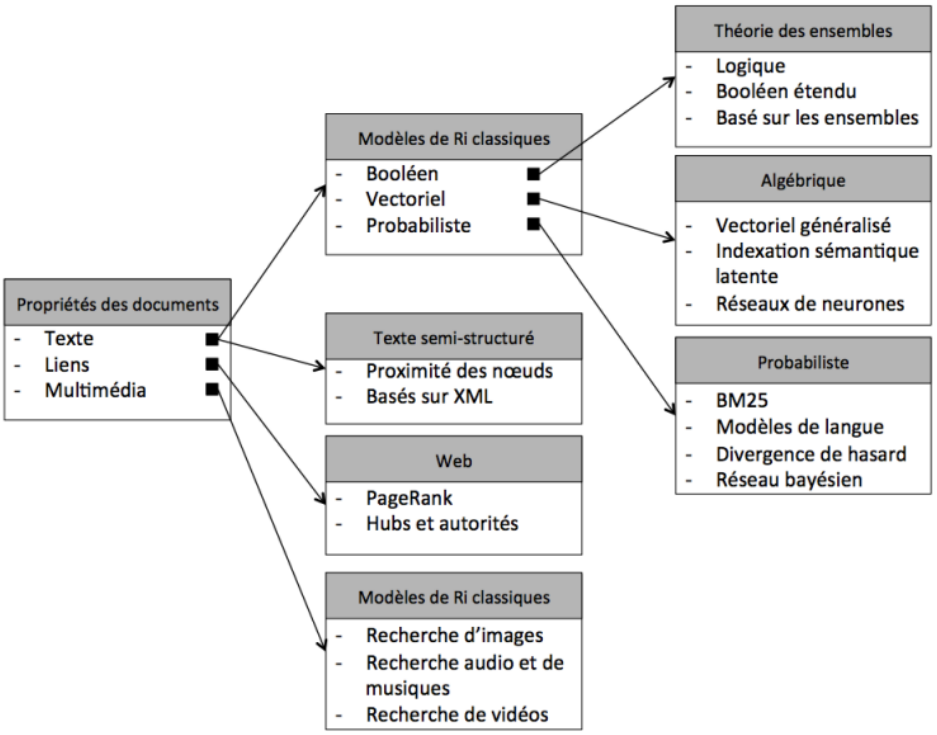



FIGURE 2.2 – Taxonomie des modèles de recherche d'information. [43]

Une présentation détaillée des modèles est proposée [44] avec une taxonomies des modèles de RI comme illustrée dans la figure 2.2.

Nous présentons seulement les modèles probabilistes et en particulier les modèles de langue avec le lissage de Dirichlet et de Jelinek-Mercer.

#### Modèles Probabilistes

Il existe principalement deux familles de modèles probabilistes : les modèles probabilistes classiques [42] et les modèles de langue [45].

#### Les modèles probabilistes classiques

Les modèles probabilistes classiques [42] s'appuient sur la distribution de probabilités de termes pour estimer la similarité entre une requête et un document. Ce

modèle permet de favoriser les documents qui ont une forte probabilité d'être pertinents et une faible probabilité d'être non pertinents. La pertinence du document $d$ par rapport à la requête utilisateur $q$ est décrite comme suit :

$$RSV(d,q) = \frac{p(p|d)}{p(\overline{p}|d} \tag{2.1}$$

avec $p(p|d)$ et $p(\overline{p}|d)$, respectivement représente la probabilité de pertinence et de non pertinence, par rapport à la requête $q$.

Plusieurs méthodes ont été proposées pour estimer cette probabilité, telles que le modèle binaire BIR [46] ou le modèle BM25 [42]. Ce dernier est le plus utilisé et ses atouts majeurs consistent en la considération de la longueur des documents dans le calcul du score de pertinence.

Le score RSV en employant le modèle $BM_{25}$ est calculé comme suit :

$$RSV(d,q) = \sum_{t_i \in q} idf_{t_i} \times \frac{tf(t_i,d).(k1+1)}{tf(t_i,d) + k_1(1-b+b\frac{|d|}{avg_{dl}})} \tag{2.2}$$

$$\tag{2.3}$$

avec $idf_{t_i}$ est la fréquence inverse de document pondérant le terme $t_i$ de la requête, $tf(t_i,d)$ représente la fréquence d'apparition du terme $t_i$ de la requête dans le document $d$, $|d|$ représente la longueur du document et $avg_{dl}$ représente la longueur moyenne des documents. Le paramètres $b$ permet de contrôler la normalisation par la longueur des documents et le paramètre $k_1$ contrôle l'effet de la saturation au niveau des occurrences des termes du document. Les valeurs par défaut des deux paramètres sont $k_1 \in [1.2; 2.0]$ et $b = 0.75$.

#### Les modèles de langue

Le principe de base des modèles de langue [45] en RI est d'ordonner chaque document $d$ de la collection $C$ suivant leur capacité à générer la requête $q$. Ainsi, il s'agit d'estimer la probabilité de génération $p(q|d)$. Pour simplifier, on suppose en que les mots qui apparaissent dans la requête sont indépendants. Ainsi, pour une requête $q = \{t_1, t_2, ..., t_n\}$, cette probabilité de génération est estimée comme suit :

$$p(q|\theta_d) = \prod_{t_i \in q} p(t_i|\theta_d)^{c(t_i,q)} \tag{2.4}$$

$$= \prod_{t_i \in q} \frac{tf(t_i,d)}{|d|}^{c(t_i,q)} \tag{2.5}$$

où $c(t_i,q)$ est la fréquence du terme $t$ dans la requête $q$, et $\theta_d$ est le modèle du document, qui reflète la distribution de termes dans $d$. La probabilité $p(t_i|\theta_d)$ représente la probabilité du terme $t$ dans le modèle du document $\theta_d$.

La probabilité $p(t_i|\theta_d)$ est estimée par la fréquence des termes de la requête $q$ dans le document $d$. Cette probabilité peut être nulle pour les documents ne contenant pas tous les termes de la requête. Dans ce cas, la probabilité $p(q|\theta_d)$ est nulle alors que le document pourrait partiellement répondre au besoin en information de l'utilisateur formulé par la requête $q$.

Afin d'éviter les probabilités nulles, il existe plusieurs méthodes de lissage, par exemple : la méthode de lissage *Jelinek-Mercer* et la méthode de lissage *Dirichlet*.

— ***Lissage de Jelinek-Mercer***
Le score de pertinence d'un document pour une requête définit par la probabilité $p(d|q)$ est estimée comme suit :

$$p(d|q) = \frac{p(q|d)p(d)}{p(q)} \tag{2.6}$$
$$\propto p(d) \prod_{t_i \in q} p(t_i|d)^{c(t_i,q)} \tag{2.7}$$
$$\propto \prod_{t_i \in q} [\lambda p(t_i|d) + (1-\lambda) p(t_i|C)] \tag{2.8}$$
$$\propto \prod_{t_i \in q} [\lambda \frac{tf(t_i,d)}{|d|} + (1-\lambda) \frac{tf(t_i,C)}{|C|}] \tag{2.9}$$

où , $tf(t_i,d)$ et $tf(t_i,C)$ représentent la fréquence d'apparition du terme $t_i$ de la requête dans le document $d$ et dans la collection $C$ respectivement, $|d|$ et $|C|$ représentent la taille du document $d$ et la collection $C$ respectivement, et $\lambda$ est le paramètre de lissage de Jelinek-Mercer. La valeur classique de $\lambda$ est égale à 0.15.

— ***Lissage de Dirichlet***
Le score de pertinence d'un document pour une requête définit par la probabilité $p(d|q)$ est estimée comme suit :

$$p(d|q) = \frac{p(q|d)p(d)}{p(q)} \tag{2.10}$$
$$\propto p(d) \prod_{t_i \in q} p(t_i|d)^{c(t_i,q)} \tag{2.11}$$
$$\propto \prod_{t_i \in q} [\frac{tf(t_i,d) + \mu \frac{tf(t_i,C)}{|C|}}{|d| + \mu}] \tag{2.12}$$

où $tf(t_i,d)$ et $tf(t_i,C)$ représentent la fréquence d'apparition du terme $t_i$ de la requête dans le document $d$ et dans la collection $C$ respectivement, $|d|$ et $|C|$ représentent la taille du document $d$ et la collection $C$ respectivement, et $\mu$ est le paramètre de lissage de Dirichlet. La valeur classique de $\mu$ est égale à 2500

## 2.3 Représentation de l'information

Dans cette section, nous présentons les modèles de représentation de l'information : les modèles thématiques des documents (Topic Models) et les modèles de représentation vectorielle continue des termes (Word Embedding).

### 2.3.1 Représentation thématique

Plusieurs modèles probabilistes thématiques latents ont été proposés afin d'analyser les contenus des documents [24, 47-50]. Par exemple, l'analyse sémantique latente (LSA) [50], qui a pour objectif de trouver des mots sémantiquement liés. Cette

méthode est basée sur l'hypothèse que les mots qui apparaissent dans des morceaux de texte similaires ont une signification similaire.

Pour pallier certaines limites du modèle LSA, certains travaux [47] proposent le modèle probabiliste pLSA. Ce modèle est basé sur un mélange de décomposition dérivé d'un modèle à classe latente, tandis que le LSA est basé sur le SVD. Le modèle pLSA est un modèle génératif des documents de l'ensemble où il est estimé. Cependant, il n'est pas un modèle génératif de nouveaux documents.

Pour résoudre ce problème, l'allocation de Dirichlet latente (LDA) a été proposée [24]. La figure 2.3 présente quelques exemples de thèmes découverts par le modèle LDA.

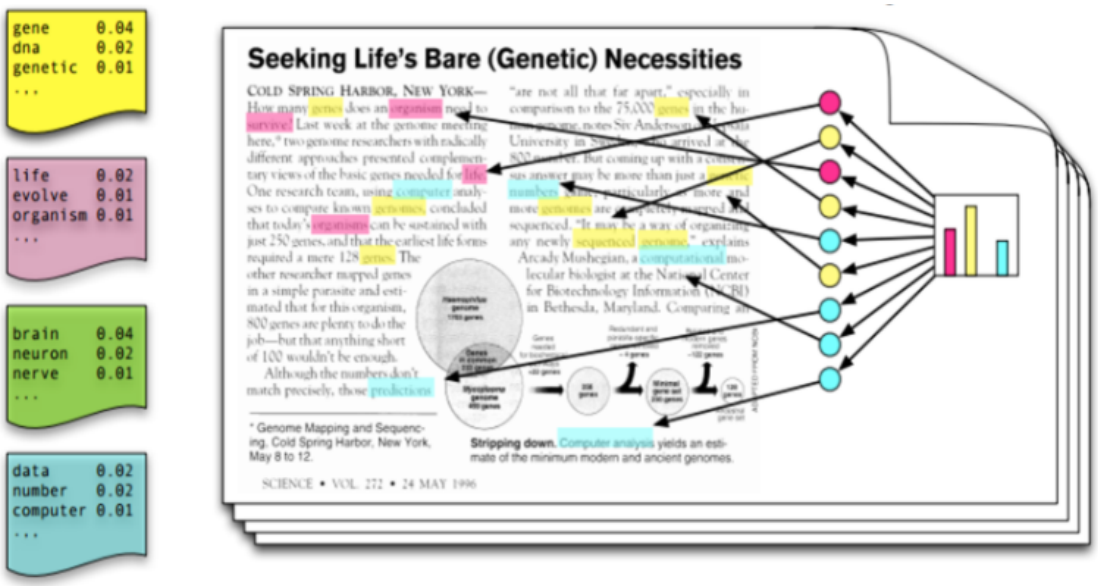



FIGURE 2.3 – Exemple de distribution de thèmes LDA [24]

Comme l'illustre la figure 2.3, les thèmes de la collection des documents sont représentés à gauche et la distribution des thèmes dans le document est présentée à droite sous forme d'histogramme. la distribution des thèmes montre clairement que le document se concentre sur le thème "les aspects génétiques" représenté en couleur jaune, comme représenté dans l'histogramme, car la plupart des mots du document sont attribués au sujet de la générique. Tandis que la distribution de mots pour le sujet "informatique" représenté en vert est faible.

La modélisation par thème, en particulier le LDA, est l'une des approches les plus populaires pour déduire les thèmes d'un corpus (collection de documents) [51, 52]. Le modèle LDA est employé dans d'innombrables domaines. Par exemple, la recherche d'information [53], la désambiguisation des mots [54], l'analyse des sentiments [55] et la synthèse multi-documents [56].

### Le modèle LDA

LDA est un modèle probabiliste génératif offrant un moyen robuste pour identifier les sujets d'un corpus de document.

L'histoire générative de LDA comprend les étapes suivantes :

— Pour chaque thème $k \in [1, K]$, tirer une distribution de mots $\phi_k \sim Dir(\beta)$
— Pour chaque document $d_i, i \in [1, M]$ :
  — Tirer une distribution de thèmes $\theta_i \sim Dir(\alpha)$

— Pour chaque position du mot n $\in d_i$, tel que $n \in [1, N_i]$
  — Tirer un thème $z_{i,n} \sim Multi(1, \theta_{d_i})$
  — Tirer un mot $w_{i,n} \sim Multi(1, \phi_{z_{i,n}})$

L'histoire générative de LDA est un processus qui est appliqué aux termes de corpus $w_i$. Le nombre de thèmes $K$ ainsi que $\alpha, \beta : \alpha \in \mathcal{R}^K et \beta \in \mathcal{R}^V$, sont des paramètres du *prior* de Dirichlet symétrique sur $\theta$ et $\phi$, respectivement.

Premièrement, les distributions des mots $\phi_k$ sont échantillonnées sur l'ensemble du corpus. Pour y parvenir, pour chaque document $d_i$, une proportion $\theta_i$ est échantillonnée sur l'ensemble du corpus. Pour chaque position de mot, un thème $z_{i,n}$ est tiré suivant la loi multinomiale paramétrée par le vecteur $\theta_{d_i}$ et qui désigne le thème qui générera le mot. Enfin, le mot est tiré suivant la loi multinomiale paramétrée par le vecteur $\phi_{z_{i,n}}$.

**L'inférence :** L'histoire générative décrit un processus itératif qui est supposé avoir généré une collection de documents, alors que l'inférence essaie d'obtenir le contraire.

Au lieu de générer le corpus, l'inférence vise à découvrir les paramètres de LDA lors de l'observation des mots d'un corpus. Les paramètres du modèle sont les distributions de thèmes par document $\theta_i$ et les distributions de thèmes par mot $\phi$. Les estimer est équivalent à découvrir les thèmes latents d'une collection. En particulier, étant donné que $\phi$ on peut identifier les mots avec la probabilité la plus élevée pour un thème tandis que $\phi_i$ est une représentation vectorielle de $d_i$ dans l'espace des thèmes. Par conséquent, les documents avec des distributions de thèmes similaires devraient être sémantiquement similaires. Les deux stratégies d'inférence les plus populaires sont l'inférence variationnelle [24] et l'échantillonnage de Gibbs réduit [57] [1].

L'algorithme Gibbs obtient des échantillons postérieurs en balayant chaque bloc de variables et en échantillonnant à partir de leur probabilités conditionnelles, tandis que les blocs restants sont fixes. En pratique, pour LDA l'algorithme initialise aléatoirement les thèmes des mots. Ensuite, pendant les itérations de Gibbs et jusqu'à la convergence, l'algorithme tire des thèmes pour les mots apparaissant dans les documents au fur et à mesure. Les probabilités du tirage multinomiale pour l'échantillonnage du thème d'une position de mot $i$ où le mot $t$ est observé [57, 58], sont données par la formule suivante :

$$p(z_i = k | z_{\neg i}, w) = \frac{\Psi^{(t)}_{k,\neg i} + \beta_t}{\sum_{t=1}^{V} \Psi^{(t)}_{k,\neg i} + \beta_t} (\Omega^{(k)}_{m,\neg i} + \alpha_k) \tag{2.13}$$

ou $\neg_i$ est un indice d'une variable de comptage, $\beta_t$, $\alpha_k$ sont respectivement la t-ième et la k-ième coordonnée de $\beta \in \mathcal{R}^V$ et $\alpha \in \mathcal{R}^K$. $\Psi$ et $\Omega$ sont des variables de comptage.

L'algorithme d'échantillonnage de Gibbs est alors un processus itératif sur les mots d'une collection $C$. L'équation 2.13 est appliquée pour découvrir les thèmes et les distributions des mots dans les thèmes, jusqu'à la convergence de l'algorithme.

**L'inférence sur de nouveaux documents :** Comme LDA est un modèle génératif, il est donc capable d'identifier les distributions de thèmes de nouveaux documents en exécutant le processus d'inférence par échantillonnage de Gibbs sur ces nouveaux

1. Nous présentons ici que l'algorithme d'échantillonnage de Gibbs

documents. En règle générale, très peu d'itérations (<10) sont nécessaires pour déduire les distributions des thèmes de nouveaux documents [58].

### 2.3.2 Représentation vectorielle continue

La représentation vectorielle continue des mots a couramment été utilisée dans le domaine de TAL (traitement automatique de la langue) [59]. L'hypothèse générale derrière ces méthodes est que les mots qui occurrent dans des contextes similaires partagent une relation ou une similarité sémantique .

Les modèles de représentation basés sur la prédiction sont enracinés dans l'idée de la modélisation du langage : prédire la probabilité d'occurrence d'un terme, étant donné l'observation d'un autre terme lorsqu'ils coexistent dans une fenêtre de contexte. Plusieurs études ont été proposées pour l'estimation de ces probabilités en utilisant des techniques de réseaux de neurones [60, 61].

Plus récemment, une approche basée sur un réseau de neurones non profond appelée word2vec est proposée [25]. Le modèle word2vec propose deux modèles de prédiction : le modèle Skip-Gram, qui prédit les termes de contexte d'un terme à partir de l'occurrence du terme, et le modèle CBOW, qui prédit l'occurrence d'un terme, compte tenu de ses termes de contexte. Cette différence est également visible dans l'architecture du modèle CBOW et Skip-Gram, représentée dans la figure 2.4 ci-dessous.

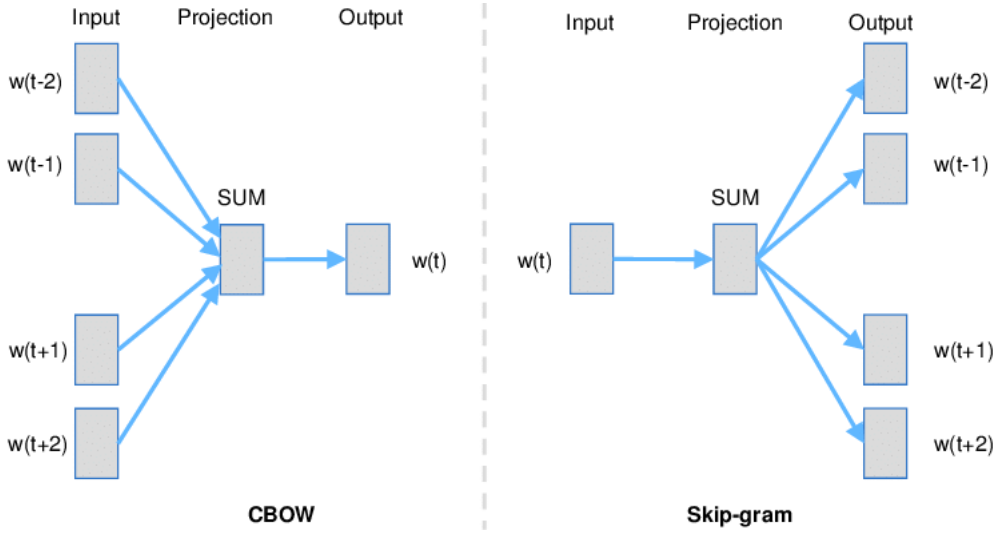



FIGURE 2.4 – Architecture du modèle CBOW et du modèle Skip-Gram

#### 2.3.2.1 Le modèle Skip-Gram

Le modèle Skip-Gram suppose qu'un mot peut être utilisé pour générer les mots qui l'entourent dans une séquence de texte. Par exemple, pour la séquence de texte "Le père aime son fils", le mot «aime» est un mot cible central et la taille de la fenêtre contextuelle est sur 2 mots. Comme présenté dans la figure 2.5, étant donné le mot cible central «aime», le modèle Skip-Gram estime la probabilité conditionnelle de générer les mots du contexte, «le», «père», «son» et «fils», qui sont à une distance d'au plus 2 mots, notée par la probabilité :

$$p(\text{"le"}, \text{"pere"}, \text{"son"}, \text{"fils"} | \text{"aime"}) \tag{2.14}$$

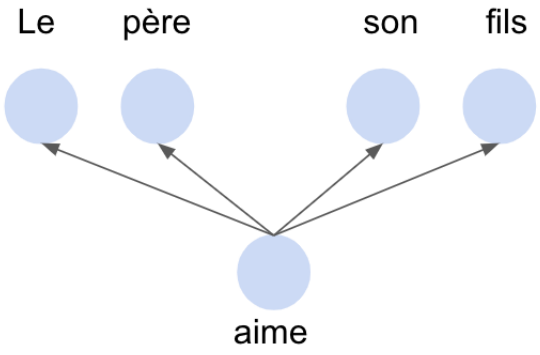



FIGURE 2.5 – Exemple : Skip-Gram

Étant donné le mot cible central, les mots de contexte sont générés indépendamment les uns des autres. Dans ce cas, la formule ci-dessus peut être réécrite comme suit :

$$p(''le''|''aime'')p(''pere''|''aime'')p(''son''|''aime'')p(''fils''|''aime''). \tag{2.15}$$

Dans le modèle Skip-Gram, chaque mot est représenté par deux vecteurs de dimension $d$, qui sont utilisés pour calculer la probabilité conditionnelle. Le mot est indexé à la position $i$ dans le dictionnaire, son vecteur est représenté par $v_i \in \mathcal{R}^d$ lorsqu'il s'agit du mot cible central, et $u_i \in \mathcal{R}^d$ lorsqu'il s'agit d'un mot de contexte.

Soient le mot cible central $w_c$ et le mot de contexte $w_o$. La probabilité conditionnelle de générer le mot de contexte pour le mot cible central donné, peut être obtenue en effectuant une opération softmax sur le produit vectoriel :

$$p(w_o|w_c) = \frac{exp(u_o^T v_c)}{\sum_{i \in V} exp(u_i^T v_c)} \tag{2.16}$$

où le vocabulaire est défini par $V = \{0, 1, \ldots, |V|\}$.

Pour une séquence de texte de longueur $T$, un mot au pas de temps $t$ test noté $w^{(t)}$, les mots de contexte soient générés indépendamment à partir des mots centraux, la taille de la fenêtre de contexte $m$, la fonction de vraisemblance du modèle Skip-Gram est la probabilité conjointe de générer tous les mots de contexte étant donné un mot central, est définit comme suit :

$$\prod_{t=1}^{T} \prod_{\substack{-m \leq j \leq m \\ j \neq 0}} \log p(w^{(t+j)}|w^t) \tag{2.17}$$

Ici, Pour tout pas de temps inférieur à 1 ou supérieur à $T$, ce pas peut être ignoré.

**Apprentissage du modèle Skip-Gram :** Les paramètres du modèle Skip-Gram sont le vecteur de mot cible central et le vecteur de mot de contexte pour chaque mot. Dans le processus d'apprentissage, les paramètres du modèle en maximisant la fonction de vraisemblance sont appris. Cela revient à minimiser la fonction de perte suivante :

$$\mathcal{L} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-c\leq j\leq c\\ j\neq 0}} \log p(w^{(t+j)}|w^{(t)}) \tag{2.18}$$

Pour l'apprentissage, l'algorithme de la descente de gradient est employé. A la fin de l'étape de l'apprentissage, pour tout mot du dictionnaire, les deux vecteurs de mots $v_i$ et $u_i$ sont générés.

#### 2.3.2.2 Le modèle CBOW

Le modèle CBOW est similaire au modèle Skip-Gram. La plus grande différence est que le modèle CBOW suppose que le mot cible central est généré sur la base des mots de contexte avant et après dans la séquence de texte. Pour la séquence de texte «le», «l'homme», «aime», «son» et «son», dans laquelle «aime» est le mot cible central, étant donné une taille de fenêtre contextuelle de 2, le modèle CBOW estime la probabilité conditionnelle de générer le mot cible «aime» sur la base des mots de contexte «le», «l'homme», «son» et «son» (comme illustré par la figure 2.6), comme suit : p("le", père, "son", "fils" | "aime").

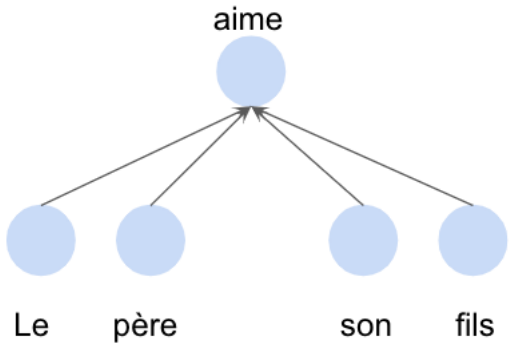



FIGURE 2.6 – Exemple : CBOW

Comme il y a plusieurs mots de contexte dans le modèle CBOW, la moyenne des vecteurs de mots est appliquée, puis la même méthode que le modèle skip-gram est utilisée pour calculer la probabilité conditionnelle.

Soient le mot cible central $w_c$ indexé comme $C$, et les mots contextuels $w_{o_1}, w_{o_2}, ..., w_{o_2m}$ être indexés comme $o_1, o_2, ..., o_2m$ dans le dictionnaire. Nous supposons que $v_i \in \mathcal{R}^d$ et $v_i \in \mathcal{R}^d$ sont le vecteur de mot de contexte et le vecteur de mot cible central du mot d'index $i$ dans le dictionnaire (notez que les symboles sont opposés à ceux du modèle skip-gram). La probabilité conditionnelle de générer un mot cible central à partir du mot de contexte est comme suit :

$$p(w_c|w_{o_1}, w_{o_2}, ..., w_{o_2m}) = \frac{exp(\frac{1}{2m}u_c^T(v_{o_1} + v_{o_2} + ..., v_{o_2m}))}{\sum_{i\in V} exp(\frac{1}{2m}u_i^T(v_{o_1} + v_{o_2} + ..., v_{o_2m}))} \tag{2.19}$$

Pour plus de clarté, nous notons $W_o = \{w_{o_1}, w_{o_2}, ..., w_{o_2m}\}$ et $\tilde{v}_o = \{v_{o_1}, v_{o_2}, ..., v_{o_2m}\}$. L'équation ci-dessus peut être simplifiée comme suit :

$$p(w_c|W_o) = \frac{exp(u_c^T\tilde{v}_o)}{\sum_{i\in V} exp(u_i^T\tilde{v}_o)} \tag{2.20}$$

Étant donné une séquence de texte de longueur $T$, le mot au pas de temps $t$ noté $w^t$ et que la taille de la fenêtre contextuelle est $m$, la fonction de vraisemblance du modèle CBOW est la probabilité de générer un mot cible central à partir des mots de contexte est présenté par la formule suivante :

$$\prod_{i=1}^{T} p(w^t|w^{t-m},...,w^{t-1},w^{t+1},...,w^{t+m}) \tag{2.21}$$

**Apprentissage du modèle CBOW :** L'apprentissage du modèle CBOW est assez similaire au modèle Skip-Gram. L'estimation du maximum de vraisemblance du modèle CBOW consiste à minimiser la fonction de perte suivante :

$$-\sum_{i=1}^{T} logp(w^t|w^{t-m},...,w^{t-1},w^{t+1},...,w^{t+m}) \tag{2.22}$$

où

$$logp(w_c|W_o) = u_c^T \tilde{v}_o - log(\sum_{i \in V} exp(u_c^T \tilde{v}_o)) \tag{2.23}$$

La même méthode d'apprentissage du modèle Skip-Gram est utilisée pour obtenir les vecteurs des mots.

## 2.4 Conclusion

Dans ce chapitre, nous avons présenté les concepts de bases de la recherche d'information et les modèles de représentation de l'information, les modèles thématiques des documents et les modèles de représentation vectorielle continue des termes.

Le modèle LDA [24] capture les relations sémantiques entre les mots au travers de l'apprentissage des thèmes latents d'une collection de documents qui sont représentés par la distribution des différents mots de la collection de documents. L'hypothèse est que les termes d'un même thème sont proches sémantiquement. Plusieurs travaux de recherche ont exploités les modèles LDA pour la tâche de recherche d'information [62-66], comme par exemple dans l'estimation des modèles de langues des documents [62, 63] ou dans l'expansion de requêtes [65, 66]. Les conclusions de ces travaux, rapportent que les modèles thématiques et principalement le modèle LDA, permettent d'améliorer les performances des systèmes.

Plusieurs travaux de recherches ont employés les plongements de mots pour améliorer les performances des systèmes de RI [67-70], ou pour la reformulation des requêtes [71]. Le modèle word2vec [25] permet de capturer les relations sémantiques entre les termes. Les résultats des expérimentations des travaux exploitant les plongements de mots ont démontré l'efficacité de cette méthode à sélectionner les termes similaires.

Principalement, dans le cadre de cette thèse, nous proposons d'employer les modèles thématiques LDA et les modèles de plongements des mots dans nos contributions. Plus précisément, nous exploitons les modèles thématiques LDA [24] pour extraire les thèmes latents dans notre collection de documents. De plus, nous employons aussi les modèles de plongements de mots words2vec [25] pour mesurer les similarités entre les termes.

# Chapitre 3

# État de l'art - Modélisation de l'utilisateur pour la RISP

## 3.1 Introduction

Dans ce chapitre, nous allons fournir un aperçu des différents travaux réalisés dans le domaine de la recherche d'information sociale personnalisée. Nous commençons d'abord par introduire la recherche d'information sociale en section 3.2. Ensuite, en section 3.3, nous présentons les travaux de l'état de l'art de la recherche d'information sociale personnalisée. La section 3.4 présente comment le profil de l'utilisateur est intégré dans un système de RISP. En section 3.5, nous présentons les collections de test. Enfin, nous présentons en section 3.6 notre positionnement par rapport aux travaux de l'état de l'art.

## 3.2 La recherche d'information sociale

Une large gamme de services et de plateformes rendent l'utilisateur de plus en plus interactif avec le web. De nombreuses informations qui concernent à la fois les utilisateurs et les ressources (documents web, images, vidéos, commentaires, tweets, tags, etc.) sont constamment générées. Ces informations peuvent être très utiles dans les tâches de recherche d'information, pour la modélisation des utilisateurs et des ressources.

Cependant, les modèles classiques de recherche d'information n'intègrent pas le contexte social de l'utilisateur et des ressources. Par conséquent, de nombreuses recherches se sont intéressées à combiner ces deux domaines qui sont la recherche d'information et les réseaux sociaux, ce qui a donné lieu à des modèles de recherche d'information sociale [72].

La recherche d'information sociale vise à fournir des contenus et des informations pertinents aux utilisateurs dans plusieurs domaines tels que la recherche d'information sociale personnalisée, la recommandation, la recherche collaborative, etc. Plusieurs plateformes [1] [2] existantes étudient cette piste afin d'améliorer le paradigme de recherche.

Le grand nombre de travaux dans ce domaine est certainement un bon indicateur de l'intérêt qu'on lui porte. Ainsi, ces travaux sont organisés en plusieurs catégories, qui sont principalement :

— **La Recherche sur le web social**. Dans ce domaine, les informations sociales sont utilisées afin d'améliorer le processus de recherche d'information classique.

1. https ://google-social-search/
2. https ://www.social-searcher.com/

Pour améliorer le processus RI classique et réduire la quantité de documents non pertinents, il y a principalement trois pistes d'amélioration : (a) reformulation des requêtes, utilisant l'expansion ou la réduction de la requête [6, 73-76], (b) le reclassement des documents récupérés (basé sur le profil ou le contexte de l'utilisateur) [4, 11, 12, 14, 77, 78], et (c) l'amélioration du modèle de RI, c'est-à-dire la façon dont les documents et les requêtes sont représentés et appariés pour quantifier leurs similitudes [12, 16, 26].

— **Recommandation sociale**, dans laquelle le réseau social de l'utilisateur est utilisé pour fournir une meilleure recommandation. Fondamentalement, la recommandation vise à prédire l'intérêt que les utilisateurs porterait à une ressource qu'il n'avait pas encore pris en compte explicitement.
Il existe deux méthodes principales de recommandation : (1) une approche "basée sur le contenu" qui a pour objectif la suggestion des éléments similaires à ceux pour lesquels l'utilisateur a manifesté son intérêt dans le passé, et (2) une approche appelée "filtrage collaboratif", qui vise à recommander des articles à l'utilisateur en fonction d'autres personnes qui ont des préférences ou des centres d'intérêts similaires.
Ces deux familles d'approches sont exploitées par exemple, pour la recommandation de ressources [79, 80], et la recommandation des utilisateurs [81, 82].

Plusieurs contributions étroitement liées au domaine de la recherche d'informations sociales sont discutées dans [83-85]. L'objectif ici n'est pas de présenter l'ensemble des contributions mais de désigner celles qui sont proches des problématiques étudiées dans cette thèse. Nous présentons principalement les travaux exploitant les plateformes sociales tels que les Microblogs (Twitter) et les folksonomies (Delicious).

## 3.3 Modélisation de l'utilisateur pour la RISP

La modélisation des utilisateurs est le processus d'acquisition, d'extraction et de représentation des centres d'intérêt des utilisateurs [15]. Le profil est employé pour présenter un contenu relatif à un utilisateur et il contient généralement des informations démographiques et des mots-clés ou des concepts qui représentent les centres d'intérêt de l'utilisateur. Enfin, les profils utilisateur construits sont évalués et peuvent être utilisés dans des applications spécifiques telles que les systèmes de recommandation ou de recherche d'information personnalisée.

Dans ce qui suit, nous discutons principalement deux dimensions du processus de modélisation des utilisateurs : (1) les données des utilisateurs et (2) la représentation des profils des utilisateurs.

### 3.3.1 Les données pour la construction des profils utilisateurs

Dans cette section, nous présentons les différents travaux exploitant les différentes types de données pour construire le profil de l'utilisateur. Les informations utilisées pour la modélisation des utilisateurs sont importantes car elles pourrait affecter directement les étapes ultérieures telles que la représentation et la qualité des profils finaux.

Nous distinguons trois types et sources d'information : (1) les contenus générés par l'utilisateur, (2) les informations provenant du réseau social de l'utilisateur et (3) les informations externes en lien avec les contenus générés par les utilisateurs.

#### 3.3.1.1 Exploitation des contenus générés par l'utilisateur

Un moyen simple d'extraire les centres d'intérêts des utilisateurs consiste à tirer parti des informations des activités des utilisateurs dans les réseaux sociaux. Dans le cas du réseau Twitter, un utilisateur peut publier (tweeter), re-tweeter, aimer ou répondre à un tweet et suivre d'autres personnes. Par conséquent, toutes les informations générées par ces activités peuvent être exploitées pour déduire les centres d'intérêts des utilisateurs [86-88].

Certaines approches proposent d'analyser les contenus et extraire des termes des tweets [86] ou des mots-clefs (hashtags) [89]. D'autres approches ont extrait des entités Wikipédia des tweets des utilisateurs [90].

En outre, les centres d'intérêts des utilisateurs peuvent être également déduits indirectement en agrégeant et en analysant les publications ou les adhésions des utilisateurs à des listes ou à des groupes thématiques. Certains travaux supposent qu'un utilisateur est intéressé par le sport s'il suit (follow) le compte Twitter de *@BEINSPORTS* [86, 87].

Pareillement, les documents annotés par les utilisateurs et les tags sont exploités pour construire les profils des utilisateurs dans les plateformes de folksonomies tel que Delicious [4, 6, 9]. Il existe deux principales approches, soit exploiter les tags des utilisateurs et les utiliser pour construire les profils [4, 6, 14, 27] ou soit exploiter les documents annotés par les utilisateurs et extraire des termes qui peuvent représenter les centres d'intérêts des utilisateurs [7, 21-23, 91].

Pour déduire les centres d'intérêts des utilisateurs à partir de leurs publications (commentaires, tweets, tags), les utilisateurs doivent être actifs, c'est-à-dire générer du contenu en continu. De plus en plus d'utilisateurs exploitent les plateformes sociales pour rechercher les informations dont ils ont besoin, sans forcément générer du contenu. Sur Facebook, deux utilisateurs sur cinq ne parcourent que les publications sans aucune interaction avec la plate-forme [92]. Certaines études ont rapporté qu'une partie importante des utilisateurs sur Twitter sont des utilisateurs passifs, qui consomment des informations sans générer de contenu [93].

Pareillement, dans Delicious, plusieurs utilisateurs sauvegardent les documents sans les avoir annotés [3]. Par conséquent, il est également important de déduire les profils des utilisateurs pour ces utilisateurs passifs.

D'autres études ont souligné que l'exploration de tweets pour déduire les intérêts des utilisateurs est instable en raison de l'évolution des centres d'intérêts des utilisateurs [94, 95].

Pour pallier a ces différentes problématiques, certains travaux proposent d'analyser les informations du cercle social des utilisateurs tels que les followers et les amis.

#### 3.3.1.2 Exploitations des relations sociales des utilisateurs

Les informations provenant des réseaux sociaux tels que les tweets, les commentaires, les publications, les tags et le cercle social des utilisateurs, peuvent être utilisées d'une part pour déduire les centres d'intérêts des utilisateurs passifs et d'autre part pour enrichir et affiner les profils des utilisateurs actifs.

Par exemple, [86, 87] ont exploré les tweets de l'utilisateur et ceux de ses abonnés pour déduire ses centres d'intérêts. Dans [96], les auteurs ont proposé de tirer parti des bibliographies des abonnés pour extraire des entités au lieu de faire correspondre les abonnés aux entités Wikipédia.

Une nouvelle approche [88], modélise un utilisateur et ses amis en employant les modèles thématique LDA. Les auteurs proposent de fusionner deux modèles ; un modèle relatif à l'utilisateur et un modèle social basé sur les amis de l'utilisateur. Étant donné que différents amis ont différentes influences sur l'utilisateur, un poids est attribué à chaque ami, composé de quatre facteurs : la popularité, l'affinité, l'interaction et le lien avec la requête de l'utilisateur.

Des approches similaires ont été proposées dans les folksonomies. Par exemple, une approche propose de générer les profils utilisateurs à partir des tags des amis de l'utilisateur [17]. D'autres approches exploitent les amis pour raffiner les poids des termes du profil de l'utilisateur [6, 21].

L'utilisation des listes Twitter peut permettre de créer de bons groupements d'utilisateurs en fonction de leurs intérêts à des thématiques particulières [8]. Par conséquent, si un utilisateur fait partie d'une liste liée à la thématique «apprentissage automatique», donc cet utilisateur est intéressé par ce sujet [96, 97].

Les centres d'intérêts des utilisateurs peuvent suivre les actualités et les tendances. Une approche [98] propose un entrelacement des tendances et des intérêts personnels des utilisateurs. En plus de tirer parti des tweets d'un utilisateur cible pour déduire les intérêts de l'utilisateur, les auteurs ont construit un profil de tendance basé sur tous les tweets de l'ensemble de données dans une certaine période de temps. Ensuite, le profil de l'utilisateur final a été construit en combinant les deux profils. Les résultats ont montré que les profils des utilisateurs combinés peuvent améliorer les performances dans la tâche de recommandations d'articles d'actualité.

Les relations sociales et les listes sont très utiles dans l'inférence des centres d'intérêt de l'utilisateur. En revanche, il est difficile de distinguer les activités des abonnés d'un utilisateur qui sont pertinents par rapport aux intérêts de cet utilisateur. Par exemple, les abonnés d'un utilisateur peuvent tweeter un large éventail de sujets qui les intéressent, et que l'utilisateur n'est pas forcément intéressé par tous ces sujets.

#### 3.3.1.3 Exploitation des données externes

L'un des défis liés à la déduction des centres d'intérêts des utilisateurs à partir des réseaux sociaux est que le contenu généré est souvent court, bruyant et ambigu- [99, 100].

Pour mieux comprendre les contenus courts des services de microblogage tels que les tweets, des informations externes de la plate-forme ont été explorées. Une approche [101, 102] propose de lier les tweets aux articles de presse et d'extraire les centres d'intérêts des utilisateurs en fonction de leurs tweets ainsi que du contenu des articles de presse associés. D'autres approches ont exploité les URLs contenus dans les tweets [103, 104].

Parallèlement, les tags aussi apportent souvent du bruit et de l'ambiguïté dans les profils des utilisateurs [100]. Ainsi, des approches proposent d'utiliser les documents annotés par les utilisateurs car ces documents apportent plus d'information et sont moins ambiguës que les tags [21-23].

Avec la popularité des réseaux sociaux, les utilisateurs ont tendance à avoir plusieurs comptes sur différentes plateformes [105]. Dans ce contexte, certains travaux ont étudié l'exploitation des profils des utilisateurs à partir d'autres réseaux sociaux.

Par exemple, une approche [89] a étudié la modélisation des utilisateurs en combinant des données de Twitter et de Flickr[3]. Cette approche donne de meilleures performances par rapport à la modélisation des utilisateurs séparément sur la base d'une seule plate-forme. Le résultat est conforme à une autre étude réalisée qui a agrégé les profils des utilisateurs sur les systèmes de marquage social tels que Delicious et Flickr [106].

Les données externes telles que le contenu des URLs intégrées dans un tweet, les documents annotés par les utilisateurs ou l'agrégation de tous leurs comptes sur les différentes plate-forme, peuvent fournir une meilleure modélisation des centres d'intérêts des utilisateurs. Néanmoins, l'analyse des données externes nécessite un effort supplémentaire et n'est pas toujours disponible. De plus, les données externes peuvent également avoir un contenu non pertinent par rapport aux intérêts des utilisateurs et peuvent donc introduire du bruit.

### 3.3.2 Représentation du profil de l'utilisateur

Jusqu'à présent, nous avons axé notre discussion sur la collecte de données provenant de diverses sources pour déduire les centres d'intérêts des utilisateurs. Dans cette section, nous présentons les différentes représentations possibles pour le profil de l'utilisateur dans le cadre de RISP.

#### 3.3.2.1 Représentation basée sur les mots clés

La représentation des centres d'intérêts des utilisateurs à l'aide de mots clés ou de groupes de mots clés a été largement utilisée dans la recherche d'information personnalisée classique mais également employée dans les réseaux sociaux [83].

Cette représentation se base sur un vecteur de mots pondérés, où chaque mot représente un centre d'intérêt. Chaque mot va avoir un poids mesurant le degré d'intérêt de ce mot pour l'utilisateur. Certaines approches [86, 97] proposent de représenter les profils des utilisateurs en utilisant des vecteurs de mots clés pondérés provenant des tweets. D'autres approches ont étudié les profils des utilisateurs basés sur les hashtags contenus dans les tweets des utilisateurs [101, 107, 108].

Pareillement, cette représentation est la plus utilisée dans les folksonomies. Les profils des utilisateurs sont basés sur les tags que les utilisateurs ont employés pour annoter les ressources web (documents, images, vidéos, etc.) [6, 10, 12, 13, 16, 18, 26, 109].

Au-delà des représentations par les mots pondérés, les approches thématiques telles que le LDA sont également populaires pour représenter les profils des utilisateurs. Par exemple, un sujet lié à la technologie de l'information peut avoir certains mots associés comme «google, twitter, apple, web». [88, 110] ont utilisé le modèle thématique LDA pour représenter chaque utilisateur comme une distribution de probabilité sur les thèmes construit sur un corpus de Tweets.

Les représentations thématiques ont également été exploitées pour représenter les utilisateurs par les thèmes latents des documents qu'ils ont annotés [22, 23, 111]. Les résultats montrent que ces représentations permettent d'améliorer les résultats.

1. **Profil basé sur les tags** :
   Dans cette catégorie de travaux, le défi est de proposer une meilleure estimation des poids des tags des utilisateurs, permettant de refléter de façon

3. https ://www.flickr.com/explore

pertinente le degré d'intérêt de chaque tag pour l'utilisateur. Le profil basé sur les tags est représenté comme suit :

$$Profil(u)_{tags} = \{< tg_1, w_{tg_1} >, < tg_2, w_{tg_2} >, ..., < tg_Y, w_{tg_Y} >\} \quad (3.1)$$

— Le schéma de pondération le plus utilisé repose sur la fréquence d'apparition du tag dans le profil de l'utilisateur, pour indiquer l'importance de ce tag pour l'utilisateur [2], qui peut être formulé comme suit :

$$w_{tg_i} = TF_{tg_i,u} \quad (3.2)$$

où $TF_{tg_i,u}$ désigne le nombre de fois que l'utilisateur $u$ a utilisé le tag $tg_i$ pour annoter ses documents.

— D'autres approches proposent de pondérer les tags en utilisant la fréquence du tag mais avec une normalisation par la taille du profil de l'utilisateur en terme de nombre de documents annotés [13, 112]. La formule de pondération est décrite comme suit :

$$w_{tg_i} = \frac{TF_{tg_i,u}}{D^u_{tg_i}} \quad (3.3)$$

où $TF_{tg_i,u}$ désigne le nombre de fois que l'utilisateur $u$ a utilisé le tag $tg_i$ pour annoter ses documents et $D^u_{tg_i}$ est le nombre de document annotés par l'utilisateur $u$ avec le tag $tg_i$.

— Pour tenir compte de la spécificité du tag, des approches proposent de combiner la fréquence du tag avec la fréquence inverse du document IDF [10, 14]. Ainsi, le poids d'un tag est calculé comme suit :

$$w_{tg_i} = TF_{tg_i,u} * IDF \quad (3.4)$$

où $TF_{tg_i,u}$ désigne le nombre de fois que l'utilisateur $u$ a utilisé le tag $tg_i$ pour annoter ses documents et $IDF$ qui désigne la fréquence inverse de document (Inverse Document Frequency) qui favorise les tags peu fréquents dans la collection des documents.

— D'autres approches de modélisation de l'utilisateur en utilisant les tags qu'il a attribué aux différents documents en ajoutant la popularité de ces tags pour les autres utilisateurs du réseau social. Ainsi, le poids $w_{tg_i}$ du tag $tg_i$ est estimé comme suit :

$$w_{tg_i} = tf(tg_i, u) \times \log(\frac{|U|}{|U_{tg_i}|}) \quad (3.5)$$

où $tf(tg_i, u_j)$ désigne le nombre de fois que l'utilisateur $u$ a utilisé le tag $tg_i$ pour annoter ses documents, $U$ est le nombre d'utilisateur et $|U_{tg_i}|$ le nombre d'utilisateur ayant utilisé le tag $tg_i$
[6, 12, 18, 113]

2. **Profil basé sur les documents** :
Plusieurs travaux ont été proposés pour modéliser l'utilisateur au travers des documents dans le cadre de la recherche d'information personnalisée. Par contre très peu de travaux ont été proposés pour modéliser l'utilisateur avec les documents dans la recherche d'information personnalisée dans les folksonomies. A notre connaissance seuls les travaux [7, 21-23, 91] ont été proposés dans ce cadre.

Les documents annotés par les utilisateurs sont une source riche d'informations. Ces documents offrent un contenu plus large que les tags et ce contenu pourrait encore aider à identifier les centres d'intérêts de l'utilisateur et ainsi améliorer les performances des systèmes de recherche. De plus, un utilisateur pourrait sauvegarder un document sans forcément l'avoir annoté ou utiliser très peu de tags. Donc, ceci indique que potentiellement le contenu des documents est particulièrement intéressant et peut fournir plus d'informations liées aux centres d'intérêts de l'utilisateur.

— Une première approche propose de prendre en compte tous les documents d'un utilisateur pour le représenter et propose ainsi un profil basé sur les termes des documents Les auteurs [7]. La pondération des termes du profil est estimée par modèle de langue standard.
— D'autres approches proposent de modéliser l'utilisateur par des thèmes latents en utilisant les modèles thématiques comme LDA [21]. Dans [22, 23], les auteurs proposent d'utiliser aussi les modèles thématiques latents (LDA) mais en intégrant les tags des utilisateurs dans l'estimation des thèmes latents des documents.

#### 3.3.2.2 Représentation basée sur les concepts

Certaines approches ont proposé de tirer parti des concepts des bases de connaissances tels que DBpedia pour représenter les centres d'intérêts des utilisateurs. L'un des avantages de l'exploitation des bases de connaissances est que nous pouvons exploiter les connaissances de base de ces concepts pour déduire les centres d'intérêts des utilisateurs qui pourraient ne pas être capturés si nous utilisons des approches basées sur des mots clés. Par exemple, un utilisateur intéressé par de la société Apple serait intéressé par tout nouveau produit d'Apple même si les noms de ces nouveaux produits n'ont jamais été mentionnés dans le profil de l'utilisateur [114].

Les concepts des différents types de bases de connaissances ont été exploités pour différentes fins de modélisation des utilisateurs, telles l'exploitation des taxonomies de concept pour la recommandation des actualités [115], ou les bases de connaissances spécifiques à un domaine tel que le médical [116-118].

Différentes stratégies de représentation à l'aide des concepts ont été explorées :

— **Représentation par les entités** : cette approche extrait les entités à partir des données des utilisateurs et utilise ces entités pour représenter les centres d'intérêts d'un utilisateur [119-121].
— **Représentation par les catégories** : cette approche exploite les catégories de DBpedia qui permettent d'avoir une représentation plus générale des centres d'intérêts des utilisateurs à la différence de la représentation par les entités [90, 122-124].
— **Représentation hybride** : cette approche combine la représentation des profils par les entités et par les catégories [90, 104, 125, 126].

Les approches basées sur les concepts exploitent la sémantique entre les concepts et peuvent tirer parti des connaissances de base sur les concepts pour construire les profils des utilisateurs. Par contre, ces approches reposent sur des bases de connaissances préexistantes ou pré-construites qui ne sont pas toujours disponibles ou manquent de couverture pour certains domaines.

## 3.4 Intégration du profil utilisateur dans un système de RISP

Une fois que le profil de l'utilisateur est construit, ce profil est intégré au système de RI. Il existe principalement trois méthodes : (1) modification de la requête, (2) intégration du profil pendant la recherche, et (3) reclassement des résultats de la recherche. Dans cette partie, nous allons présenter ces trois méthodes pour RISP dans le cadre des folksonomies.

### 3.4.1 Modification de la requête

Cette méthode permet de modifier la requête originale de l'utilisateur en ajoutant des termes du profil de l'utilisateur. Certaines approches [4, 9, 74, 112], proposent de fusionner la requête originale avec les termes du profil de l'utilisateur avec une simple combinaison. Ainsi, tout le profil de l'utilisateur est pris en compte dans le calcul des scores.

Certaines approches proposent de sélectionner les termes du profils par des méthodes de sélection sémantique. Donc, pour chaque terme de la requête, les top-k termes similaires aux termes sont sélectionnés. Les méthodes peuvent d'appuyer sur des approches exploitant des matrices de cooccurrences [15, 76, 127], des plongement de mots (word2vec) [27] ou des modèles thématiques [15].

D'autres approches proposent d'inclure le cercle social de l'utilisateur [6, 17, 31] pour enrichir le profil de l'utilisateur ou étendre la requête avec des termes du profil du cercle social.

#### 3.4.1.1 Intégration du profil pendant la recherche

Dans cette méthode, le profil est intégré pendant l'étape d'appariement. C'est-à-dire, une fusion entre le score de correspondance de la requête avec le document RSV(q,d) avec le score de correspondance du document avec le profil RSV(d,u) [11, 13, 14, 16, 21, 28]. Généralement, une combinaison linéaire est employée pour calculer le score final. Cette méthode permet de contrôler les deux scores, le score RSV(q,d) et le score RSV(d,u).

Certaines approches [26] proposent de calculer d'abord un score RSV(q,d). Ensuite, sélectionner le top-k document. Et enfin, faire une combinaison linéaire entre le RSV(q,d) et le RSV(d,u) sur les topk-k documents. Cette méthode, permet d'une part, de réduire le nombre de documents et d'autre part, de s'assurer que la requête est toujours prise en compte. Cette méthode peut être efficace et apporter de bons résultats dans le cas où les profils des utilisateurs sont longs.

#### 3.4.1.2 Reclassement des résultats

Plusieurs approches ont été proposées pour personnaliser le classement des résultats de recherche en utilisant les informations sociales [2, 10, 13, 14, 18, 109, 128]. Le processus repose sur deux étapes. En première étape, le système calcule un score de pertinence entre la requête de l'utilisateur et les documents. La seconde étape consiste à reclasser les top-k documents suivant leur score de pertinence avec le profil utilisateur. La première étape est généralement commune à toutes les approches de reclassement des résultats. Par contre, plusieurs travaux proposent des approches différentes pour la seconde étape.

Des approches [109] ont étudié le reclassement des résultats personnalisés en fonction des relations sociales de l'utilisateur. D'autres approches ont exploré la collecte de données à partir de plusieurs systèmes sociaux pour la personnalisation [128].

D'autres approches [12, 18], proposent d'utiliser les données sociales et les relations avec les utilisateurs dans le reclassement des résultats.

## 3.5 Évaluation des systèmes de RISP

L'évaluation est une étape permettant de mesurer la capacité des systèmes de recherche d'information à satisfaire les besoins d'information des utilisateurs.

Habituellement, un système de RI classique est évalué au travers d'une collection de test comprenant 3 éléments essentiels qui sont :

— **Un ensemble de requêtes** qui représentent le besoin en information des utilisateurs. Cet ensemble est fourni afin d'évaluer pour chaque requête la capacité du système à retrouver les documents pertinents.
— **Un ensemble de documents** qui contiennent les informations pertinentes qui répondent à l'ensemble des requêtes des utilisateurs.
— **Les jugements de pertinence** qui permettent d'identifier les documents pertinents pour chaque requête. Ces jugements sont fournis généralement par des utilisateurs.

L'évaluation d'un système de RI sociale personnalisée est un énorme défi. Les jugements de pertinences doivent être exclusivement et impérativement fournis par les utilisateurs ayant soumis les requêtes au système de recherche d'information personnalisée.

La difficulté est qu'il existe très peu de collection de tests standard pour l'évaluation de l'efficacité des systèmes de recherche d'information sociale personnalisée. Fournir de telles collections est très complexe par le fait de la difficulté de la mise en place d'un système qui permet de collecter les informations des utilisateurs, leurs requêtes et ensuite les jugements de pertinence pour tous les documents pour chaque requête pour chaque utilisateur. Certains travaux se sont intéressés à fournir un cadre expérimental complet pour les systèmes de folksonomies [19, 129, 130]. Ce cadre expérimental sera présenté dans le chapitre 6, que nous allons suivre pour construire notre collection de test.

### 3.5.1 Collections de test

Dans cette section, nous présentons les deux principales collection de test dans le cadre de la RISP, qui sont : CLEF avec la tâche : Social Book Search [4] et TREC avec la tâche : Contextual Suggestion [5].

#### 3.5.1.1 TREC Contextual Suggestion

La suggestion contextuelle consiste à rechercher des besoins d'information complexes qui dépendent à la fois du contexte et des intérêts des utilisateurs.

Cette tâche est définie sous la forme de la tâche de recommandation de point d'intérêt personnalisé, dans laquelle le système de recommandation fournit une liste

4. http ://social-book-search.humanities.uva.nl/#/overview
5. https ://sites.google.com/site/treccontext/

classée de suggestions en fonction du profil de l'utilisateur et du contexte dans lequel l'utilisateur recherche la suggestion. Un exemple de profil utilisateur et son contexte est représenté dans la figure 3.2.

Dans la suggestion contextuelle, étant donné les informations des utilisateurs, y compris leur âge, leur sexe et l'ensemble des lieux ou activités notés selon les préférences de l'utilisateur (les notes sont comprises entre -1 et 4), la tâche consiste à générer une liste de suggestions classées à partir de un ensemble d'attractions candidates, en donnant à l'utilisateur des informations ainsi que des informations sur le contexte, notamment le lieu du voyage, la saison du voyage, la durée du voyage et le type de groupe avec lequel la personne voyage.

Les suggestions sont extraites de plusieurs ressources web comme décrit dans la figure 3.1.

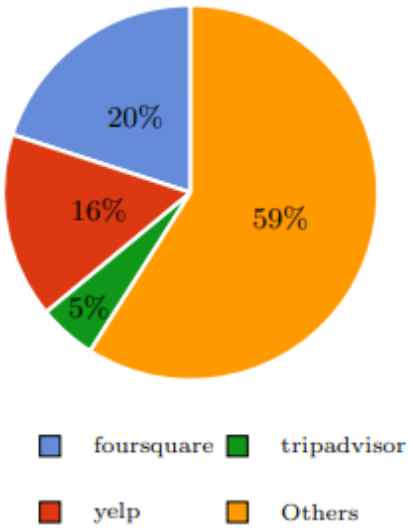



FIGURE 3.1 – Domaines les plus populaires dans la collection TREC Contextual Suggestion (2016) [131]

La collection de tests de suggestions contextuelle TREC 2016 comprend plus de 956.437 documents, plus de 60 profils utilisateurs et des jugements de pertinence décrits par des notes (de -1 à 4) assignées par l'utilisateur.

```
{"id":743,
"body": {
    "group": "Friends",
    "season":"Summer",
    "trip_type":"Holiday",
    "duration":"Weekend trip",
    "location":{
        "state":"TX",
        "id":306,
        "name":"Waco",
        "lat":31.54933,
        "lng":-97.14667},
    "person": {
                "gender": "Male",
                "age": 28,
                "id": 15012,
                "preferences":[
                    {
                    "rating":4,
                    "documentId":"TRECCS-00211395-161",
                    "tags":[
                         "Beer",
                         "Culture",
                         "Cocktails",
                         "Restaurants",
                         "Food",
                         "pub-hopping",
                         "cocktails",
                         "bar-hopping"
                         ]},
                    ...
                    ]
                    }},
"candidates":[
    {"documentId":"TRECCS-00267253-306",
    "tags":[
        "Beer",
        "Cocktails",
        "Family Friendly",
        "Restaurants",
        "Food"
        ]},
    {"documentId":"TRECCS-00294259-306",
    "tags":[
    "Tourism",
    "Bar-hopping",
    "Restaurants",
    "Entertainment",
    "Live Music"
    ]},
     ...
    ]
    }
```


FIGURE 3.2 – Exemple de requête TREC Contextual Suggestion

#### 3.5.1.2 CLEF Social Book Search

La recherche sociale de livre est une tâche qui a pour but d'évaluer les approches pour aider les utilisateurs à rechercher des collections de livres sur la base des métadonnées et contenu généré par l'utilisateur associé.

La collection 2016 contient 2,8 millions de livres sous la forme des méta-données au format XML collectés sur le site Amazon.com, 380 profils d'utilisateurs et 680 topics extraits du forum de réseau social d'amateurs de littérature LibraryThing [6] et 8.918 jugements de pertinence [132].

La figure 3.3 présente un exemple de requête et un profil utilisateur. Comme présenté dans la figure 3.3, l'utilisateur est représenté par un ensemble de livres achetés (par exemple "Daughter of the Forest". Chaque livre est associé à un ensemble de tags et de notes attribuées par les différents utilisateurs qui ont acheté ce livre.

### 3.5.2 Les mesures d'évaluation

Il existe de nombreuses métriques pour évaluer la qualité des systèmes de RISP. Nous présentons dans cette section, les métriques utilisées dans cette thèse.

— Le **Rappel** est une mesure qui estime la capacité d'un système à retourner ***tous*** les documents pertinents comme décrit par la formule suivante :

$$Rappel = \frac{\text{Nombre de document pertinents retournés par le système}}{\text{Nombre de document pertinents dans la collection}} \tag{3.6}$$

— La **Précision** est une mesure qui estime la capacité d'un système à retourner ***que*** les documents pertinents comme décrit par la formule suivante :

$$Prcision = \frac{\text{Nombre de document pertinents retournés par le système}}{\text{Nombre de document retourné par le système}} \tag{3.7}$$

— La **F-mesure** est une mesure qui estime une moyenne harmonique pondérée de la précision et du rappel et est décrite par la formule suivante :

$$F - mesure = \frac{2 \times Precision \times Rappel}{Precision + Rappel} \tag{3.8}$$

— La **MAP (Mean Average Precision)** est une mesure qui estime la moyenne des précisions moyennes sur l'ensemble des requêtes permettant ainsi de mesurer la performance global d'un système, et est décrite par la formule suivante :

$$MAP = \frac{1}{|Q|} \sum_{q_i \in Q} \frac{1}{r} \sum_{r \in R} Precision(q_i)@R \tag{3.9}$$

tel que, Q représente l'ensemble des requêtes, r représente le nombre de document pertinents pour la requête $q_i$ et R le rang du document pertinent.

— Le **P-Gain** mesure le gain personnalisé d'un système par rapport à un autre système. Les taux sont calculés aux niveaux des performances au niveau de la requête en relevant les différences en valeurs de précision moyenne AP (Average Precision) obtenues entre deux systèmes. Cette mesure est estimée comme suit :

$$P\text{-}Gain = \frac{\#Q_+ - \#Q_-}{\#Q_+ + \#Q_-} \tag{3.10}$$

où $\#Q_+$ représente le nombre de requêtes améliorées, $\#Q_-$ représente le nombre de requêtes détériorées par rapport à un système.

6. https ://www.librarything.com/

```
<topics>
<topic>
    <topicid>107277</topicid>
<request>Greetings! I'm looking for suggestions of fantasy novels whose heroines are creative in some way and have some sort of talent in art, music, or literature. I've seen my share of "tough gals" who know how to swing a sword or throw a punch but have next to nothing in the way of imagination. I'd like to see a few fantasy-genre Anne Shirleys or Jo Marches.
Juliet Marillier is one of my favorite authors because she makes a point of giving most of her heroines creative talents. Even her most "ordinary" heroines have imagination and use it to create. Clodagh from "Heir to Sevenwaters," for example, may see herself as being purely domestic, but she plays the harp and can even compose songs and stories. Creidhe of "Foxmask" can't read, but she can weave stories and make colors. The less ordinary heroines, like Sorcha from "Daughter of the Forest" and Liadan from "Son of the Shadows," are good storytellers. I'm looking for more heroines like these. Any suggestions?

</request>
    <group>FantasyFans</group>
    <title>Fantasy books with creative heroines?</title>
<examples>
      <work>
        <booktitle>Daughter of the Forest</booktitle>
        <author>Juliet Marillier</author>
        <workid>6442</workid>
      </work>
      <work>
        <booktitle>Foxmask</booktitle>
        <author>Juliet Marillier</author>
        <workid>349475</workid>
      </work>
      <work>
        <booktitle>Son of the Shadows</booktitle>
        <author>Juliet Marillier</author>
        <workid>6471</workid>
      </work>
      <work>
        <booktitle>Heir to Sevenwaters</booktitle>
        <author>Juliet Marillier</author>
        <workid>5161003</workid>
      </work>
    </examples>
    <catalogue>
      <work>
        <tags/>
        <rating>0.0</rating>
        <publication-year>2002</publication-year>
        <booktitle>Blue Moon (Anita Blake, Vampire Hunter, Book 8)</booktitle>
        <cataloging-date>2011-08</cataloging-date>
        <author>Laurell K. Hamilton</author>
        <workid>10868</workid>
      </work>
    <catalogue>
  </topic>
```

FIGURE 3.3 – Exemple de requête CLEF Social Book Search

## 3.6 Discussion

Dans cette section, nous présentons notre positionnement par rapport à l'état de l'art et argumentons nos propositions.

Nous allons discuter les trois points principaux : (1) les données que nous exploitons pour représenter le profil de l'utilisateur, (2) la représentation des profils utilisateurs que nous utilisons, et (3) les formules de pondération des termes du profils que nous proposons.

### 3.6.1 Les données de l'utilisateur

Comme nous l'avons présenté en section 3.3.1, certains travaux proposent d'utiliser les données qui sont générées par les utilisateurs eux-mêmes. Des travaux proposent d'exploiter les tweets postés par les utilisateurs [87-90], d'autres travaux exploitent les tags que les utilisateurs ont employés pour annoter des documents [6, 11, 14, 17], ou les documents annotés par les utilisateurs [7].

Certains travaux [17, 21, 88] proposent d'exploiter les données du cercle social de l'utilisateur. L'exploitation du cercle social de l'utilisateur à deux avantages. D'une part, il permet d'enrichir les profils des utilisateurs, et d'autre part, de construire des profils pour les utilisateurs passifs.

Une autre catégorie de travaux [103, 104] propose d'intégrer des informations externes extraites du contenu des URLs intégrées dans un tweet.

Malgré que nous soyons conscient de l'apport des relations sociales des utilisateurs dans l'estimation des profils des utilisateurs, cependant, nous ne les prenons pas en compte dans nos travaux. La raison principale de ce choix est le manque de données sur ces relations dans les collections de test dont nous disposons.

De plus, nous n'exploitons pas les données externes car d'une part, nous ne disposons pas de ces données et d'autre part, les travaux exploitant ces données n'ont pas conclu de façon sûre l'avantage d'exploitation des données externes. De plus, ces données peuvent introduire le bruit si le contenu des URLs n'est pas pertinent pour un utilisateur.

Donc, dans cette thèse, nous proposons d'exploiter seulement les données générées par les utilisateurs eux-mêmes. Plus précisément, nous proposons des profils qui se basent sur les tags des utilisateurs et des profils des utilisateurs basés sur les termes des documents annotés.

### 3.6.2 Représentation du profil de l'utilisateur

En section 3.3.2, nous avons présenté deux principales approches de modélisation des utilisateurs. Une première représentation est basé sur les mots clés [6, 10, 12, 13, 16, 18, 26, 109]. Dans cette approche, nous distinguons une famille d'approche qui utilisent des termes et une seconde famille qui exploitent les modèles probabilistes latents comme le LDA [22, 23, 111]. Nous n'avons pas noté de travaux exploitant les thèmes latents sur les tags des utilisateurs.

Une seconde approche qui exploite les concepts des bases de connaissances tels que DBpedia pour représenter les centres d'intérêts des utilisateurs [90, 114, 122-124].

Même si les approches basées sur les concepts exploitent la sémantique entre les concepts pour construire les profils des utilisateurs. Par contre, ces approches

reposent sur des bases de connaissances préexistantes qui ne sont pas toujours disponibles. Les profils représentés par les mots-clés sont les plus simples à créer et ne s'appuient pas sur des connaissances externes issues d'une base de connaissances.

Dans le cadre de cette thèse, nous employons la représentation par mots-clés pour construire les profils des utilisateurs. Cette méthode est la plus utilisée dans les travaux de folksonomies et qui est la plus adéquate à nos propositions.

### 3.6.3 Pondération des termes du profil de l'utilisateur

Comme présenté dans les deux points précédents, nous proposons des modèles utilisateurs qui se basent sur les folksonomies. Dans la section suivante, nous présentons la problématique que nous traitons dans cette thèse.

#### Définition de la problématique

Soit un utilisateur $u$ ayant annoté un ensemble de document $D_u = \{d_1^u, d_2^u, ..., d_N^u\}$. L'utilisateur a attribué un ensemble de tags $T_{d_i}^u = \{tg_1, tg_2, ..., tg_M\}$ pour chaque document $d_i^u \in D_u$.

L'objectif principal est de construire :

1. Un profil utilisateur basé sur les tags et d'attribuer pour chaque tag $tg_i$ de l'ensemble de tags de l'utilisateur $V_{TG}^u$ un poids $w_{tg_i}$ qui doit refléter au mieux les centres d'intérêts de l'utilisateur.
   Donc le profil de l'utilisateur $P_{TG}^u$ basé sur les tags va être représenté comme suit :

$$P_{TG}^u = \{< tg_1, w_{tg_1} >, < tg_2, w_{tg_2} >, ..., < tg_Y, w_{tg_Y} >\} \tag{3.11}$$

2. Un profil pour l'utilisateur basé sur les termes des documents et d'attribuer pour chaque terme $t_o$ de l'ensemble du vocabulaire de la collection $V_T$ un poids $w_{t_o}$.
   Donc le profil de l'utilisateur $P_C^u$ basé sur le contenu va être représenté comme suit :

$$P_C^u = \{< t_1, w_{t_1} >, < t_2, w_{t_2} >, ..., < t_S, w_{t_S} >\} \tag{3.12}$$

#### Proposition 1 : profil basé sur les tags du document

Dans le cadre de recherche d'information personnalisée dans les folksonomies, les tags des utilisateurs employés pour annoter les documents sont souvent employé pour déterminer leurs centres d'intérêts [2, 6, 7, 9-16].

Certaines approches proposent une représentation vectorielle des tags de l'utilisateur [2]. Le poids de chaque tag suit un schéma de pondération basé sur la fréquence d'utilisation de ce tag par l'utilisateur (TF). En plus d'une pondération basée seulement sur la fréquence des tags, d'autres travaux combinent la fréquence du tag (TF) et la fréquence utilisateur inversée (IUF : Inverse User Frequency) qui est une adaptation du modèle TF-IDF [10]. Une approche hybride [14] propose de combiner le modèle TF-IUF avec le modèle BM25 pour pondérer les tags de l'utilisateur.

Toutes ces méthodes de pondération tentent de trouver la meilleure fonction de pondération des tags de l'utilisateur afin de donner plus d'importance aux tags qui représentent mieux les centres d'intérêts de l'utilisateur. Dans cette optique, des travaux [11, 13] étudient les différentes méthodes de pondérations des tags proposées

pour représenter un utilisateur. Les auteurs ont comparé et détaillé les limites des méthodes basées sur TF, TF-IUF, et BM25. Ainsi, ils proposent une nouvelle fonction de pondération NTF (Normalized Tag Frequency) qui souligne la préférence de l'utilisateur à utiliser un tag pour annoter ses documents. Donc, le poids de chaque tag doit tenir compte non seulement de la fréquence du tag dans le profil de l'utilisateur mais aussi du nombre de documents annoté avec ce tag pour le document, ce qui reflète en quelque sorte, la popularité du tag.

Plusieurs études ont été menées pour comprendre les structures des systèmes de folksonomies [3]. Les auteurs ont constaté que 50% des documents annotés par un tag particulier, ce tag est un terme du document. Ceci démontre que certains utilisateurs ont tendance à employer des termes du document comme tags pour les annoter car de leur point de vue ces termes décrivent bien le sujet du document. Une étude plus approfondie [19] a rapporté que les tags des utilisateurs sont corrélés avec leurs motivations d'annotation. Les auteurs ont identifié que les utilisateurs ont différentes motivations dans leur activité d'annotation. Un utilisateur peut annoter un document dans le but de le retrouver par la suite par une recherche de mots-clés en employant le tag comme requête, pour une classification de tâche, une note pour une tâche à faire "to read", etc. Par conséquent les tags d'un utilisateur peuvent être de différents types suivant ses motivations.

Pour tenir compte des variétés des types de tags que pourrait avoir un utilisateur, des travaux [20] proposent une classification des tags d'un utilisateur en plusieurs catégories : Tags libres, tags généraux, tags spécifiques, tags synonymes, tags contextuels, tags subjectifs. Cette classification de tags est couramment utilisée dans la tâche de recommandation.

En plus des différents types de tags, un autre problème souligné est le bruit qui peut être généré par les tags, comme les tags mal orthographiés, tags combinés, etc. Pour réduire ce bruit et que l'utilisateur ne soit pas représenté par des tags non pertinents, certains travaux se sont intéressés à inclure le réseau social de l'utilisateur. Par exemple, une version de TF-IUF est proposée et tient compte du réseau social des utilisateurs [12, 18]. L'idée globale est que si plusieurs utilisateurs attribuent un même tag à un même document, alors potentiellement ce tag est un tag non erroné et donc éventuellement, ce tag traite du sujet du document.

Toutes les approches citées ci-dessus tentent de trouver la meilleure pondération des tags de l'utilisateur afin d'avoir la meilleure représentation des centres d'intérêts de l'utilisateur et les sujets auxquels il s'intéresse. Toutes les pondérations des tags présentées ici ignorent un élément important qui est le document. Plus précisément, la non prise en compte du lien entre les tags et le document. En effet, la pondération des tags est prise de façon extrinsèque du contenu du document. c'est-à-dire, le poids d'un tag est considéré indépendamment du contenu du document auquel il est associé. Un tag peut avoir un poids important et peut être considéré comme étant pertinent pour représenter l'utilisateur même s'il n'a aucune relation avec le contenu thématique du document et inversement.

Un autre point important est la pondération des tags. Si nous supposons que les tags sont bien sélectionnés et décrivent le document, ces tags ne doivent pas avoir tous la même importance et le même degré de description du contenu du document auxquels ils sont associés. Si un utilisateur a attribué des tags à un document, naturellement, le tag qui décrit le contenu document est plus à même de décrire le centre d'intérêt de l'utilisateur. De plus, même si ces tags décrivent tous le contenu du document, ces tags peuvent ne pas avoir la même importance et le même degré de description du contenu du document, du point de vue de l'utilisateur. Donc,

nous devons prendre en compte deux éléments importants : le fait qu'un tag décrit le contenu du document et la pondération de ce tag.

Nous rappelons que notre objectif est de construire des profils utilisateurs afin de les intégrer dans des systèmes de recherche d'information personnalisée. Le type d'information que nous souhaitons mettre en avant est de type contenu afin de retourner des documents qui répondent aux requêtes de type contenu, aux utilisateurs. Comme les profils que nous souhaitons construire sont de types contenu et sont basés sur les tags, alors l'importance des tags doit être en corrélation avec les thématiques des documents annotés par l'utilisateur.

La contribution principale de nos travaux réside dans la définition d'un modèle utilisateur représenté par les tags, tel que ces tags doivent couvrir les sujets des documents. Ainsi, nous proposons différentes méthodes de pondérations des tags associés à un document. Cette proposition sera présentée en détails dans le chapitre 4.

**Proposition 2 : profil basé sur les termes du document**

La plupart des travaux de recherche dans les folksonomies modélisent les utilisateurs au travers des tags qu'ils ont attribué aux différents documents [2, 6, 7, 9-16, 18].

Les documents annotés par les utilisateurs sont une source riche d'informations. Ces documents offrent un contenu plus large que les tags et ce qui pourrait encore aider à identifier les centres d'intérêts de l'utilisateur et ainsi améliorer les performances des systèmes de recherche d'information. De plus, un utilisateur pourrait être intéressé par un document et le sauvegarder sans forcement l'avoir annoté ou pourrait avoir utilisé très peu de tags pour l'annoter. Donc, ceci indique que potentiellement que le contenu des documents est particulièrement intéressant et peut donc fournir plus d'information liée aux centres d'intérêts de l'utilisateur.

Certains travaux [7] prennent en compte tous les documents d'un utilisateur pour le représenter et proposent ainsi un profil basé sur les termes des documents. La pondération des termes du profil est estimée par un modèle de langue standard. D'autres approches proposent de modéliser l'utilisateur par des thèmes latents en utilisant les modèles thématiques comme LDA [21].

D'autres approches [22, 23] proposent d'utiliser aussi les modèles thématiques latents (LDA) mais en intégrant les tags des utilisateurs dans l'estimation des thèmes latents des documents.

Dans ces travaux de l'état de l'art, les auteurs proposent d'employer les documents pour représenter les centres d'intérêts des utilisateurs. Cependant, les utilisateurs ayant annoté les mêmes documents vont avoir des profils similaires même si ces utilisateurs utilisent des tags différents pour annoter ces documents.

De plus, un utilisateur pourrait avoir des intérêts différents pour un même document. C'est-à-dire, pour un document couvrant 3 thèmes, l'utilisateur peut n'être intéressé que par une seule thématique et donc potentiellement n'annoterai le document qu'avec des tags couvrant cette thématique. Donc, les autres thématiques du document n'intéressent pas trop l'utilisateur et donc les termes de ces thématiques ne devraient donc pas être pris en compte ou du moins avoir des poids moins importants par rapport aux autres termes décrivant les thématiques importantes pour l'utilisateur.

Notre intuition est que les utilisateurs peuvent être intéressés par des sujets différents pour un même document, et que leur intérêt par le document est reflété par le tags qu'ils emploient pour l'annoter.

Par exemple, un utilisateur qui fait une recherche sur les méthodes de lissage, il annote par exemple, un document qui traite de ce sujet avec les tags "to read", "Dirichet", "Smoothing". Notre hypothèse estime que les tags de l'utilisateur sont un bon moyen de déterminer les termes importants du document. Par contre dans le cas de l'exemple, notre hypothèse ne vas pas marcher et peut donc générer des erreurs dans l'estimation des termes importants. Car, des termes importants du document vont voir leur poids diminués voir éliminés vu qu'ils de ne disposent pas de lien avec le tag "to read". Ce problème nous l'avons constaté dans le modèle parcimonieux que nous avons proposé [16].

Donc, nous proposons de tenir compte de ces caractéristiques et ainsi nous proposons un modèle utilisateur représenté par les documents qu'il a annoté. Spécifiquement, nous proposons de ne prendre en compte que les termes du document qui sont en lien avec les tags de l'utilisateur qu'il a attribués à ce document.

La contribution décrite dans le chapitre 5 est de proposer des modèles d'estimation des poids pour les termes des documents annotés par les utilisateurs, donc les termes qui vont décrire le profil de l'utilisateur. Ces termes sont ceux qui sont liés aux tags attribués au document par cet utilisateur. De plus, les tags que nous prenons en compte doivent être représentatifs des centres d'intérêts de l'utilisateur. Ainsi, nous intégrons notre contribution précédente qui est l'estimation des tags importants d'un document.

# Troisième partie

# Contributions

# Chapitre 4

# Modèle de l'utilisateur basé sur les tags - Exploitation des documents pour la pondération des tags

## 4.1 Introduction

Les systèmes de folksonomies deviennent de plus en plus populaires. Par exemple, Delicious permet aux utilisateurs d'annoter des documents et de les partager avec d'autres utilisateurs de la plateforme. Les documents et les tags postés par les utilisateurs sont supposés être fortement liés à leurs centres d'intérêts. En particulier, les tags émis par les utilisateurs fournissent des informations riches permettant de créer des profils utilisateurs. Compte tenu des caractéristiques des systèmes de folksonomies, les chercheurs considèrent que la construction des profils des utilisateurs à partir des tags est essentielle pour la recherche d'information sociale personnalisée.

Dans ce chapitre, nous présentons notre première contribution de modèle utilisateur dans lequel nous nous intéressons à la représentation des centres d'intérêts de l'utilisateur au travers de leurs tags. Nous proposons une nouvelle approche de construction de profil utilisateur en intégrant les documents dans l'estimation des poids des tags de l'utilisateur. En outre, nous proposons trois modèles permettant d'identifier les tags qui décrivent de façon pertinente les centres d'intérêts de l'utilisateur.

La suite du chapitre est comme suit : la section 4.2 détaille l'architecture de notre approche, les différents modèles utilisateur que nous proposons et les notations utilisées dans ce chapitre. En section 4.3, nous présentons nos propositions d'estimation des modèles de tags des documents pour identifier les tags pertinents. La construction du profil de l'utilisateur est décrite en section 4.4. Enfin, en section 4.5, nous présentons comment nous intégrons le profil utilisateur dans le système de RISP.

## 4.2 Approche de modélisation de l'utilisateur basé sur les tags

Soit un utilisateur $u$ ayant annoté un ensemble de document $D_u = \{d_1^u, d_2^u, ..., d_N^u\}$. L'utilisateur a attribué un ensemble de tags $T_{d_i}^u = \{tg_1, tg_2, ..., tg_M\}$ pour chaque document $d_i^u \in D_u$.

L'objectif principal est de construire un profil utilisateur basé sur les tags et d'attribuer pour chaque tag $tg_i$ de l'ensemble de tags de l'utilisateur $V_{TG}^u$ un poids $w_{tg_i}$ qui doit refléter au mieux les centres d'intérêts de l'utilisateur.

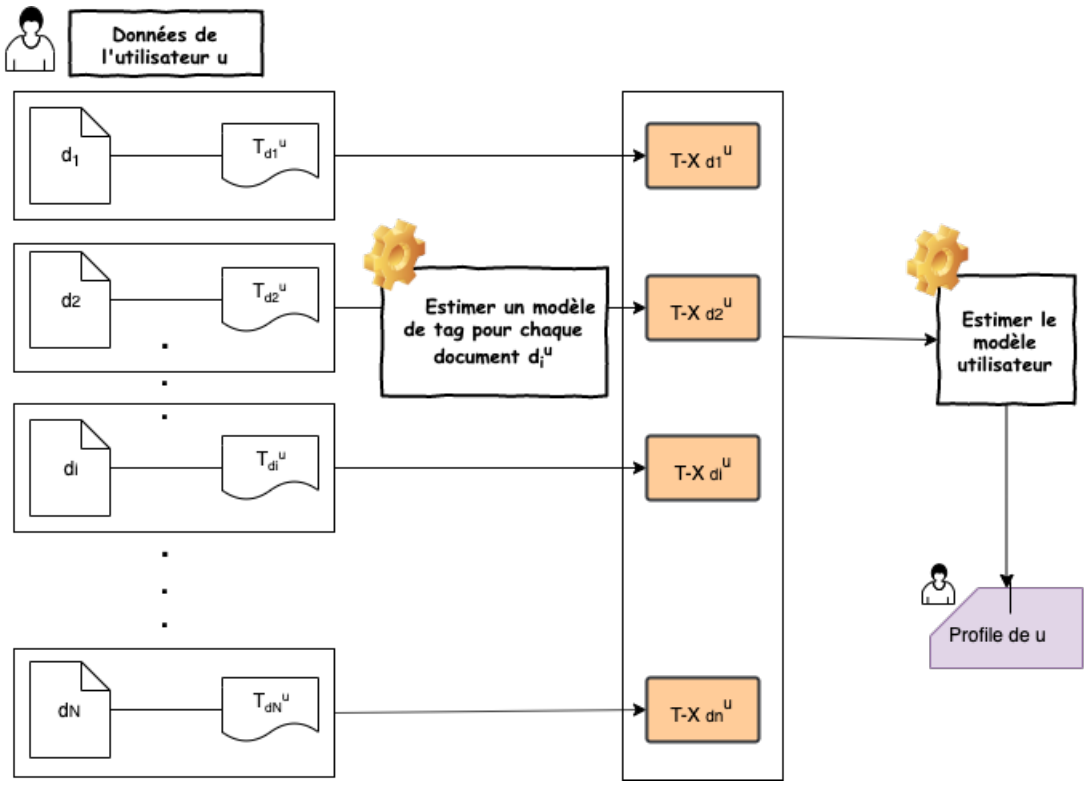



FIGURE 4.1 – Approche globale de construction du profil de l'utilisateur basé sur les tags

Donc le profil de l'utilisateur $P^u_{TG}$ basé sur les tags va être représenté comme suit :

$$P^u_{TG} = \{< tg_1, w_{tg_1} >, < tg_2, w_{tg_2} >, ..., < tg_Y, w_{tg_Y} >\} \tag{4.1}$$

L'architecture globale de notre approche qui est présentée dans la figure 4.1.
Comme présenté dans la figure 4.1 :

**Information de l'utilisateur**

Dans notre approche, les informations que nous exploitons pour modéliser l'utilisateur sont les documents que l'utilisateur lui-même a annoté et les tags qu'il a attribués à chaque document.

Plus formellement, un utilisateur $u$ est associé à un ensemble de document $D_u = \{d^u_1, d^u_2, ..., d^u_N\}$ qu'il a annoté, où chaque document $d^u_i \in D_u$ est associé à un ensemble de tag $T^u_{d_i} = \{tg_1, tg_2, ..., tg_M\}$ attribué par l'utilisateur $u$.

**Représentation du profil de l'utilisateur**

Pour représenter le profil de l'utilisateur, nous employons la représentation vectorielle qui est la plus utilisée dans les travaux de l'état de l'art [2, 10-14, 18, 84].

La modélisation que nous proposons repose sur un profil utilisateur basé sur les tags. En conséquence le profil de l'utilisateur est représenté par un vecteur de tag, où chaque tag $tg_i$ va avoir un poids $w_{tg_i}$ qui estime l'importance du tag pour l'utilisateur comme présenté dans 4.1 .

**Construction du profil**

La construction du profil de l'utilisateur est réalisée en deux étapes.

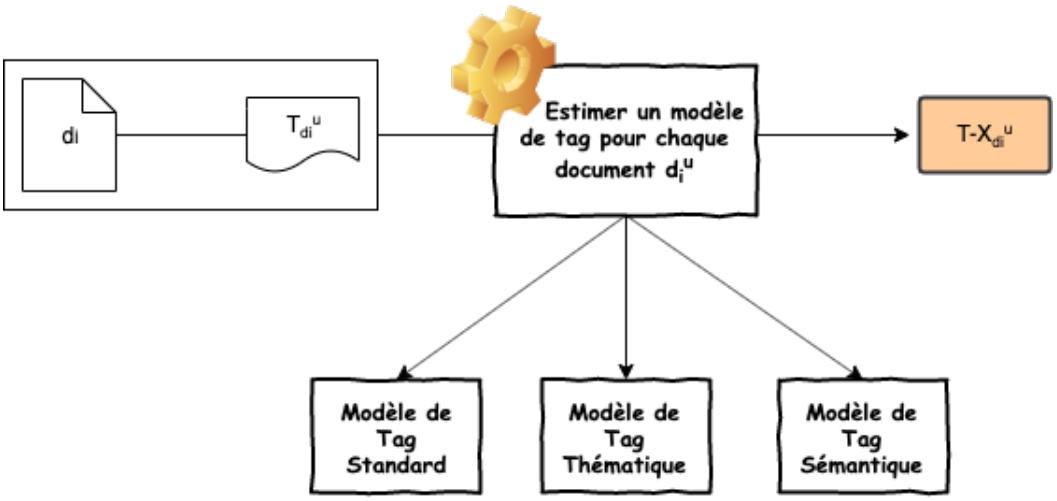



FIGURE 4.2 – Les différents modèles de Tag d'un document

1. La première étape consiste à estimer les modèles de tags des documents annotés par l'utilisateur. Donc, pour chaque couple $< d_i, T^u_{d_i} >$, nous construisons un modèle de Tag pour le document $T\text{-}X_{d^u_i}$. Cette étape a pour but d'identifier les tags pertinents pour un document. Chaque tag va avoir un poids qui exprime la pertinence du tag pour le document, c'est-à-dire à quel point le tag décrit le contenu du document.

   Nous proposons trois différentes méthodes d'estimation du modèle de Tag $T\text{-}X_{d^u_i}$, tel que $X \in \{S, T, W\}$, et chaque lettre désigne une approche spécifique. Ces approches sont présentées dans la figure 4.2.
   - Un modèle de Tag Standard $T\text{-}S_{d^u_i}$ qui sera présenté dans la section suivante 4.3.1
   - Un modèle de Tag Thématique $T\text{-}T_{d^u_i}$ fera l'objet de la section 4.3.2
   - Un modèle de Tag Sémantique $T\text{-}W_{d^u_i}$ sera présenté dans la section 4.3.3
2. La seconde étape présentée à droite de la figure 4.1 de notre approche globale, est consacrée à définir le profil de l'utilisateur qui est représenté par ses tags, où nous agrégeons tous les modèles des tags pour avoir le profil de l'utilisateur.

   Ainsi, nous proposons trois modèles de profil utilisateur. Chaque modèle va s'appuyer sur un modèle de tag de document comme suit :

   (a) Un profil utilisateur Standard qui se base sur le modèle de Tag Standard

   (b) Un profil utilisateur Thématique qui se base sur le modèle de Tag Thématique

   (c) Un profil utilisateur Sémantique qui se base sur le modèle de Tag Sémantique

Nous proposons dans le tableau 4.1 toutes les notations employées dans ce chapitre.

| | |
|---|---|
| $V_T$ | L'ensemble des termes des documents |
| $V_{TG}$ | L'ensemble des tags des documents |
| $u$ | Un utilisateur |
| $t \in V_T$ | Un terme |
| $tg \in V_{TG}$ | Un tag |
| $D_u = \{d_1^u, ..., d_N^u\}$ | Ensemble de documents annotés par $u$ |
| $d_i^u$ | Un document annoté par $u$ |
| $T_{d_i}^u = \{tg_1, ..., tg_M\}$ | L'ensemble de tags assignés par $u$ au $d_i^u$ |
| $\theta_{d_i,Z} = \{\{\theta_{d_i,z_1}\}, ..., \{\theta_{d_i,z_K}\}\}$ | Distribution des thèmes dans le document $d_i^u$ |
| $\phi_Z = \{\{\phi_{z_1}\}, ..., \{\phi_{z_K}\}\}$ | Distribution des termes dans les thèmes Z. |
| $\mathcal{M}(\overrightarrow{t_o}, \overrightarrow{tg_j})$ | La matrice des plongements de termes-tags |
| $T\text{-}S_{d_i^u}$ | Modèle de Tag Standard de $d_i^u$ |
| $T\text{-}T_{d_i^u}$ | Modèle de Tag Thématique de $d_i^u$ |
| $T\text{-}W_{d_i^u}$ | Modèle de Tag Sémantique de $d_i^u$ |
| $P_{T\text{-}S}^u$ | Profil de $u$ basé sur le modèle de Tag Standard |
| $P_{T\text{-}T}^u$ | Profil de $u$ basé sur le modèle de Tag Thématique |
| $P_{T\text{-}W}^u$ | Profil de $u$ basé sur le modèle de Tag Sémantique |

TABLE 4.1 – Notations utilisées dans le chapitre 4

## 4.3 Estimation des Modèles de Tag pour les documents

Cette section est consacrée à la description des modèles qui estiment les poids des tags et leurs importance pour l'utilisateur (la première étape de la construction du profil utilisateur), et qui est représentée dans la figure 4.2.

Notre principale contribution réside dans l'estimation du modèle de Tag de chaque document $d_i^u$ annoté par l'utilisateur $u$ en exploitant le lien entre les tags et le document. Nous proposons de donner une définition au lien entre le document et les tags et qui est porté par notre hypothèse principale qui est :

*"**Seuls les tags qui décrivent les sujets des documents doivent être pris en compte**"*.

Partant de cette hypothèse globale, nous proposons trois sous hypothèses :

— **H1** : *Seuls les tags de l'utilisateur qui sont des termes du document sont des tags pertinents et décrivent le contenu du document d'après l'utilisateur.*

— **H2** : *Seuls les tags de l'utilisateur qui sont dans le même espace latent que les thèmes du document sont des tags pertinents et décrivent le contenu du document d'après l'utilisateur.*

— **H3** : *Seuls les tags de l'utilisateur qui sont dans le même espace sémantique que les termes du document sont des tags pertinents et décrivent le contenu du document d'après l'utilisateur.*

Pour chaque hypothèse nous proposons un modèle de Tags, à savoir le Modèle de Tag Standard $T\text{-}S_{d_i^u}$ présenté en section 4.3.1 répondant à l'hypothèse **H1**, le Modèle de Tag Thématique $T\text{-}T_{d_i^u}$ présenté en section 4.3.3 répondant à l'hypothèse **H2** et le Modèle de Tag Sémantique $T\text{-}W_{d_i^u}$ présenté en section 4.3.2 répondant à l'hypothèse **H3**.

### 4.3.1 Modèle de Tag Standard $T\text{-}S_{d_i^u}$ : basé sur le contenu du document

Dans le modèle de Tag Standard $T\text{-}S_{d_i^u}$ la distribution des tags repose sur l'hypothèse **H1**, que seuls les tags de l'utilisateur qui sont des termes du document sont des tags pertinents et décrivent le contenu du document d'après l'utilisateur. Ceci implique, d'une part, que tous les autres tags utilisés par l'utilisateur pour annoter un document et qui ne sont pas des termes de ce document seront considérés comme non pertinents et donc sont éliminés du modèle. D'autre part, les tags qui sont des termes des documents vont avoir une importance qui est relative à leur fréquence d'apparition dans le document auquel ils sont associés.

Donc, pour un document $d_i^u$ annoté par l'utilisateur $u$, nous calculons le poids de chaque tag $tg_j \in T_{d_i}^u$, qui est traduit par la probabilité que ce tag soit présent dans le document $P(tg_j|d_i^u)$ (la probabilité que le tag est un terme du document). Nous estimons cette probabilité par un maximum de vraisemblance (qui satisfait l'hypothèse) comme suit :

$$P(tg_j|d_i^u) = \frac{tf(tg_j, d_i^u)}{|d_i^u|} \tag{4.2}$$

où $tf(tg_i, d_i^u)$ représente la fréquence du tag $tg_j$ dans le document $d_i^u$ et $|d_i^u|$ représente la taille du document $d_i^u$.

L'estimation finale du modèle de Tag Standard $T\text{-}S_{d_i^u}$ est illustrée dans l'algorithme 1, et les notations utilisées sont détaillées dans le Tableau 4.2.

| | |
|---|---|
| $u$ | Un utilisateur |
| $d_i^u$ | Un document annoté par l'utilisateur $u$ |
| $T_{d_i}^u = \{tg_1, tg_2, ..., tg_M\}$ | L'ensemble de tags assignés par l'utilisateur $u$ au document $d_i^u$ |

TABLE 4.2 – Notations utilisées dans l'algorithme d'estimation du Modèle de Tag Standard $T\text{-}S_{d_i^u}$

**Algorithm 1** Estimation du Modèle de Tag Standard $T\text{-}S_{d_i^u}$

**Require:**
$d_i^u = \{t_1, t_2, t_3, ..., t_N\}$
$T_{d_i}^u = \{tg_1, tg_2, ..., tg_M\}$
**Ensure:**
$T\text{-}S_{d_i^u}$
1: **for each** $tg_j \in T_{d_i}^u$ **do**
2: $P(tg_j|T\text{-}S_{d_i^u}) = \frac{P(tg_j|d_i^u)}{\sum_{tg_k \in T_{d_i}^u} P(tg_k|d_i^u)}$ where $P(tg_j|d_i^u) = \frac{tf(tg_j, d_i^u)}{|d_i^u|}$
3: **end for**

L'hypothèse **H1** sur laquelle est construit le modèle de Tag Standard $T\text{-}S_{d_i^u}$ permet d'une part de ne prendre en compte que les tags qui décrivent le sujet du document et d'autre part potentiellement élimine les tags qui sont erronés (bruit) et ne décrivant pas le contenu du document. Par contre, cette hypothèse véhicule une contrainte et une limite. En effet, les tags qui peuvent être importants mais qui ne sont pas des termes du document seront éliminés par le modèle du fait que ce dernier reflète l'hypothèse qu'un utilisateur n'utilise que les termes du document comme tags pour décrire le sujet du document.

Cependant, cette hypothèse n'est pas toujours vraie, car un utilisateur peut totalement utiliser des tags qui décrivent le contenu du document sans qu'ils ne soient des termes du document. Donc, cette contrainte peut potentiellement éliminer certains tags qui sont pertinents et qui décrivent eux aussi le sujet du document.

Nous rappelons que la principale motivation de nos propositions est de ne prendre en compte que les tags qui décrivent le contenu du document. Dans ce but, nous pouvons utiliser des méthodes qui capturent la sémantique entre les tags et le contenu du document. Nous présentons dans la section suivante la méthode permettant de pallier à ce problème en employant des méthodes sémantiques.

### 4.3.2 Modèle de Tag Thématique $T\text{-}T_{d_i^u}$ : basé sur les thèmes du document

La limite du modèle de Tag Standard réside dans le fait qu'un tag important et qui couvre le sujet du document attribué par l'utilisateur pourrait être ignoré et éliminé du modèle s'il n'apparaît pas dans le document (le tag n'est pas un terme du document). Pour tenir compte de ces tags qui traitent des sujets du document mais ne sont pas des termes du document, nous proposons d'employer des modèles probabilistes thématiques.

Dans notre contexte, nous proposons d'exploiter le modèle probabiliste thématique LDA pour estimer l'importance des tags pour un document. Donc, un tag est considéré pertinent et couvre les sujets du document s'il traite des thèmes de ce dernier.

**Découvertes des thèmes des documents**

Nous employons le modèle thématique probabiliste LDA [24] pour découvrir les thèmes latents dans la collection de documents. Pour chaque document $d_i$, l'algorithme LDA fournit une distribution des thèmes $Z = \{z_1, z_2, ...z_K\}$ dénoté $\theta_{d_i,z_k}$ qui mesure la probabilité que le document $d_i$ traite du thème $z_k$. L'algorithme va aussi fournir une distribution des termes pour chaque thème $\phi_{t_o,z_k}$ mesurant la probabilité qu'un terme $t_o$ apparaît dans le thème $z_k$.

**Estimation des probabilités d'apparition des tags dans les thèmes des documents**

Chaque document $d_i^u$ annoté par l'utilisateur est associé à un ensemble de tags $T_{d_i}^u$.

Nous estimons la probabilité $P(tg_j|d_i^u)$ qui mesure la pertinence d'un tag $tg_j \in T_{d_i}^u$ pour le document $d_i^u$, qui reflète la probabilité que le tag traite des thèmes du document. Nous mesurons cette probabilité comme suit :

$$P(tg_j|d_i^u) = \sum_{z_k=1}^{K} P(tg_j|z_k)P(z_k|d_i^u) \tag{4.3}$$

$$= \sum_{z_k=1}^{K} \phi_{tg_j,z_k}\theta_{z_k,d_i^u} \tag{4.4}$$

où $P(tg_j|z_k)$ représente la probabilité que le tag $tg_j$ apparaît dans le thème $z_k$ qui est représenté par $\phi_{tg_j,z_k}$ et la probabilité $P(z_k|d_i^u)$ représente la probabilité que le document $d_i^u$ traite du thème $z_k$ représenté par $\theta_{z_k,d_i^u}$. Ces probabilités sont calculées sur l'ensemble des thèmes $Z$.

L'estimation finale de modèle de Tag Thématique $T\text{-}T_{d_i^u}$ est illustrée dans l'algorithme 2, et les notations utilisées sont détaillées dans le tableau 4.3.

| | |
|---|---|
| $u$ | Un utilisateur |
| $d_i^u$ | Le document annoté par l'utilisateur $u$ |
| $\theta_{d_i,Z} = \{\{\theta_{d_i,z_1}\},...,\{\theta_{d_i,z_K}\}\}$ | Distribution des thèmes dans le document $d_i^u$ |
| $\phi_Z = \{\{\phi_{z_1}\},...,\{\phi_{z_K}\}\}$ | Distribution des termes dans l'ensemble des thèmes Z. |
| $T_{d_i}^u = \{tg_1, tg_2, ..., tg_M\}$ | L'ensemble de tags assigné par l'utilisateur $u$ au document $d_i^u$ |

TABLE 4.3 – Notations utilisées dans l'algorithme d'estimation du modèle de Tag Thématique $T\text{-}T_{d_i^u}$

**Algorithm 2** Estimation du modèle de Tag Thématique $T\text{-}T_{d_i^u}$

**Require:**
$T_{d_i}^u, \theta_{d_i,Z}, \phi_Z$
**Ensure:**
$T\text{-}T_{d_i^u}$
1: **for each** $tg_j \in T_{d_i^u}$ **do**
2: $P(tg_j|T\text{-}T_{d_i^u}) = \frac{P(tg_i|d_i^u)}{\sum tg_k \in T_{d_i}^u P(tg_k|d_i^u)}$ avec $P(tg_j|d_i) = \sum_{z_k=1}^{K} P(tg_j|z_k)P(z_k|d_i^u)$
3: $= \sum_{z_k=1}^{K} \phi_{tg_j,z_k}\theta_{z_k,d_i^u}$
4: **end for**

Le modèle de Tag Thématique $T\text{-}T_{d_i^u}$ repose sur l'hypothèse **H2** que les tags importants sont ceux qui traitent des thèmes des documents auxquels ils sont associés. Ceci est traduit par une correspondance sémantique entre les tags et le document en passant par des modèles thématiques (LDA). Au delà des similarités document-termes [63, 64], on peut calculer des similarités entre termes (terme-terme) [71, 133-135].

Nous proposons dans la section suivante une autre approche d'estimation des tags importants pour un document en employant des modèles sémantiques qui sont plus orientés sur des calculs de similarité terme-terme. Précisément, nous mesurons par une similarité sémantique plus fine l'importance d'un tag avec chaque terme du document en employons les plongements de mots (word embeddings).

### 4.3.3 Modèle de Tag Sémantique $T\text{-}W_{d_i^u}$ : basé sur les plongement de mots

Le modèle de Tag Sémantique $T\text{-}W_{d_i^u}$ que nous proposons estime la similarité sémantique entre un tag et chaque terme du document à la différence du modèle de Tag Thématique $T\text{-}T_{d_i^u}$ qui estime la similarité entre le tag et les thèmes du document.

Plusieurs travaux se sont intéressés à définir des méthodes pour estimer la similarité sémantique entre deux termes.

Les approches les plus répandues reposent sur le principe de cooccurrence des termes dans les documents. Plus les termes co-occurrent dans les mêmes documents, plus ils sont proches sémantiquement et donc potentiellement ont la même signification. Pour estimer cette sémantique, nous employons les modèles de plongement de mots [25].

Nous nous inspirons du modèle de langue de translation (Translation Language Model) pour estimer la probabilité qu'un tag soit pertinent pour le document annoté par l'utilisateur. La pertinence d'un tag est fonction de sa similarité sémantique avec chaque terme du document et de l'importance du terme dans le document. Pour mesurer la similarité entre le tag et le terme du document, nous employons le modèle word2vec [25].

Le modèle de Tag Sémantique $T\text{-}W_{d_i^u}$ pour le document $d_i^u$ annoté par l'utilisateur $u$ est estimé sur l'ensemble de tags $T_{d_i}^u$ associés au document, où la probabilité $P(tg_j|d_i^u)$ est estimée comme suit :

$$P(tg_j|d_i^u) = \frac{\sum_{t_o \in d_i} P(tg_j|t_o)P(t_o|d_i)}{\sum_{tg_k \in T_{d_i}^u} P(tg_k|d_i^u)} \tag{4.5}$$

La probabilité $P(tg_j|t_o)$ présente la similarité sémantique entre le tag $tg_j$ et le terme $t_o$ que nous estimons par la fonction Sigmoid [67] comme suit :

$$P(tg_j|t_o) = \delta(\overrightarrow{tg_j}, \overrightarrow{t_o}) = \frac{1}{1 + exp(-a(cos(\overrightarrow{tg_j}, \overrightarrow{t_o}) - c))} \tag{4.6}$$

où $cos(\overrightarrow{tg_j}, \overrightarrow{t_o})$ représente le cosinus entre le tag $tg_j$ et le terme $t_o$ estimé par le modèle word2vec, $a$ et $c$ sont des paramètres de la fonction.

La probabilité $P(t_o|d_i^u)$ présente l'importance du terme dans le document que nous estimons par un maximum de vraisemblance comme suit :

$$P(t_o|d_i^u) = \frac{tf(t_o, d_i^u)}{|d_i^u|} \tag{4.7}$$

avec $tf(t_o, d_i^u)$ la fréquence du terme $t_o$ dans le document $d_i^u$ et $|d_i^u|$ la taille du document $d_i^u$.

L'estimation finale du modèle de Tag Sémantique $T\text{-}W_{d_i^u}$ est présenté dans l'algorithme 3, et les notations utilisées sont détaillées dans le Tableau 4.4.

| | |
|---|---|
| $u$ | Un utilisateur |
| $d_i^u$ | Le document annoté par l'utilisateur $u$ |
| $T_{d_i}^u = \{tg_1, tg_2, ..., tg_M\}$ | L'ensemble de tags assigné par l'utilisateur $u$ au document $d_i^u$ |
| M | La matrice de prolongement de termes-tags |

TABLE 4.4 – Notations utilisées dans l'algorithme d'estimation du modèle de Tag Sémantique $T\text{-}W_{d_i^u}$

**Algorithm 3** Estimation du modèle de Tag Sémantique $T\text{-}W_{d_i^u}$

**Require:**
$d_i = \{t_1, t_2, t_3, ..., t_N\}$
$T_{d_i}^u = \{tg_1, tg_2, ..., tg_M\}$
$\mathcal{M}(\overrightarrow{t_o}, \overrightarrow{tg_j})$
**Ensure:**
$\mathcal{W}_{T_{d_i}^u}$
1: **for each** $tg_i \in T_{d_i}^u$ **do**
2: $P(tg_j|T\text{-}W_{d_i^u}) = \frac{P(tg_j|d_i)}{\sum_{tg_k \in T_{d_i}^u} P(tg_k|T\text{-}W_{d_i^u})} = \frac{\sum t_o \in d_i P(tg_j|t_o)P(t_o|d_i)}{\sum_{tg_k \in T_{d_i}^u} P(tg_k|T\text{-}W_{d_i^u})}$
3: where $P(tg_j|t_o) = \delta(\overrightarrow{tg_j}, \overrightarrow{t_o}) = \frac{1}{1+exp(-a(sim(\overrightarrow{tg_j}, \overrightarrow{t_o})-c))}$
4: and $P(t_o|d_i) = \frac{tf(t_o,d_i)}{|d_i|}$
5: **end for**

## 4.4 Construction du profil de l'utilisateur

La construction du profil de l'utilisateur que nous proposons est réalisée en deux étapes comme suit :

**Identification des tags importants** consiste à déterminer les tags pertinents qui reflètent les centres d'intérêts de l'utilisateur en exploitant ses documents annotés avec les tags associés à chaque document.
Donc, pour chaque couple $< d_i^u, T_{d_i}^u >$ (document $d_i^u$ annoté par l'utilisateur $u$ avec les tags $T_{d_i}^u$) , nous estimons le poids de chaque tag de l'ensemble $T_{d_i}^u$. Ce poids reflète l'importance (la pertinence que le tag traite les sujets du document) du tag pour le document. Nous construisons un Modèle de Tag pour chaque document $d_i^u$ qui propose une distribution des tags associés au document.
Dans la section précédente, nous avons détaillé les trois méthodes d'estimation des modèles de tags où chaque modèle propose une pondération différente des tags : le modèle de Tag Standard $T\text{-}W_{d_i^u}$, le modèle de Tag Thématique $T\text{-}W_{d_i^u}$ et le modèle de Tag Sémantique $T\text{-}W_{d_i^u}$.

**Construction du profil utilisateur** réalisée en exploitant les modèles de tags des documents estimés dans la première étape. Pour chaque tag $tg_j$ employé par l'utilisateur, le poids final de ce tag correspond à sa distribution moyenne dans tous les modèles des tags des documents auxquels il a été assigné par l'utilisateur.

Plus formellement, pour l'utilisateur $u$ ayant annoté l'ensemble de documents $D_u$, où chaque document $d_i^u \in D_u$ est associé à l'ensemble de tag $T_{d_i}^u$ attribué par l'utilisateur $u$, nous définissons son profil sur l'ensemble de son vocabulaire de tags $V_{TG}^u$ comme suit :

$$P(tg_j|P_{T\text{-}X}^u) = \frac{\sum_{d_i \in D_u} P(tg_j|T\text{-}X_{d_i^u})}{\sum_{tg_p \in V_{TG}^u} \sum_{d_i \in D_u} P(tg_p|T\text{-}X_{d_i^u})} \tag{4.8}$$

$P_{T\text{-}X}^u$ représente le profil de l'utilisateur où $X$ peut être S pour désigner le profil standard $P_{T\text{-}S}^u$, T pour le profil thématique $P_{T\text{-}T}^u$, et W pour le profil sémantique $P_{T\text{-}W}^u$.

Nous définissons ainsi trois profils pour l'utilisateur qui exploitent chaque modèle de Tags de document comme suit :

— **Le Profil Utilisateur Standard** notée $P_{T\text{-}S}^u$ où l'estimation du poids de chaque tag est basée sur le modèle de Tag Standard décrit en section 4.3.1, comme suit :

$$P(tg_j|P_{T\text{-}S}^u) = \frac{\sum_{d_i \in D_u} P(tg_j|T\text{-}S_{d_i^u})}{\sum_{tg_p \in V_{TG}^u} \sum_{d_i \in D_u} P(tg_p|T\text{-}S_{d_i^u})} \tag{4.9}$$

— **Le Profil Utilisateur Thématique** $P_{T\text{-}T}^u$ avec une estimation basée sur le modèle de Tag Thématique présenté en section 4.3.2 pour chaque tag de l'utilisateur :

$$P(tg_j|P_{T\text{-}T}^u) = \frac{\sum_{d_i \in D_u} P(tg_j|T\text{-}T_{d_i^u})}{\sum_{tg_p \in V_{TG}^u} \sum_{d_i \in D_u} P(tg_p|T\text{-}T_{d_i^u})} \tag{4.10}$$

— **Le Profil Utilisateur Sémantique** $P_{T\text{-}W}^u$ basé sur le modèle de Tag Sémantique 4.3.3

$$P(tg_j|P_{T\text{-}W}^u) = \frac{\sum_{d_i \in D_u} P(tg_j|T\text{-}W_{d_i^u})}{\sum_{tg_p \in V_{TG}^u} \sum_{d_i \in D_u} P(tg_p|T\text{-}W_{d_i^u})} \tag{4.11}$$

## 4.5 Modèle de recherche d'information sociale personnalisée

Nous plaçons nos propositions dans un contexte de recherche d'information personnalisée dans les folksonomies.

L'objectif principal de nos contributions est l'évaluation de l'apport des profils des utilisateurs lors du traitement de requêtes. De ce fait, nous employons un modèle d'ordonnancement qui est basé sur une combinaison linéaire du score de pertinence du document à la requête et un score de pertinence du document pour le profil de l'utilisateur. Ce modèle est largement employé dans les travaux de l'état de l'art [11, 13, 14, 16, 21, 28].

Pour un utilisateur $u$ qui soumet la requête $q$, le document $d$ va avoir un score d'ordonnancement estimé comme suit :

$$RSV(q,d,u) = \beta_{TG} RSV(q,d) + (1-\beta_{TG}) RSV(d,u) \tag{4.12}$$

avec $RSV(q,d)$ représente le score correspondance entre le document $d_i^u$ et la requête $q$ et $RSV(u,d)$ représente le score de correspondance entre le profil utilisateur et le document respectivement, et $\beta_{TG}$ un paramètre dans [0,1].

Le profil de l'utilisateur $u$ est estimé en utilisant les modèles que nous avons proposés dans les sections précédentes. Nous proposons ainsi 3 modèles de RISP qui sont :

— Le modèle $\mathbf{MT^u_S}$ où le profil de l'utilisateur est estimé avec le modèle $P^u_{T\text{-}S}$ sous l'hypothèse que les tags qui sont présents dans les documents annotés par l'utilisateur décrivent mieux les thématiques de l'utilisateur.

$$RSV(q,d,u) = \beta_{TG} RSV(q,d) + (1-\beta_{TG}) RSV(d,P^u_{T\text{-}S}) \tag{4.13}$$

— Le modèle $\mathbf{MT^u_T}$ où le profil de l'utilisateur est définit en employant le modèle $P^u_{T\text{-}T}$ qui suppose que les tags qui sont dans le même espace sémantique des documents estimé par le modèle LDA sont les plus pertinent pour modéliser l'utilisateur.

$$RSV(q,d,u) = \beta_{TG} RSV(q,d) + (1-\beta_{TG}) RSV(d,P^u_{T\text{-}T}) \tag{4.14}$$

— Le modèle $\mathbf{MT^u_W}$ où le profil de l'utilisateur est modélisé par le modèle $P^u_{T\text{-}W}$ avec l'hypothèse que les tags qui sont dans le même espace sémantique définit par les plongements des mots sont les plus à même pour caractériser l'utilisateur.

$$RSV(q,d,u) = \beta_{TG} RSV(q,d) + (1-\beta_{TG}) RSV(d,P^u_{T\text{-}W}) \tag{4.15}$$

Les scores $RSV(q,d)$ et $RSV(u,d)$ peuvent être calculés en utilisant un modèle classique de RI (modèle vectoriel, modèle probabiliste BM25, modèle de langue Dirichlet, Hiemstra, etc). Nous proposons d'employer le modèle de langue avec lissage de Dirichlet. Le paramètre de lissage $\mu$ est fixé, dans toutes nos expérimentations, à la valeur classique 2500, qui est celle par défaut de systèmes comme Terrier.

# Chapitre 5

# Modèle de Document - Exploitation des tags pour la pondération des termes du document

## 5.1 Introduction

Dans ce chapitre, nous présentons notre deuxième contribution de modèle utilisateur. Ce modèle consiste à représenter les centres d'intérêts de l'utilisateur qui sont extraits des documents annotés. L'hypothèse globale du modèle est que les termes importants d'un document sont ceux qui sont en relation avec les tags attribués par cet utilisateur à ce même document.

Notre proposition permet d'identifier dans quelle mesure les tags des documents peuvent être exploités pour identifier les termes les plus importants et les plus représentatifs des centres d'intérêts de l'utilisateur.

Le chapitre est organisé comme suit : premièrement dans la section 5.2 nous décrivons l'architecture globale de notre approche de modélisation de l'utilisateur et les notations utilisées dans ce chapitre. En section 5.3, nous détaillons les différents modèles de contenu de documents. En section 5.4 nous présentons la construction du profil de l'utilisateur. La construction du profil de l'utilisateur est décrite en section 4.4. Enfin, en section 5.5, nous présentons comment nous intégrons le profil utilisateur dans le système de RISP.

## 5.2 Approche de modélisation de l'utilisateur basé sur les termes du document

Soit un utilisateur $u$ ayant annoté un ensemble de document $D_u = \{d_1^u, d_2^u, ..., d_N^u\}$. L'utilisateur a attribué un ensemble de tags $T_{d_i}^u = \{tg_1, tg_2, ..., tg_M\}$ pour chaque document $d_i^u \in D_u$.

L'objectif est de construire un profil pour l'utilisateur basé sur les termes des documents et d'attribuer pour chaque terme $t_o$ de l'ensemble du vocabulaire de la collection $V_T$ un poids $w_{t_o}$. Donc le profil de l'utilisateur $P_C^u$ basé sur le contenu va être représenté comme suit :

$$P_C^u = \{< t_1, w_{t_1} >, < t_2, w_{t_2} >, ..., < t_S, w_{t_S} >\} \tag{5.1}$$

L'architecture globale de notre approche qui est présentée dans la figure 5.1.

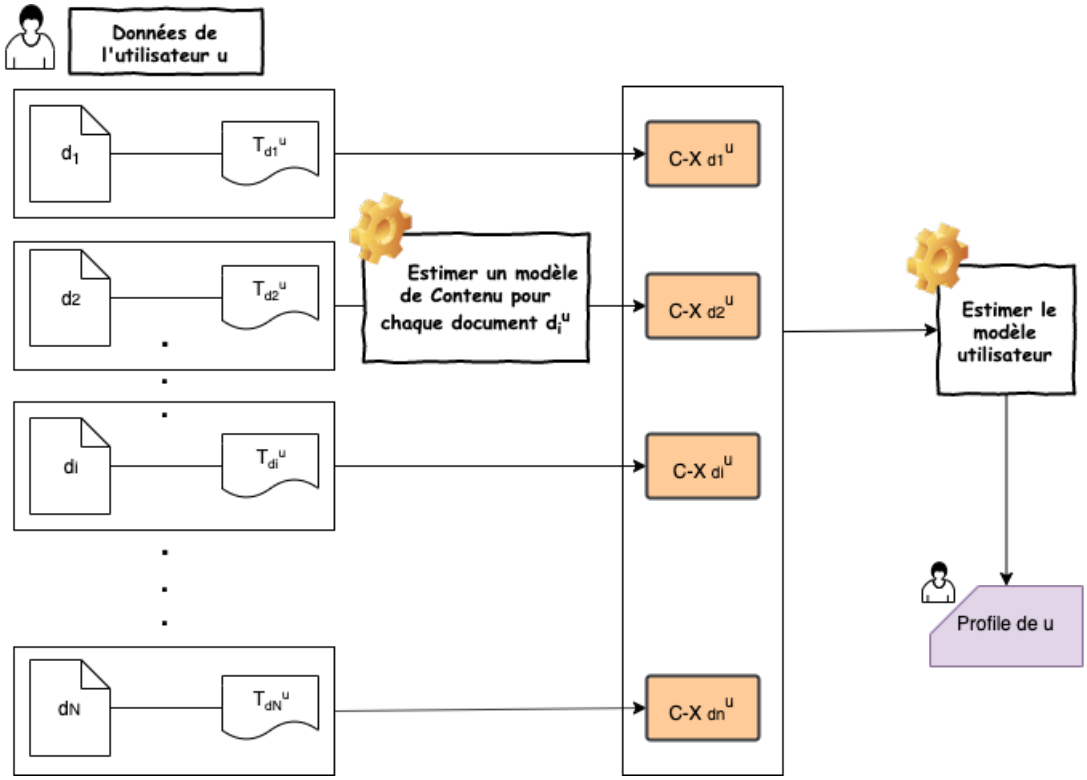



FIGURE 5.1 – Approche globale de construction du profil de l'utilisateur basé sur les documents

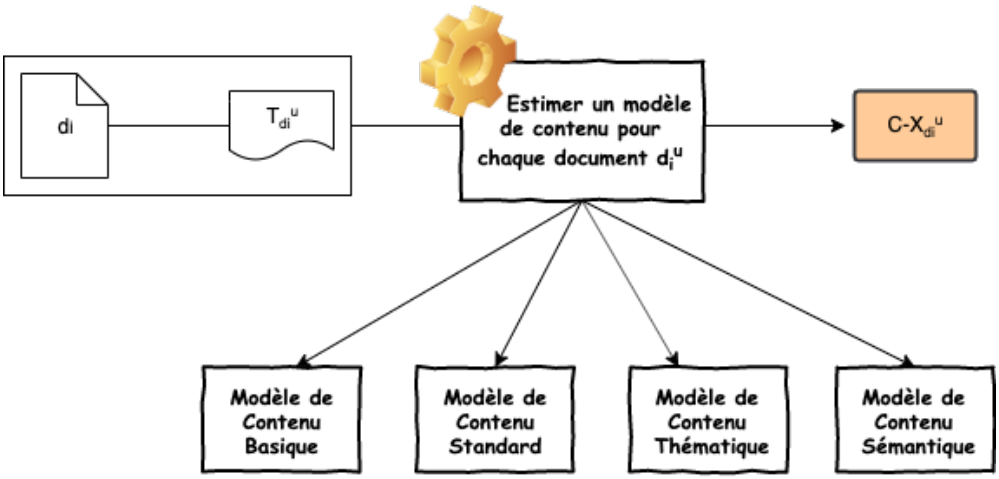



FIGURE 5.2 – Les différents modèles de Contenu d'un document

**Information de l'utilisateur**

Dans notre approche, les informations que nous exploitons pour modéliser l'utilisateur sont les documents que l'utilisateur lui-même a annoté et les tags qu'il a attribués à chaque document.

Plus formellement, un utilisateur $u$ est associé à un ensemble de document $D_u = \{d_1^u, d_2^u, ..., d_N^u\}$ qu'il a annoté, où chaque document $d_i^u \in D^u$ est associé

à un ensemble de tag $T^u_{d_i} = \{tg_1, tg_2, ..., tg_M\}$ attribué par l'utilisateur $u$.

**Représentation du profil de l'utilisateur**

Pour représenter le profil de l'utilisateur, nous employons la représentation vectorielle.

La modélisation que nous proposons repose sur un profil de l'utilisateur basé sur les termes des documents qu'il a annoté. En conséquence le profil de l'utilisateur est représenté par un vecteur de terme, où chaque terme $t_o$ va avoir un poids $w_{t_o}$ qui estime l'importance du terme pour l'utilisateur comme présenté par la formule 5.1.

**Construction du profil**

La construction du profil de l'utilisateur est réalisée en deux étapes.

1. La première étape consiste à estimer les modèles de contenu pour les documents annotés par l'utilisateur. Cette étape a pour but d'identifier les termes pertinents pour un document.

   Donc, pour chaque document $d^u_i$, nous construisons un modèle de contenu où chaque terme va avoir un poids d'importance qui exprime la pertinence du terme pour le document, c'est-à-dire à quel point le terme décrit le contenu du document.

   Nous proposons quatre différentes méthodes d'estimation de modèle de contenu pour chaque document $d^u_i$ annoté par l'utilisateur $u$ comme présenté dans la figure 5.2 :
   — Un modèle de Contenu Basique qui sera présenté dans la section 5.3.1
   — Un modèle de Contenu Standard qui sera présenté dans la section 5.3.2
   — Un modèle de Contenu Thématique qui sera présenté dans la section 5.3.3
   — Un modèle de Contenu Sémantique qui sera présenté dans la section 5.3.4

2. La seconde étape, quant à elle, est consacrée à définir le profil de l'utilisateur qui est représenté par les termes des documents qu'il a annoté, où nous agrégeons tous les modèles des contenu des documents pour avoir le profil de l'utilisateur.

   Ainsi, nous proposons quatre différents modèles de profil pour l'utilisateur. Chaque modèle va s'appuyer sur un modèle de contenu du document comme suit :

   (a) Un profil utilisateur Basique qui se base sur le modèle de Contenu Basique

   (b) Un profil utilisateur Standard qui se base sur le modèle de Contenu Standard

   (c) Un profil utilisateur Thématique qui se base sur le modèle de Contenu Thématique

   (d) Un profil utilisateur Sémantique qui se base sur le modèle de Contenu Sémantique

Nous proposons dans le tableau 5.1 toutes les notations employées dans ce chapitre.

| | |
|---|---|
| $V_T$ | L'ensemble des termes des documents |
| $V_{TG}$ | L'ensemble des tags des documents |
| $u$ | un utilisateur |
| $t \in V_T$ | un terme |
| $tg \in V_{TG}$ | un tag |
| $V^u_{TG}$ | L'ensemble des tags de l'utilisateur $u$ |
| $D_u$ | Ensemble de documents annotés par l'utilisateur $u$ |
| $d^u_i$ | Un document annoté par l'utilisateur $u$ |
| $T^u_{d_i}$ | L'ensemble de tags assignés par l'utilisateur $u$ au document $d^u_i$ |
| $T\text{-}B_{d^u_i}$ | Modèle de Tag Basique $d^u_i$ |
| $T\text{-}S_{d^u_i}$ | Modèle de Tag Standard $d^u_i$ |
| $T\text{-}T_{d^u_i}$ | Modèle de Tag Thématique $d^u_i$ |
| $T\text{-}W_{d^u_i}$ | Modèle de Tag Sémantique $d^u_i$ |
| $C\text{-}B_{d^u_i}$ | Modèle de Contenu Basique $d^u_i$ |
| $C\text{-}S_{d^u_i}$ | Modèle de Contenu Standard $d^u_i$ |
| $C\text{-}T_{d^u_i}$ | Modèle de Contenu Thématique d $d^u_i$ |
| $C\text{-}W_{d^u_i}$ | Modèle de Contenu Sémantique $d^u_i$ |
| $P^u_{C\text{-}B}$ | Profil de $u$ basé sur le modèle de Contenu Basique |
| $P^u_{C\text{-}S}$ | Profil de $u$ basé sur le modèle de Contenu Standard |
| $P^u_{C\text{-}T}$ | Profil de $u$ basé sur le modèle de Contenu Thématique |
| $P^u_{C\text{-}W}$ | Profil de $u$ basé sur le modèle de Contenu Sémantique |

TABLE 5.1 – Notations utilisées dans ce chapitre 5

## 5.3 Estimation du modèle de Contenu pour les documents

Dans cette section, nous détaillons l'estimation du modèle. La particularité de ce modèle est de faire dépendre les termes du document non seulement du contenu textuel du document mais également des tags attribués par l'utilisateur à ce document. Le but est de déterminer les termes importants du document qui reflètent les centres d'intérêts de l'utilisateur.

Comme décrit précédemment, un terme du document est pondéré en combinant deux sources d'évidence :

— La première étant le document lui même. L'importance du terme pour le document, qui est représenté par la probabilité du terme dans le document $P(t_o|d^u_i)$.

— La seconde est la proximité du terme aux tags du document attribués par l'utilisateur. Cette proximité est estimée par la probabilité du terme dans le modèle de tag du document $P(t_o|T_{d^u_i})$.

Ainsi, nous proposons d'estime l'importance d'un terme $t_o$ du document dans le modèle de contenu $C\text{-}X_{d^u_i}$ en combinant deux scores, un premier score $P(t_o|d^u_i)$ qui reflète l'importance du terme $t_o$ dans le document $d^u_i$ et un second score qui reflète l'importance du terme $t_o$ dans l'ensemble des tag associé $T_{d^u_i}$ au document $d^u_i$. La formule finale est donc écrite comme suit :

$$P(t_o|C\text{-}X_{d^u_i}) = \frac{\alpha P(t_o|d^u_i) + (1-\alpha)P(t_o|T_{d^u_i})}{\sum_{t_k \in d^u_i} P(t_k|C\text{-}X_{d^u_i})} \tag{5.2}$$

**Importance du terme dans le document** L'importance d'un terme $t_o$ dans le document est décrite par la probabilité $P(t_o|d_i^u)$ que nous estimons par un maximum de vraisemblance comme suit :

$$P(t_o|d_i^u) = \frac{tf(t_o, d_i^u)}{|d_i^u|} \tag{5.3}$$

où $tf(t_o, d_i^u)$ représente la fréquence du terme $t_o$ dans le document $d_i^u$, $|d_i^u|$ représente la taille du document $d_i^u$.

**Importance du terme dans le modèle de Tag du Document** intègre à la fois la proximité du terme par rapport aux tags associés au document $d_i^u$ attribués par l'utilisateur $u$ et l'importance des tags dans le modèle des tags du document. Cette importance $P(t_o|T_{d_i^u})$ est exprimée comme suit :

$$P(t_o|T_{d_i}^u) = \frac{\sum_{tg_j \in T_{d_i^u}} P(t_o|tg_j) P(tg_j|T\text{-}X_{d_i^u})}{\sum_{t_k \in {d_i}^u} P(t_k|T_{d_i}^u)} \tag{5.4}$$

— La probabilité $P(t_o|tg_j)$ est estimée par une proximité sémantique entre le terme du document $t_o$ et le tag $tg_j$. Nous employons les plongements de mots pour mesurer cette proximité sémantique. Nous utilisons la fonction Sigmoid [68] pour calculer la similarité comme suit :

$$P(t_o|tg_j) = \delta(\overrightarrow{tg_j}, \overrightarrow{t_o}) = \frac{1}{1 + exp(-a(cos(\overrightarrow{tg_j}, \overrightarrow{t_o}) - c))} \tag{5.5}$$

où $cos(\overrightarrow{tg_j}, \overrightarrow{t_o})$ représente le cosinus entre le tag $tg_j$ et le terme $t_o$ estimé par le modèle word2vec, $a$ et $c$ sont des paramètres de la fonction.

— La probabilité $P(tg_j|T\text{-}X_{d_i^u})$ représente l'importance du tag $tg_j$ suivant un modèle de Tag $T\text{-}X_{d_i^u}$ du document.

Globalement, plus le terme d'un document est proche du tag de l'utilisateur $P(t_o|tg_j)$ et plus ce tag est important $P(tg_j|T\text{-}X_{d_i^u})$ alors le terme va avoir poids plus important. Ainsi, la $P(tg_j|T\text{-}X_{d_i^u})$ dépend du modèle de Tag $T\text{-}X_{d_i^u}$ où $X$ représente le modèle de Tag que nous avons présenté dans le chapitre précédent.

Donc, la formule globale d'estimation des termes importants d'un document est comme suit :

$$P(t_o|C\text{-}X_{d_i^u}) = \frac{\alpha P(t_o|d_i^u) + (1-\alpha) P(t_o|T_{d_i^u})}{\sum_{t_k \in d_i^u} P(t_k|C\text{-}X_{d_i^u})} \tag{5.6}$$

$$= \frac{\alpha \frac{tf(t_o, d_i^u)}{|d_i^u|} + (1-\alpha) \frac{\sum_{tg_j \in T\text{-}X_{d_i^u}} P(t_o|tg_j) \boxed{P(tg_j|T\text{-}X_{d_i^u})}}{\sum_{t_k \in d_i^u} P(t_k|T_{d_i}^u)}}{\sum_{t_k \in d_i^u} P(t_k|C\text{-}X_{d_i^u})} \tag{5.7}$$

L'estimation de la distribution des termes du document dans le modèle dépend du score d'importance du terme dans le document et de son importance pour le modèle de Tag du document.

Nous avons proposé dans le chapitre précédent des modèles de tags pour le document qui fournissent une distribution de tags de l'utilisateur pour chaque document reflétant l'importance du tag pour ce dernier.

Nous proposons des instantiation du modèle pour tenir en compte des modèles de tags, où la différence entre chaque instantiation est l'estimation du modèle de Tag d'un document représenté par $\boxed{P(tg_j|T\text{-}X_{d_i^u})}$ (encadré dans chaque formule).

### 5.3.1 Modèle de Contenu Basique $C\text{-}B_{d_i^u}$

Le modèle de contenu de document intègre le modèle de Tag dans l'estimation des poids des termes du document. La première instanciation que nous proposons est le modèle de Contenu Basique $P(t_o|C\text{-}B_{d_i^u})$ comme suit :

$$P(t_o|C\text{-}B_{d_i^u}) = \frac{\alpha P(t_o|d_i^u) + (1-\alpha)P(t_o|T\text{-}B_{d_i^u})}{\sum_{t_k \in d_i^u} P(t_k|C\text{-}B_{d_i^u})} \tag{5.8}$$

$$= \frac{\alpha \frac{tf(t_o,d_i^u)}{|d_i^u|} + (1-\alpha)\frac{\sum_{tg_j \in T_{d_i}^u} P(t_o|tg_j)\boxed{P(tg_j|T\text{-}B_{d_i^u})}}{\sum_{t_k \in d_i^u} P(t_k|T_{d_i}^u)}}{\sum_{t_k \in d_i^u} P(t_k|C\text{-}B_{d_i^u})} \tag{5.9}$$

$$= \frac{\alpha \frac{tf(t_o,d_i^u)}{|d_i^u|} + (1-\alpha)\frac{\sum_{tg_j \in T_{d_i}^u} P(t_o|tg_j)\boxed{\frac{tf(tg_j,T_{d_i}^u)}{|T_{d_i}^u|}}}{\sum_{t_k \in d_i^u} P(t_k|T_{d_i}^u)}}{\sum_{t_k \in d_i^u} P(t_k|C\text{-}B_{d_i^u})} \tag{5.10}$$

où la probabilité $P(tg_j|T\text{-}B_{d_i^u})$ est calculée par la fréquence d'apparition du tag $tg_j$ dans l'ensemble des tags $T_{d_i}^u$, qui représente le nombre de fois que l'utilisateur a utilisé le tag $tg_j$ pour annoter le document $d_i^u$. La fréquence d'un un tag pour un même document est égale à 1 car généralement l'utilisateur emploie une fois le même tag pour annoter un document.

### 5.3.2 Modèle de Contenu Standard $C\text{-}S_{d_i^u}$

Un modèle de Contenu Standard noté $C\text{-}S_{d_i^u}$ où la probabilité $P(tg_j|T\text{-}S_{d_i^u})$ est estimé suivant le modèle de Tag Standard présenté dans le chapitre 4, et la formule finale du modèle $C\text{-}S_{d_i^u}$ est comme suit :

$$P(t_o|C\text{-}S_{d_i^u}) = \frac{\alpha P(t_o|d_i^u) + (1-\alpha)P(t_o|T\text{-}S_{d_i^u})}{\sum_{t_k \in d_i^u} P(t_k|C\text{-}S_{d_i^u})} \tag{5.11}$$

$$= \frac{\alpha \frac{tf(t_o,d_i^u)}{|d_i^u|} + (1-\alpha)\frac{\sum_{tg_j \in T\text{-}S_{d_i^u}} P(t_o|tg_j)\boxed{P(tg_j|T\text{-}S_{d_i^u})}}{\sum_{t_k \in d_i^u} P(t_k|T_{d_i}^u)}}{\sum_{t_k \in d_i^u} P(t_k|C\text{-}S_{d_i^u})} \tag{5.12}$$

$$= \frac{\alpha \frac{tf(t_o,d_i^u)}{|d_i^u|} + (1-\alpha)\frac{\sum_{tg_j \in T\text{-}S_{d_i^u}} P(t_o|tg_j)\boxed{\frac{\frac{tf(tg_j,d_i^u)}{|d_i|}}{\sum_{tg_r \in T_{d_i}^u}\frac{tf(tg_r,d_i^u)}{|d_i|}}}}{\sum_{t_k \in d_i^u} P(t_k|T_{d_i}^u)}}{\sum_{t_k \in d_i^u} P(t_k|C\text{-}S_{d_i^u})} \tag{5.13}$$

où la $tf(tg_j|d_i^u)$ présente la fréquence du tag $tg_j$ dans le document $d - i^u$ et $|d_i^u|$ représente la taille du document $d_i$.

### 5.3.3 Modèle de Contenu Thématique $C\text{-}T_{d_i^u}$

Un modèle de Contenu Thématique noté $C\text{-}T_{d_i^u}$ où la probabilité $P(tg_j|T\text{-}T_{d_i^u})$ est estimé suivant le modèle de Tag Thématique présenté dans la chapitre 4, et la formule finale du modèle $C\text{-}T_{d_i^u}$ est comme suit :

$$P(t_o|C\text{-}T_{d_i^u}) = \frac{\alpha P(t_o|d_i^u) + (1-\alpha)P(t_o|T\text{-}T_{d_i^u})}{\sum_{t_k \in d_i^u} P(t_k|C\text{-}T_{d_i^u})} \tag{5.14}$$

$$= \frac{\alpha \frac{tf(t_o,d_i^u)}{|d_i^u|} + (1-\alpha)\frac{\sum_{tg_j \in T\text{-}T_{d_i^u}} P(t_o|tg_j)\boxed{P(tg_j|T\text{-}T_{d_i^u})}}{\sum_{t_k \in d_i^u} P(t_k|T_{d_i}^u)}}{\sum_{t_k \in d_i^u} P(t_k|C\text{-}T_{d_i^u})} \tag{5.15}$$

$$= \frac{\alpha \frac{tf(t_o,d_i^u)}{|d_i^u|} + (1-\alpha)\frac{\sum_{tg_j \in T\text{-}T_{d_i^u}} P(t_o|tg_j)\boxed{\frac{\sum_{z_k=1}^{K} \phi_{tg_j,z_k}\theta_{z_k,d_i^u}}{\sum_{tg_r \in T_{d_i}^u}\sum_{z_k=1}^{K} \phi_{tg_r,z_k}\theta_{z_k,d_i^u}}}}{\sum_{t_k \in d_i^u} P(t_k|T_{d_i}^u)}}{\sum_{t_k \in d_i^u} P(t_k|C\text{-}T_{d_i^u})} \tag{5.16}$$

où $\phi_{tg_j,z_k}$ représente la probabilité du tag $tg_j$ dans le thème $z_k$ et $\theta_{z_k,d_i^u}$ représente la probabilité que le document traite $d_i^u$ du thème $z_k$. Ces probabilités sont calculées sur l'ensemble des thèmes.

### 5.3.4 Modèle de Contenu Sémantique $C\text{-}W_{d_i^u}$

Un modèle de Contenu Sémantique noté $C\text{-}W_{d_i^u}$ où la probabilité $P(tg_j|T\text{-}W_{d_i^u})$ est estimé suivant le modèle de Tag Thématique présenté dans la chapitre 4, et la formule finale du modèle $C\text{-}W_{d_i^u}$ est comme suit :

$$P(t_o|C\text{-}W_{d_i^u}) = \frac{\alpha P(t_o|d_i^u) + (1-\alpha)P(t_o|T\text{-}W_{d_i^u})}{\sum_{t_k \in d_i^u} P(t_k|C\text{-}W_{d_i^u})} \tag{5.17}$$

$$= \frac{\alpha \frac{tf(t_o,d_i^u)}{|d_i^u|} + (1-\alpha)\frac{\sum_{tg_j \in T\text{-}W_{d_i^u}} P(t_o|tg_j)\boxed{P(tg_j|T\text{-}W_{d_i^u})}}{\sum_{t_k \in d_i^u} P(t_k|T_{d_i}^u)}}{\sum_{t_k \in d_i^u} P(t_k|C\text{-}W_{d_i^u})} \tag{5.18}$$

$$= \frac{\alpha \frac{tf(t_o,d_i^u)}{|d_i^u|} + (1-\alpha)\frac{\sum_{tg_j \in T\text{-}W_{d_i^u}} P(t_o|tg_j)\boxed{\frac{\sum_{t_p \in d_i^u} P(tg_j|t_p)P(t_p|d_i^u)}{\sum_{tg_r \in T_{d_i^u}}\sum_{t_p \in d_i^u} P(tg_r|t_p)P(t_p|d_i^u))}}}{\sum_{t_k \in d_i^u} P(t_k|T_{d_i}^u)}}{\sum_{t_k \in d_i^u} P(t_k|C\text{-}W_{d_i^u})} \tag{5.19}$$

où la probabilité $P(tg_j|t_p)$ présente la similarité sémantique entre le tag $tg_j$ et le terme $t_p$ et $P(t_p|d_i^u)$ représente la probabilité du terme $t_p$ dans le document $d_i^u$. Cette probabilité est calculée par un maximum de vraisemblance $(tf(t_p, d_i^u)||d_i^u|)$.

## 5.4 Construction du profil de l'utilisateur

La construction du profil de l'utilisateur que nous proposons est réalisée en deux étapes comme suit :

1. **Identification des termes importants** consiste à déterminer les termes pertinents pour chaque document annoté par l'utilisateur et qui potentiellement reflètent les centres d'intérêts de cet utilisateur.
   Pour chaque couple $< d_i^u, T_{d_i}^u >$ ( document $d_i^u$ annoté par l'utilisateur $u$ avec les tags $T_{d_i}^u$), nous estimons le poids de chaque terme du document $d_i^u$.
   Dans la section précédente, nous avons détaillé les quatre méthodes d'estimation des modèles de contenu pour chaque document de l'utilisateur :
   (a) Le modèle de Contenu Basique $C\text{-}X_{d_i^u}$
   (b) Le modèle de Contenu Standard $C\text{-}S_{d_i^u}$
   (c) Le modèle de Contenu Thématique $C\text{-}T_{d_i^u}$
   (d) Le modèle de Contenu Sémantique $C\text{-}W_{d_i^u}$
2. **Construction du profil utilisateur** est réalisée en exploitant les modèles de Contenu des documents estimés dans la première étape.
   Sur l'ensemble des documents annotés par l'utilisateur $d_i^u \in D_u$, nous estimons les poids des termes sur l'ensemble des modèles de contenu de l'utilisateur.
   Pour chaque termes $t_o$ du vocabulaire, le poids final de ce terme correspond à sa distribution moyenne dans tous les modèles des Contenu des documents.

Plus formellement, un utilisateur $u$ est représenté par un vocabulaire de termes $V_T$ définit sur l'ensemble de termes des documents $D^u$ qu'il a annoté. Nous définissons un profil utilisateur sur l'ensemble de son vocabulaire $V_T$, où le poids de chaque terme $t_o \in V_T$ est estimé comme suit :

$$P(t_o|P_{C\text{-}X}^u) = \frac{\sum_{d_i \in D_u} P(t_o|C\text{-}X_{d_i^u})}{\sum_{t_p \in V_T} \sum_{d_i \in D_u} P(t_p|C\text{-}X_{d_i^u})} \tag{5.20}$$

$P_{C\text{-}X}^u$ représente le profil de l'utilisateur où $X$ qui peut être B pour le profil basique $P_{C\text{-}B}^u$, S pour désigner le profil standard $P_{C\text{-}S}^u$, T pour le profil thématique $P_{C\text{-}T}^u$, et W pour le profil sémantique $P_{C\text{-}W}^u$.

Nous définissons ainsi quatre profils pour l'utilisateur qui exploitent chaque modèle de contenu de document comme suit :

— **Le Profil Utilisateur Basique** notée $P_{C\text{-}B}^u$ basé sur le modèle de Contenu Basique décrit en section 5.3.1, comme suit :

$$P(t_o|P_{C\text{-}B}^u) = \frac{\sum_{d_i \in D_u} P(t_o|C\text{-}B_{d_i^u})}{\sum_{t_p \in V_T} \sum_{d_i \in D_u} P(t_p|C\text{-}B_{d_i^u})} \tag{5.21}$$

— **Le Profil Utilisateur Standard** notée $P_{C\text{-}S}^u$ basé sur le modèle de Contenu Standard décrit en section 5.3.2, comme suit :

$$P(t_o|P_{C\text{-}S}^u) = \frac{\sum_{d_i \in D_u} P(t_o|C\text{-}S_{d_i^u})}{\sum_{t_p \in V_T} \sum_{d_i \in D_u} P(t_p|C\text{-}S_{d_i^u})} \tag{5.22}$$

— **Le Profil Utilisateur Thématique** $P^u_{C\text{-}T}$ avec une estimation basée sur le modèle de Contenu Thématique présenté en section 5.3.3 pour chaque tag de l'utilisateur :

$$P(t_o|P^u_{C\text{-}T}) = \frac{\sum_{d_i \in D_u} P(t_o|C\text{-}T_{d^u_i})}{\sum_{t_p \in V_T} \sum_{d_i \in D_u} P(t_p|C\text{-}T_{d^u_i})} \tag{5.23}$$

— **Le Profil Utilisateur Sémantique** $P^u_{C\text{-}W}$ basé sur le modèle de Contenu Sémantique 5.3.4

$$P(t_o|P^u_{C\text{-}W}) = \frac{\sum_{d_i \in D_u} P(t_o|C\text{-}W_{d^u_i})}{\sum_{t_p \in V_T} \sum_{d_i \in D_u} P(t_p|C\text{-}W_{d^u_i})} \tag{5.24}$$

## 5.5 Modèle de recherche d'information sociale personnalisée

Dans cette section, nous présentons les différents modèles de RISP pour évaluer les profils utilisateurs que nous avons proposés dans ce chapitre.

Similairement au chapitre 5, nous employons un modèle d'ordonnancement qui est basé sur une combinaison linéaire du score de pertinence du document à la requête et un score de pertinence du document pour le profil de l'utilisateur.

Pour un utilisateur $u$ qui soumet la requête $q$, le document $d$ va avoir un score d'ordonnancement estimé comme suit :

$$RSV(q,d,u) = \beta_C RSV(q,d) + (1-\beta_C) RSV(d,u) \tag{5.25}$$

avec $RSV(q,d)$ représente le score correspondance entre le document $d^u_i$ et la requête $q$ et $RSV(u,d)$ représente le score de correspondance entre le profil utilisateur et le document respectivement, et $\beta_C$ un paramètre dans [0,1].

Le profil de l'utilisateur $u$ est estimé en utilisant les modèles que nous avons proposés dans les sections précédentes. Nous proposons ainsi 4 modèles de RISP qui sont :

— Le modèle $\mathbf{MC^u_B}$ où le profil de l'utilisateur est estimé avec le modèle $P^u_{C\text{-}B}$.

$$RSV(q,d,u) = \beta_c RSV(q,d) + (1-\beta_c) RSV(d, P^u_{C\text{-}B}) \tag{5.26}$$

— Le modèle $\mathbf{MC^u_S}$ où le profil de l'utilisateur est estimé avec le modèle $P^u_{C\text{-}S}$.

$$RSV(q,d,u) = \beta_c RSV(q,d) + (1-\beta_c) RSV(d, P^u_{C\text{-}S}) \tag{5.27}$$

— Le modèle $\mathbf{MC^u_T}$ où le profil de l'utilisateur est défini en employant le modèle $P^u_{C\text{-}T}$.

$$RSV(q,d,u) = \beta_c RSV(q,d) + (1-\beta_c) RSV(d, P^u_{C\text{-}T}) \tag{5.28}$$

— Le modèle $\mathbf{MC^u_W}$ où le profil de l'utilisateur est modélisé par le modèle $P^u_{C\text{-}W}$.

$$RSV(q,d,u) = \beta_c RSV(q,d) + (1-\beta_c) RSV(d, P^u_{C\text{-}W}) \tag{5.29}$$

Les scores $RSV(q,d)$ et $RSV(u,d)$ sont estimés avec le modèle de langue avec lissage de Dirichlet. Le paramètre de lissage $\mu$ est fixé, dans toutes nos expérimentations, à la valeur classique 2500, qui est celle par défaut de systèmes comme Terrier.

## 5.6 Conclusion

Dans le chapitre précédent 4, nous avons présenté une première approche de modélisation de l'utilisateur basé sur les tags en exploitant les documents pour la pondération des tags. Nous avons proposé trois modèles utilisateurs et chaque modèle exploite un modèle de Tags de document comme suit :

— Le modèle $\mathbf{MT^u_S}$ où le profil de l'utilisateur est estimé avec le modèle de Tag Standard $P^u_{T\text{-}S}$.
— Le modèle $\mathbf{MT^u_T}$ où le profil de l'utilisateur est définit en employant le modèle de Tag Thématique $P^u_{T\text{-}T}$.
— Le modèle $\mathbf{MT^u_W}$ où le profil de l'utilisateur est modélisé par le modèle de Tag Sémantique $P^u_{T\text{-}W}$.

L'évaluation de ces modèles est présentée dans le chapitre 8.

Dans ce chapitre, nous avons présenté une approche de modélisation l'utilisateur basé sur les documents en exploitant les tags pour la pondération des termes du document. Précisément, cette approche propose de faire dépendre les termes du document non seulement du contenu textuel du document mais également des tags attribués par l'utilisateur à ce document. Le but est de déterminer les termes importants du document qui reflètent les centres d'intérêts de l'utilisateur.

Nous avons proposé quatre modèles utilisateurs, où chaque modèle se base sur un modèle de Tag spécifique :

— Le modèle $\mathbf{MC^u_B}$ où le profil de l'utilisateur est estimé avec le modèle $P^u_{C\text{-}B}$. Le profil de l'utilisateur $P^u_{C\text{-}B}$, prend en compte tous les tags de l'utilisateur.
— Le modèle $\mathbf{MC^u_S}$ où le profil de l'utilisateur est estimé avec le modèle $P^u_{C\text{-}S}$. Le profil de l'utilisateur $P^u_{C\text{-}S}$, intègre le modèle de Tag Standard.
— Le modèle $\mathbf{MC^u_T}$ où le profil de l'utilisateur est défini en employant le modèle $P^u_{C\text{-}T}$. Le profil de l'utilisateur $P^u_{C\text{-}T}$, intègre le modèle de Tag Thématique.
— Le modèle $\mathbf{MC^u_W}$ où le profil de l'utilisateur est modélisé par le modèle $P^u_{C\text{-}W}$. Le profil de l'utilisateur $P^u_{C\text{-}W}$, intègre le modèle de Tag Sémantique.

L'évaluation de ces modèles est présentée dans le chapitre 7.

# Quatrième partie

# Expérimentations

# Chapitre 6

# Cadre Expérimental

## 6.1 Introduction

Dans ce chapitre, nous présentons les modèles de référence avec lesquelles nous comparons nos propositions en section 6.2. Les contraintes expérimentales sont discutées dans la section 6.3. Les collections de test font l'objet de la section 6.3. Enfin, les paramètres des différents systèmes sont présentés en section 6.5.

## 6.2 Modèles de références

Pour valider et comparer nos contributions, nous choisissons un certain nombre de travaux de l'état de l'art qui s'inscrivent dans la même problématique que nous souhaitons résoudre. Pour rappel, le but de cette contribution est de proposer une approche qui permet d'avoir une meilleure modélisation de l'utilisateur afin d'améliorer les performances des systèmes de recherche d'information personnalisée.

Cependant, nous ne nous focalisons dans ces expérimentations que sur l'évaluation du modèle utilisateur et non sur les modèles des systèmes de recherche d'information personnalisée. Ainsi, nous choisissons les différentes approches proposées pour modéliser l'utilisateur au travers des tags et des documents. Nous notons que les modèles utilisateurs des travaux choisis sont utilisés et repris dans la plupart des travaux de l'état de l'art.

Nous choisissons des travaux qui modélisent l'utilisateur avec les tags et des travaux qui modélisent l'utilisateur avec les documents.

1. **Travaux qui modélisent l'utilisateur avec les tags**
   Le profil de l'utilisateur est construit en utilisant les tags, que l'utilisateur a employés pour annoter ses documents, et est représenté comme suit :

$$P_T^u = \{< tg_1, w_{tg_1} >, < tg_2, w_{tg_2} >, ..., < tg_y, w_{tg_Y} >\} \tag{6.1}$$

   — "Xu" [10]. La formule de pondération du tag de l'utilisateur est basée sur la formule suivante :

$$w_{tg_i} = tf(tg_i, D_u) * idf \tag{6.2}$$

   où $tf(tg_i, D_u)$ désigne le nombre de fois que l'utilisateur $u$ a utilisé le tag $tg_i$ pour annoter ses documents et $idf$ qui désigne la fréquence inverse de document (Inverse Document Frequency).

   — "Bouadjenek" [12, 18]. La formule de pondération du tag de l'utilisateur est basée sur la formule suivante :

$$w_{tg_i} = tf(tg_i, D_u) \times \log(\frac{|U|}{|U_{tg_i}|}) \tag{6.3}$$

où $tf(tg_i, D_u)$ désigne le nombre de fois que l'utilisateur $u$ a utilisé le tag $tg_i$ pour annoter ses documents, $U$ est le nombre d'utilisateur et $|U_{tg_i}|$ le nombre d'utilisateur qui ont utilisé le tag $tg_i$ pour annoter leurs documents.

— "Cai" [11, 13]. La formule de pondération du tag de l'utilisateur est basée sur la formule suivante :

$$w_{tg_i} = \frac{tf(tg_i, D_u)}{D_u} \tag{6.4}$$

où $tf(tg_i, D_u)$ désigne le nombre de fois que l'utilisateur $u$ a utilisé le tag $tg_i$ pour annoter ses documents et $D_u$ est le nombre de documents annotés par l'utilisateur $u$.

2. **Travaux qui modélisent l'utilisateur avec les documents**
Le profil de l'utilisateur est construit en utilisant les documents qu'il a annoté, et est représenté comme suit :

$$P_T^u = \{< t_1, w_{t_1} >, < t_2, w_{t_2} >, ..., < t_S, w_{t_S} >\} \tag{6.5}$$

— "Carman" [7]. La formule de pondération des termes est basée sur le modèle de langue standard, sur l'ensemble des documents annotés par l'utilisateur.

$$w_{t_i} = \frac{tf(t_i, D_u)}{|D_u|} \tag{6.6}$$

## 6.3 Contraintes expérimentales

Dans cette section, nous présentons dans un premier lieu les contraintes expérimentales auxquelles nous sommes confrontés pour nos évaluations.

L'évaluation d'un système de RISP est un énorme défi. Les jugements de pertinence doivent être exclusivement et impérativement fournis par les utilisateurs ayant soumis les requêtes au système de recherche d'information personnalisée. La difficulté est qu'il n'existe aucune collection de test standard pour l'évaluation de l'efficacité des systèmes de recherche d'information personnalisée. Fournir de telles collections est très complexe à cause de la difficulté de la mise en place d'un système qui permet de collecter les informations des utilisateurs, leurs requêtes et ensuite les jugements de pertinence pour tous les documents pour chaque requête pour chaque utilisateur.

Plusieurs travaux se sont intéressés à établir et fournir un cadre expérimental complet pour les systèmes de folksonomies [19, 129, 130]. Dans ces travaux, les auteurs attestent que l'activité d'annotation dans ces systèmes reflète, ou peut être considérée comme étant similaire à l'activité de l'utilisateur dans un système de RI classique. Plus précisément, un utilisateur qui annote un document avec un tag aura tendance à sélectionner un document si ce document fait partie de la liste des résultats d'une recherche ayant comme requête le tag de cet utilisateur. Dans d'autres termes, les tags d'un utilisateur peuvent être utilisés comme des requêtes pour ce même utilisateur. De plus, cet utilisateur jugera pertinent tous les documents qu'il

aurait annoté avec ces tags. Plus formellement, l'idée est décrite par l'hypothèse suivante :

— *Pour une requête $Q_u = \{tg_j\}$ de l' utilisateur u tel que $tg_j$ est un tag utilisé par l'utilisateur u, tout les documents annotés par l'utilisateur u avec le tag $tg_j$ sont considérés comme étant pertinent pour l'utilisateur u pour la requête $Q_u$ .*

Partant de cette hypothèse, tout système proposant une structure de folksonomie (utilisateur, document, tag), pourrait utiliser cette hypothèse et donc fournir un cadre d'évaluation pour les systèmes de recherche d'information sociale personnalisée.

Très peu de collections de test publiques existent et sont exploitées dans les systèmes de recherche d'information personnalisée. Ces collections de test permettent d'évaluer un ou plusieurs aspects des systèmes de recherche d'information personnalisée. A notre connaissance, les seules collections existantes sont :

1. Delicious [1] : un réseau social fournissant aux utilisateurs des services d'annotations, de partages, de stockages de documents.
2. CLEF Social Book Search [2] : une collection de test fournis par la tâche d'évaluation CLEF.
3. TREC Contextual Suggestion [3] : une collection de test fournie par la tâche d'évaluation TREC.

Pour l'évaluation de nos propositions, nous pourrions éventuellement utiliser ces collections. Cependant, ces collections n'offrent pas tous les éléments indispensables pour évaluer nos approches. En effet, nous proposons des modèles qui exploitent conjointement les documents des utilisateurs et les tags que l'utilisateur a attribués à ses documents. Ce cadre est représenté par le triplet : <*utilisateur, document, tags*>. Les collections adaptées doivent satisfaire ces trois critères importants :

— L'utilisateur qui soumet les requêtes doit avoir un ensemble de documents et leurs tags associées (attribué par cet utilisateur).
— Nous devons avoir assez de données pour pouvoir construire les requêtes des utilisateurs.
— Il faut avoir la possibilité de créer des jugements de pertinence personnalisés.

Nous présentons dans le tableau 6.1 les collections de test et les contraintes à satisfaire.

| Datatset | $Tags_{u_q}$ | $Documents_{u_q}$ | Qrels | Requêtes |
|---|---|---|---|---|
| Trec Contextual Suggestion | ≅ | ✓ | ✓ | ✖ |
| Social Book Search | ≅ | ✓ | ✓ | ✓ |
| Delicious | ✓ | ✓ | ✖ | ✖ |

TABLE 6.1 – Contraintes sur les collections de test, ✓ : Un nombre important ; ✖ : Aucun, ≅ : Très peu

Le tableau 6.1 montre que aucune collection de test ne satisfait complètement les contraintes pour l'évaluation de nos propositions.

1. http ://del.icio.us/
2. http ://social-book-search.humanities.uva.nl/#/overview
3. https ://sites.google.com/site/treccontext/

— La collection de test TREC Contextual Suggestion ne fournit pas de requêtes et les utilisateurs sont représentés par un contexte, les votes et les tags. Les tags des utilisateurs sont très peu nombreux (plus de 80 % des utilisateurs n'ont pas de tags). Donc, par manque de tags, nous ne pouvons pas générer les requêtes.
— Pour la collection de test Social Book Search, le problème est similaire. Les données de l'utilisateur sont un ensemble de catalogues représentants les livres achetés par l'utilisateur et très peu de tags sont fournis (beaucoup d'utilisateurs ont des profils vides c'est-à-dire aucun catalogue et aucun tag).
— Enfin, la collection de test Delicious satisfait les deux conditions nécessaires. Cette collection n'offre pas les jugements de pertinence et les requêtes, mais comme nous disposons d'une grande quantité de tags, nous pouvons alors générer les requêtes et les jugements de pertinence automatiquement. Et donc, il est nécessaire de développer notre propre collection de test.

## 6.4 Construction de la collection de test

Pour construire notre collection de test, nous suivons le même protocole suivi dans les travaux de l'état de l'art [6, 7, 21, 109]. Les différentes étapes sont : (1) la collecte de documents, (2) la construction des requêtes et enfin (3) la génération des jugements de pertinence.

### 6.4.1 Collecte des documents

L'évaluation des systèmes de RISP requiert un ensemble de requêtes des utilisateurs, les documents, les jugements de pertinence personnalisés et les données des utilisateurs pour construire leurs profils. Comme présenté précédemment, seule la collection de test Delicious satisfait toutes les contraintes. Toutefois, cette collection ne fournit pas réellement les requêtes et les jugements de pertinence personnalisés mais compte tenu de la quantité des données, il est donc possible de les générer. Pour avoir une collection de test complète nous suivons les étapes suivantes :

1. **Étape 1 : Nettoyage préliminaire**
   Avant de procéder à la construction de la collection de test, nous effectuons un nettoyage des annotations des utilisateurs. Nous supprimons tous les tags navigationnels contenant ("http","https", "www.", ".com", ".net", ".org", ".edu").
2. **Étape 2 : Récupération des documents**
   Nous récupérons tous les documents en utilisant les urls. Nous prenons en compte seulement les documents en Anglais. A l'issue de cette étape, nous ne récupérons qu'un certain nombre de documents car certaines pages web sont erronées ou elles n'existent plus, ou ne sont plus accessibles (domaine protégé).
3. **Étape 3 : Sélection des utilisateurs**
   Afin d'avoir des profils utilisateurs variés et non vides, nous avons choisi certains critères et qui sont aussi utilisés dans la plupart des travaux de l'état de l'art [7, 11, 12, 111] :
   — Nous avons choisi de ne garder que les documents avec des tags. Les documents non annotés sont supprimés.
   — Nous avons éliminé les tags qui sont des tags de domaines comme : youtube, facebook, ...

— Seuls les utilisateurs ayant plus de 100 documents différents sont pris en compte.
— Seuls les documents ayant plus de 10 tags sont pris en compte.

| | |
|---|---|
| Nombre de signets (bookmarks) | 2.715.390 |
| Nombre de document | 211.205 |
| Nombre d'utilisateurs | 395 |
| Nombre de tags uniques | 59.758 |

TABLE 6.2 – Statistiques sur la collection de test finale

### 6.4.2 Générations des requêtes

Comme présenté précédemment dans la section 6.3, nous suivons un protocole pour générer automatiquement les requêtes des utilisateurs et les jugements de pertinence.

Nous générons les requêtes à partir des triplets de chaque utilisateur. C'est-à-dire, pour chaque utilisateur, nous sélectionnons aléatoirement un ensemble de triplets $< u, tg, d >$, ainsi la requête va être le tag $tg$ et le document pertinent va être $d$. Sur l'ensemble des utilisateurs, nous avons généré 6760 requêtes.

### 6.4.3 Génération des jugements de pertinence

Après avoir généré les requêtes des utilisateurs, nous générons automatiquement les jugements de pertinence pour notre collection de test. Comme présenté dans la section 6.3, les jugements de pertinence sont générés en se basant sur l'hypothèse que tous les documents annotés avec les termes de la requête par l'utilisateur sont considérés pertinents pour cette requête. Ainsi, de l'ensemble de requêtes générées dans la section précédente, nous générons les jugements de pertinence.

## 6.5 Paramètres des systèmes

Dans cette partie, nous détaillons les paramètres et l'apprentissage des différents paramètres de chaque modèle.

### 6.5.1 Apprentissage des plongements des mots

Pour générer les vecteurs des termes de la collection, nous avons utilisé le modèle de ***word2vec*** proposé par [25] sur notre collection de test. Pour les paramètres d'apprentissage, nous avons utilisé majoritairement les valeurs conseillées par [136]. Les valeurs des paramètres sont listées dans le Tableau 6.3.

### 6.5.2 Génération des thèmes LDA

Pour l'apprentissage du modèle LDA, nous avons utilisé Mallet[4] implémenté par [137] pour estimer les distributions des termes dans les thèmes et les distributions des thèmes dans les documents. Nous avons observé que le choix des hyperparamètres affecte très peu les performances des systèmes. Cependant, nous avons

4. https ://people.cs.umass.edu/ mccallum/mallet/

| Paramètres | Valeurs |
|---|---|
| Modèle | Skip-Gram |
| Dimensions des vecteurs | 300 |
| Dimension de la fenêtre de contexte | 8 |
| Negative sampling | k = 10 |
| Down sampling | t = $10^{-}5$ |
| cds | $\alpha$ = 0,75 |
| Nombre d'itération | 20 |
| Fréquence minimales des termes | >= 10 |

TABLE 6.3 – Paramètres d'apprentissage de word2vec

expérimenté suivant le nombre de thèmes $k = [100, 200, 300, 400, 500, 600, 700]$, pour les autres paramètres, nous avons utilisé les valeurs conseillées dans [24].

### 6.5.3 Paramètres des modèles

Pour tous les documents et les requêtes, nous avons supprimé tous les mots vides en utilisant la liste standard de mots vides [5]. Nous notons que nous n'avons pas utilisé les algorithmes de lemmatisation. Pour toutes nos expérimentations, les fonctions d'ordonnancement des documents sont calculées avec le modèle de langue avec lissage de Dirichlet où $\mu$ est fixé à 2500.

Pour chaque modèle de l'état de l'art, nous avons conduit une évaluation sur deux plis afin de prendre les meilleurs paramètres pour chaque modèle. Nous avons utilisé 1352 requêtes pour l'estimation des paramètres des modèles. Les 5408 requêtes restantes sont utilisées pour évaluer nos contributions et les comparer aux modèles de l'état de l'art. Les valeurs des paramètres des modèles seront présentés dans les chapitres 7 et 8.

## 6.6 Conclusion

Dans ce chapitre, nous avons décrit l'ensemble des paramètres expérimentaux ainsi que la construction d'une collection de test afin d'évaluer nos propositions. Nous avons également listé les propositions auxquelles nous allons nous comparer dans les deux chapitres 7 et 8 qui suivent.

5. https ://github.com/nawalouldamer/english-words

# Chapitre 7

# Évaluation du profil utilisateur basé sur les tags

## 7.1 Introduction

Dans ce chapitre nous présentons les résultats des expérimentations menées pour évaluer la qualité et la robustesse de nos propositions. Ces évaluations ont pour objectif d'évaluer les différents modèles utilisateurs basés sur les tags, que nous avons proposé dans le chapitre 4, où chaque modèle proposé s'appuie sur une hypothèse de sa capacité à identifier les tags les plus importants et représentatifs des thématiques auxquelles s'intéresse l'utilisateur. Nos modèles sont construits autour des hypothèses suivantes :

- **H1** : Les tags importants d'un document annoté par un utilisateur sont ceux qui sont présents dans le document, donc les termes du document que l'utilisateur emploie comme tag pour annoter le document. Pour répondre à cette hypothèse, nous avons proposé le modèle $MT_S^u$ qui exploite le profil utilisateur noté $P_{T\text{-}S}^u$.
- **H2** : Cette hypothèse stipule que les tags importants sont ceux qui traitent des thématiques des documents auxquels ils ont été assignés par l'utilisateur. Nous avons employé les modèles probabilistes thématiques LDA et avons proposé le modèle $MT_T^u$ où l'utilisateur est représenté par le profil thématique noté $P_{T\text{-}T}^u$.
- **H3** : Le modèle $MT_W^u$ apporte une solution à l'hypothèse qui atteste que les tags qui représentent l'utilisateur sont ceux qui sont sémantiquement similaires aux termes du documents de l'utilisateur - dans le même contexte sémantique. Nous avons employé les modèles de plongements de mots (word2vec) pour estimer le profil de l'utilisateur sémantique noté $P_{T\text{-}W}^u$.

Pour vérifier la validité de nos hypothèses, nous procédons à une évaluation globale pour chaque modèle proposé afin d'évaluer la capacité de chaque modèle à identifier les tags importants et estimer leur impact sur les performances du système. Ensuite, nous évaluons la robustesse des systèmes en évaluant leurs performances sur chaque requête de l'utilisateur.

## 7.2 Paramètres des modèles

L'objectif de cette expérimentation est d'évaluer nos modèles en comparant leurs performances à celles des modèles de l'état de l'art pour la tâche d'ordonnancement de documents personnalisé.

Pour toutes nos propositions sauf pour le modèle unifié 7.4, nous avons conduit une évaluation croisée à deux plis. Pour chaque modèle, nous avons choisi les valeurs des paramètres qui donnaient les meilleures performances. Les modèles et les valeurs des paramètres sont :

— Le modèle $\mathbf{MT^u_S}$ où le profil de l'utilisateur est estimé avec le modèle $P^u_{T\text{-}S}$.

$$RSV(q,d,u) = \beta_{TG} RSV(q,d) + (1-\beta_{TG}) RSV(d, P^u_{T\text{-}S}) \tag{7.1}$$

Où la $\beta_{TG}$ du modèle est de 0.54. Pour la partie $RSV(d, P^u_{T\text{-}S})$, le modèle $P^u_{T\text{-}B}$ présenté dans la formule 4.9 n'a aucun paramètre.

— Le modèle $\mathbf{MT^u_T}$ où le profil de l'utilisateur est défini en employant le modèle $P^u_{C\text{-}T}$.

$$RSV(q,d,u) = \beta_{TG} RSV(q,d) + (1-\beta_{TG}) RSV(d, P^u_{T\text{-}T}) \tag{7.2}$$

Où la $\beta_{TG}$ du modèle est de 0.43. Pour le la partie $RSV(d, P^u_{T\text{-}T})$, le modèle $P^u_{T\text{-}T}$ présenté dans la formule 4.10, présente un paramètre qui est le nombre de thèmes pour le LDA que nous expérimentons par la suite.

— Le modèle $\mathbf{MT^u_W}$ où le profil de l'utilisateur est modélisé par le modèle $P^u_{T\text{-}W}$.

$$RSV(q,d,u) = \beta_{TG} RSV(q,d) + (1-\beta_{TG}) RSV(d, P^u_{C\text{-}W}) \tag{7.3}$$

Où la $\beta_{TG}$ du modèle est de 0.34. Pour le la partie $RSV(d, P^u_{T\text{-}W})$, le modèle $P^u_{T\text{-}W}$ présenté dans la formule 4.11, présente un paramètre qui est le seuil de similarité entre le tag et le terme du document que nous expérimentons par la suite.

## 7.3 Évaluation globale des modèles

Cette évaluation a pour objectif d'étudier et explorer les trois hypothèses H1, H2, et H3 et mesurer l'impact et l'efficacité des modèles à générer des profils utilisateurs efficaces et par conséquent mesurer les performances par rapport aux modèles de l'état de l'art.

Dans la suite de la section, nous présentons les résultats d'expérimentation de chaque modèle séparément en étudiant la spécificité de chaque modèle. Chaque modèle ( $MT^u_S$, $MT^u_T$, $MT^u_W$) est comparé aux modèles de l'état de l'art présentés en section 6.2. Nous rapportons les résultats sur les quatre mesures d'évaluation MAP, MRR, et P@5, et nous considérons la MAP comme étant la mesure d'évaluation principale.

### 7.3.1 Évaluation du modèle Standard $MT^u_S$

Cette section est consacrée à l'évaluation du modèle $MT^u_S$ qui exploite le profil utilisateur basé sur les tags avec une estimation standard s'appuyant sur l'hypothèse H1. Les résultats de l'expérimentation sont présentés dans le tableau 7.1.

| Systèmes | MAP | MRR | P@5 |
|---|---|---|---|
| Bouadjenek | 0.3901 | 0.3957 | 0.1007 |
| Cai | 0.3597 | 0.3656 | 0.0940 |
| Xu | 0.2689 | 0.2727 | 0.0712 |
| $MT_S^u$ | **0.4024** | **0.4073** | **0.1061** |

TABLE 7.1 – Comparaison de l'efficacité de notre modèle $MT_S^u$ avec les modèles de l'état de l'art Bouadjenek, Cai, Xu. Les meilleurs résultats obtenus sont présentés en gras.

D'après les résultats présentés dans le tableau 7.1, on observe que le modèle $MT_S^u$ obtient les meilleurs résultats en comparaison avec les modèles de l'état de l'art, et cela quelle que soit la mesure d'évaluation.

Les résultats obtenus montrent que le modèle $MT_S^u$ est capable d'identifier et d'estimer de manière efficace les tags relatifs à chaque document et ainsi représenter les sujets d'intérêts de l'utilisateur, ce qui confirme notre hypothèse H1, et qui en finalité impacte positivement les performances du système de recherche.

En complément des expérimentations globales du modèle, nous avons conduit une analyse quantitative supplémentaire qui mesure le gain personnalisé P-Gain qui évalue la stabilité et la robustesse d'un système [138]. Cette mesure permet d'avoir une évaluation au niveau de chaque requête en rapportant les taux de requêtes améliorées et détériorées par nos systèmes (cf. 3.5.2). Les résultats obtenus sont présentés dans le tableau 7.2.

| Système | **#Q** | **#Q$_+$** | **#Q$_-$** | **#Q$_{==}$** | **P-Gain** |
|---|---|---|---|---|---|
| $\Delta(MT_S^u$-Bouadjenek) | 5408 | 991 (+18.3%) | 960 (-17.7%) | 3457 (64%) | 0.016 |
| $\Delta(MT_S^u$-Cai) | 5408 | 1225 (+22%) | 872 (-16%) | 3311 (62%) | 0,168 |
| $\Delta(MT_S^u$-Xu) | 5408 | 1786 (+33%) | 784 (-15%) | 2838 (52%) | 0.389 |

TABLE 7.2 – Les résultats d'évaluation au niveau de chaque requête avec le nombre et taux de requêtes améliorées et détériorées par notre système $MT_S^u$ et les taux des gains personnalisés obtenus par rapport aux modèles de l'état de l'art Bouadjenek, Cai et Xu.

Les résultats obtenus montrent que globalement le système $MT_S^u$ présente des gains personnalisés positifs par rapport à tous les systèmes de l'état de l'art et ces résultats sont en phase avec ceux mesurés dans l'évaluation précédente (Tableau 7.1). Nous notons des taux importants de 16.8% et 38.9% par rapport aux modèles de Cai et Xu respectivement. Mais nous n'obtenons qu'un gain faible de 1.6% par rapport au modèle de Bouadjenek, car le système améliore et détériore presque le même nombre de requêtes ($\#Q_+ = 991$ et $\#Q_- = 960$). Nous expliquons ces résultats par le fait, que le système $MT_S^u$ qui se base sur un profil estimé par le modèle de Tag Standard élimine des tags qui peuvent être importants car ils ne sont pas présents dans les documents (les tags ne sont pas des termes du document) ce que nous avons relevé comme limite du modèle. Ceci n'est pas la seule explication des raisons pour lesquelles notre modèle dégrade les résultats.

Nous avons fait une analyse qualitative sur les requêtes afin d'identifier le type de requête pour lequel notre système est sensible. En analysant les profils des utilisateurs associés aux requêtes pour lesquelles notre système dégrade les résultats, nous avons remarqué que certains tags avaient des poids faibles dans les profils générés

par notre modèle, par contre, ils avaient des poids importants dans les profils générés par les modèles de l'état de l'art (notre modèle leur a attribué des poids faibles à la différence des autres modèles).

De plus, en analysant les documents auxquels ils sont associés, nous avons constaté que notre modèle défavorise certains tags malgré le fait qu'ils sont des termes du document car ils n'ont pas un poids important (leur fréquence dans le document est faible). Le biais est dû à la génération des requêtes pour évaluer les systèmes. Comme les requêtes sont générées à partir des tags de l'utilisateur, donc si ces requêtes ne sont pas de type informatif (des tags de type point de vue, une tâche à faire, etc) donc, notre système va détériorer les résultats. Par exemple, pour la requête utilisateur "job search", l'utilisateur a annoté des documents avec ce tag car il est entrain de rechercher un emploi, il a annoté tous les documents relatifs à sa recherche d'emploi "job search" donc le sujet du document n'a rien à voir avec la recherche d'emploi car il décrit un emploi comme pour une annonce d'emploi. Cet utilisateur en particulier, va annoter plein de document avec "job", "search job", etc.

Notre système va attribuer un poids faible (si le terme est présent dans le document) ou nul si le tag n'est pas présent dans le document. Par contre, dans les autres approches de l'état de l'art, ce tag va avoir un poids important d'autant plus si l'utilisateur a annoté beaucoup de documents avec ce tag. En conséquence, pour la requête "job search" notre système va avoir de faibles performances à la différence des autres modèles.

Nous présumons que si une classification des requêtes ou une génération contrôlée des requêtes est faite, notre système pourrait alors avoir de meilleurs résultats. Par contre, cela ne pourra pas résorber le problème pour les tags importants mais qui sont éliminés par notre modèle dû à la contrainte véhiculée par l'hypothèse H1.

### 7.3.2 Évaluation du modèle Thématique $MT_T^u$

Dans cette section, nous évaluons le modèle thématique $MT_T^u$. Ce dernier exploite les modèles probabilistes thématiques et repose sur l'hypothèse H2. Cette hypothèse stipule que les tags doivent couvrir les thèmes du document auquel ils sont associés. Pour évaluer le modèle, dans un premier temps, nous avons fait un apprentissage sur la collection de document afin de découvrir les thèmes de la collection en employant le modèle LDA avec les paramètres par défaut [1]. L'apprentissage est réalisé pour différentes valeurs du nombre de thèmes Z = 100, 200, 300, 400, 500, 600, 700. Les résultats de cette expérimentation sont présentés dans la figure 7.1.

1. Valeurs des paramètres présentés dans le papier [24]

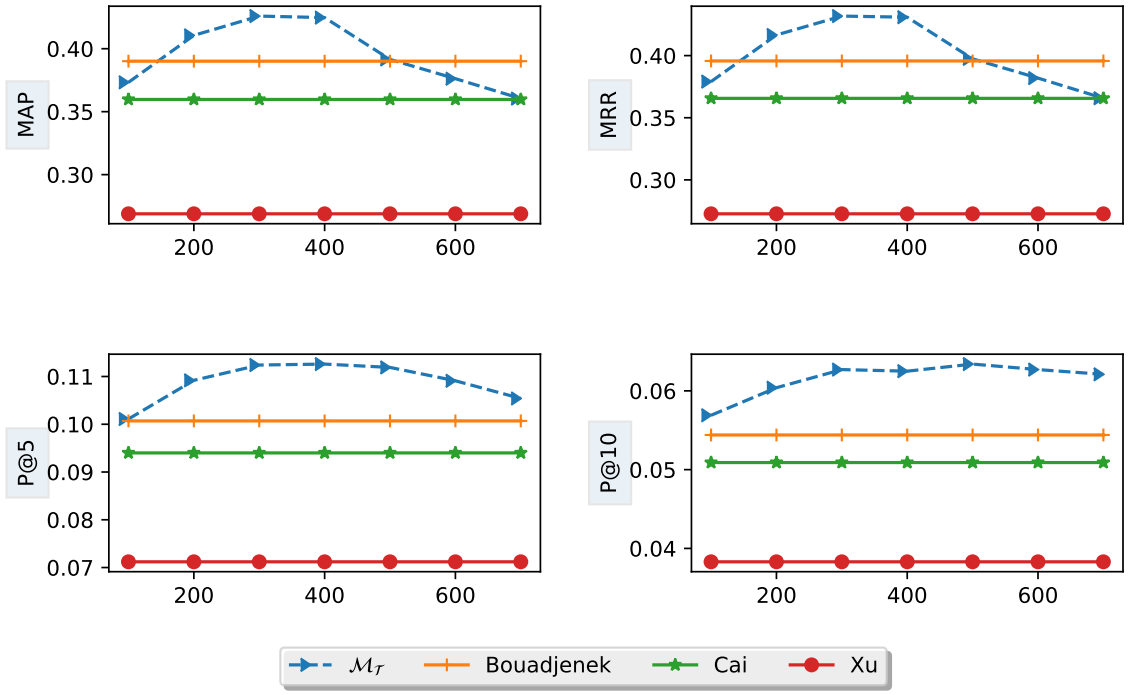



FIGURE 7.1 – Comparaison de l'efficacité du modèle $MT_T^u$ (présenté $M_T$ dans la figure) par suivant les valeurs de nombre de thème K

De façon globale, les résultats obtenus montrent que le modèle thématique $MT_T^u$ obtient les meilleurs résultats en valeur de MAP, MRR, et P@5. Nous observons que le modèle Bouadjenek obtient les meilleurs résultats en valeur de MAP et MRR pour les valeurs de thèmes Z = 600, 700. Nous pensons que cela est dû au fait que pour les nombre de thèmes est très important et donc il est nécessaire de choisir le bon nombre de thèmes.

Dans cette expérimentation, nous avons choisis arbitrairement de considérer la valeur de nombre de thèmes qui est égale à 400 thèmes. Les résultats des évaluations sont présentés dans le tableau 7.3.

| Systèmes | MAP | MRR | P@5 |
|---|---|---|---|
| Bouadjenek | 0.3901 | 0.3957 | 0.1007 |
| Cai | 0.3597 | 0.3656 | 0.0940 |
| Xu | 0.2689 | 0.2727 | 0.0712 |
| $MT_T^u$ | **0.4259** | **0.4318** | **0.1124** |

TABLE 7.3 – Évaluation de l'efficacité de notre modèle $MT_T^u$ avec le modèle LDA entraîné sur 400 thèmes, et comparaison avec les modèles de l'état de l'art Bouadjenek, Cai, Xu. Les meilleurs résultats obtenus sont présentés en gras.

Les résultats obtenus montrent que le modèle thématique $MT_T^u$ dépasse toutes les valeurs obtenues par les modèles de l'état de l'art. Ces résultats sont expliqués par le fait que notre modèle $MT_T^u$ permet de sélectionner et de donner plus d'importance aux tags de l'utilisateur qui sont dans le même espace latent que les documents annotés par cet utilisateur.

Nous avons également évalué les performances du modèle $MT_T^u$ au niveau de chaque requête et avons mesuré les gains personnalisés obtenus. Les résultats relatifs à cette évaluation sont présentés dans le tableau 7.4.

| Système | **#Q** | **#Q$_+$** | **#Q$_-$** | **#Q$_{==}$** | **P-Gain** |
|---|---|---|---|---|---|
| $\Delta(MT_T^u$-Bouadjenek) | 5408 | 1060 (+20%) | 805 (-15%) | 3543 (65%) | 0.136 |
| $\Delta(MT_T^u$-Cai) | 5408 | 1477 (+27%) | 747 (-13%) | 3377 (60%) | 0.328 |
| $\Delta(MT_T^u$-Xu) | 5408 | 1824 (+34%) | 682 (-12%) | 2902 (54%) | 0.455 |

TABLE 7.4 – Les résultats d'évaluation au niveau de chaque requête avec le nombre et taux de requêtes améliorées et détériorées par notre système $MT_T^u$ et les taux de gain personnalisés obtenus par rapport aux modèles de l'état de l'art Bouadjenek, Cai et Xu.

Nous observons que le modèle $MT_T^u$ exploitant le profil thématique obtient des gains personnalisés importants et positifs allant de 13.6% à 45.5%. Ces gains montrent que notre système est capable de bien générer les profils des utilisateurs. Nous notons aussi que le modèle obtient des gains supérieurs par rapport au modèle précédent $MT_S^u$ car il prend en compte d'une part les tags qui sont présents dans le document et d'autre part, les tags qui couvrent les thèmes du document. Néanmoins, tout comme le modèle $MT_S^u$, le modèle $MT_T^u$ dégrade les résultats de certaines requêtes par rapport à chaque modèle. L'une des explications est la même que celle du modèle précédent et l'autre vient du fait que pour les tags non présents dans la collection des documents, alors ils ne seront pas dans les thèmes générés par LDA, ce qui donnera des résultats nuls pour la requête.

### 7.3.3 Évaluation du modèle Sémantique $MT_W^u$

Dans cette partie, nous présentons et détaillons les évaluations du modèle $MT_W^u$ qui exploite un profil construit en utilisant les plongements de mots pour identifier les tags importants des documents et qui vont mieux représenter les sujets d'intérêts de l'utilisateur, et repose sur l'hypothèse H3.

L'estimation des poids des tags repose sur un calcul de similarité entre les tags et les termes du document comme présenté dans la formule 5.5 décrite en section 5.3. Dans cette formule, le poids d'un tag est déterminé globalement par la moyenne de ses scores de similarités avec les termes du document. Le point clé est de savoir à partir de quel score de similarité entre deux termes, nous pouvons dire que les deux termes dans le même espace sémantique ont la même signification.

De récentes [139], se sont intéressées à estimer le meilleur seuil de similarité entre deux termes. C'est-à-dire, quelle est la valeur de similarité entre deux termes qui détermine que ces deux termes sont similaires. Par conséquent, nous avons conduit des expériences suivant différentes valeurs de seuil de similarité pour étudier son impact sur les performances du modèle $MT_W^u$. Finalement, seuls les tags ayant une similarité supérieure à un certain seuil sont pris en compte dans l'estimation du modèle.

Dans un premier temps, nous avons réalisé une expérimentation simple pour étudier l'impact de la valeur de seuil sur les performances du système $MT_W^u$. La Figure 7.2 détaille les résultats obtenus suivant les valeurs de seuil qui sont $[0.4, 0.5, 0.6, 0.7]$ dans .

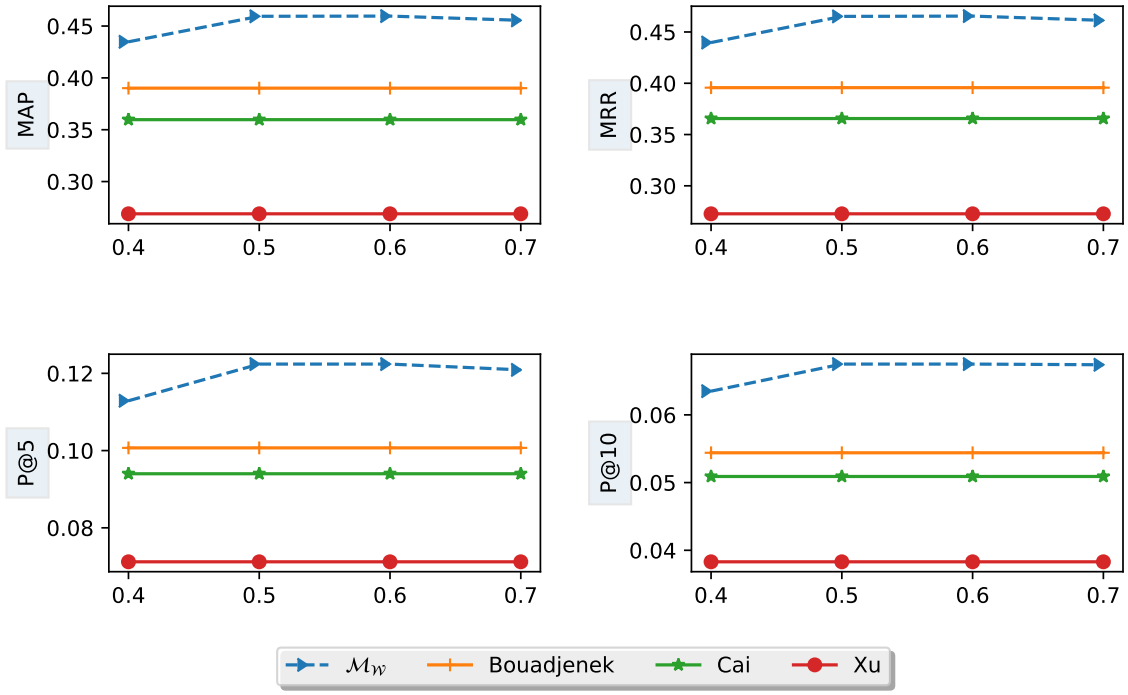



FIGURE 7.2 – Résultats suivant les valeurs de seuil de similarité du modèle $MT^u_W$ (présenté $M_W$ dans la figure)

Les résultats obtenus montrent d'une part que le seuil influe sur les performances du système. En effet, les meilleurs résultats sont obtenus avec la valeur de seuil égale 0.5 et 0.6. Ces seuils sont en adéquation avec les résultats présentés par [139]. De plus, nous constatons que même avec des seuils de similarité faibles (inférieur de 0.5), notre modèle se comporte mieux que les modèles de l'état de l'art. D'autre part, notre modèle $MT^u_W$ dépasse les valeurs obtenues par les modèles de l'état de l'art et cela quelles que soient les mesures d'évaluation.

Nous avons conduit une validation croisée à deux plis pour déterminer la meilleure valeur de seuil de similarité et évaluer notre modèle sur cette valeur. Nous avons obtenu les meilleurs résultats avec le seuil de similarité égale à 0.65. Les résultats sur les différentes mesures de MAP, MRR, et P@5 sont présentés dans le tableau 7.5.

| Systèmes | MAP | MRR | P@5 |
|---|---|---|---|
| Bouadjenek | 0.3901 | 0.3957 | 0.1007 |
| Cai | 0.3597 | 0.3656 | 0.0940 |
| Xu | 0.2689 | 0.2727 | 0.0712 |
| $MT^u_W$ | **0.4555** | **0.4613** | **0.1209** |

TABLE 7.5 – Évaluation de l'efficacité de notre modèle $MT^u_W$ avec un seuil de similarité égale à 0.65 et comparaison avec les modèles de l'état de l'art Bouadjenek, Cai, Xu. Les meilleurs résultats obtenus sont présentés en gras.

Nous observons que notre modèle $MT^u_W$ est plus efficace que tous les modèles de l'état de l'art et cela quelles que soient les mesures d'évaluation.

Ces résultats sont expliqués par le fait que notre modèle $MT^u_W$ permet de sélectionner d'attribuer plus d'importance qu'aux tags de l'utilisateur qui sont dans le même espace sémantique que les termes du document annoté par cet utilisateur.

Nous avons également évalué les performances du modèle $MT^u_W$ au niveau de chaque requête et avons mesuré les gains personnalisés obtenus. Les résultats relatifs à cette évaluation sont présentés dans le tableau 7.6.

| Système | **#Q** | **#Q$_+$** | **#Q$_-$** | **#Q$_{==}$** | **P-Gain** |
|---|---|---|---|---|---|
| $\Delta(MT^u_W$ -Bouadjenek) | 5408 | 1209 (+22%) | 756 (-14%) | 3443 (64%) | 0.230 |
| $\Delta(MT^u_W$ -Cai) | 5408 | 1448 (+26%) | 686 (-13%) | 3274 (61%) | 0.357 |
| $\Delta(MT^u_W$ -Xu) | 5408 | 1944 (+36%) | 627 (-11%) | 2837 (53%) | 0.512 |

TABLE 7.6 – Les résultats d'évaluation au niveau de chaque requête avec le nombre et taux de requêtes améliorées et détériorées par notre système $MT^u_W$ et les taux de gain personnalisés obtenus par rapport aux modèles de l'état de l'art Bouadjenek, Cai et Xu.

Nous observons que le modèle $MT^u_W$ exploitant le profil sémantique obtient des gains personnalisés importants et positifs allant de 23% à 51.2%. Ces gains montrent que notre système est capable de bien générer les profils des utilisateurs. Nous notons aussi que le modèle obtient des gains supérieurs par rapport au modèle précédent $MT^u_T$. Néanmoins, tout comme le modèle $MT^u_T$, le modèle $MT^u_W$ dégrade les résultats de certaines requêtes par rapport à chaque modèle. Les raisons sont similaires à celles des modèles précédents. Mais en plus, une raison supplémentaire vient du fait que les tags ne sont pas présents dans les documents, ainsi dans l'apprentissage des plongements des mots, les tags ne sont pas présents. Ainsi, les similarités entre les tags et les termes sont nulles, ce qui affectent la génération des profils.

## 7.4 Évaluation de la complémentarité des modèles $MT^u_S$, $MT^u_T$ et $MT^u_W$

Nous proposons une évaluation préliminaire d'un modèle unifié $MT^u_U$ qui combine les spécificités de chaque modèle. L'objectif de cette expérimentation est d'étudier l'apport et l'impact du modèle unifié $MT^u_U$ dans la tâche de modélisation de l'utilisateur dans le but d'améliorer les performances du système de recherche d'information personnalisé.

De plus, cette expérimentation permet d'évaluer dans quelle mesure la complémentarité des trois hypothèses **H1, H2 et H3** est capable d'identifier correctement les tags importants qui décrivent les centres d'intérêts de l'utilisateur.

Le poids d'un tag va être fonction du score du tag dans le profil de Tag Standard, du score du tag dans le profil de Tag Thématique et du score du tag dans le profil de Tag Sémantique comme suit :

$$P(tg_j, P^u_{T\text{-}U}) = \frac{\alpha_S P(tg_j, P^u_{T\text{-}S}) + \beta_T P(tg_j, P^u_{T\text{-}T}) + \gamma_W P(tg_j, P^u_{T\text{-}W})}{\sum_{tg_k \in T^u_{d_i}} P(tg_k, P^u_{T\text{-}U})} \tag{7.4}$$

Dans les sections précédentes (7.3.1, 7.3.2, 7.3.3), nous avons évalué chaque modèle que nous avons proposés, et nous avons conduit une évaluation à deux plis pour déterminer les valeurs des paramètres qui donnaient les meilleurs résultats de performance. Pour le modèle unifié, nous n'avons pas conduit une expérimentation pour choisir les meilleures valeurs des paramètres. Nous avons arbitrairement choisis les mêmes valeurs des paramètres pour chaque modèle composant le modèle unifié.

Ces valeurs sont celles présentées dans les sections précédentes. Les différentes valeurs des paramètres du modèle unifié sont comme suit :

— Le profil de Tag Standard $P^u_{T\text{-}S}$ ne dispose d'aucun paramètre.
— Le profil de Tag Thématique $P^u_{T\text{-}T}$) exploite le modèle probabiliste thématique LDA. Pour le nombre de thèmes $z$, que nous avons choisi de garder la même valeur égale à 400, comme présenté en section 7.3.2.
— Le profil de Tag Sémantique $P^u_{T\text{-}S}$ emploi les plongements de mots. Le seul paramètre du modèle est le seuil de similarité entre le tag et le terme (cf. formule 5.5). La valeur du seuil de similarité que nous choisissons est égale à 0.65. Cette valeur est la même que nous avons utilisé en section 7.3.3.
— Pour les paramètres $\alpha_S$, $\beta_T$, et $\gamma_W$, nous avons choisi d'attribuer le même apport pour chaque modèle. Les valeurs de ces paramètres sont : $\alpha_S = 0.33$, $\beta_T = 0.33$, et $\gamma_W = 0.33$.
— Pour le paramètre $\beta_{TG}$, nous avons choisi arbitrairement la valeur $\beta_{TG} = 0.5$.

Donc le modèle $MT^u_U$ où le profil de l'utilisateur est estimé en employant le modèle $P^u_{T\text{-}U}$ comme suit :

$$RSV(q,d,u) = \beta_{TG} RSV(q,d) + (1-\beta_{TG}) RSV(d, P^u_{T\text{-}U}) \tag{7.5}$$

Les résultats de cette expérimentation sont présentés dans le tableau 7.7.

| Systèmes | MAP | MRR | P@5 |
|---|---|---|---|
| $MT^u_S$ | 0.4024 | 0.4073 | 0.1061 |
| $MT^u_T$ | 0.4259 | 0.4318 | 0.1124 |
| $MT^u_W$ | 0.4555 | **0.4613** | 0.1209 |
| $MT^u_U$ | **0.4595** | 0.4603 | **0.1231** |

TABLE 7.7 – Évaluation du modèle $MT^u_U$ qui représente la complémentarité des modèles $MT^u_W$, $MT^u_W$ et $MT^u_W$. Les meilleurs résultats obtenus sont présentés en gras.

Les résultats obtenus montrent que le modèle $MT^u_U$ obtient légèrement de meilleurs résultats que tous les autres modèles.

Ce modèle unifié peut aussi être amélioré du point de vue conceptuel. Le profil de l'utilisateur est une agrégation des poids de tags pour chaque document (une moyenne). Nous pensons qu'une combinaison simple des scores des trois modèles n'est pas un choix judicieux. Donc, un modèle véritablement unifié pourrait mieux tirer avantage des spécificités des chaque modèle.

De plus, le modèle $MT^u_S$ obtient les résultats les plus faibles par rapport aux autres modèles et qui sont prévisibles du fait de la limite du modèle en éliminant tous les tags qui ne sont pas des termes du document. Le modèle $MT^u_T$ obtient de meilleurs résultats que le modèle $MT^u_W$. Nous pensons que ces résultats sont dû à la capacité des modèles de plongements de mots à identifier les termes les plus proches d'un terme par rapport aux modèles probabilistes thématiques.

Malgré les bons résultats obtenus, nous avons toutefois un ensemble de requêtes pour lequel le modèle unifié dégrade les résultats. Nous pensons que les raisons découlent des limites de chaque modèle.

## 7.5 Évaluation de l'impact de la taille du profil de l'utilisateur

Les expérimentations présentées dans les sections précédentes sont basées sur l'exploitation globale du profil de l'utilisateur. Dans les travaux de l'état de l'art, la distinction du profil long-terme et court-terme est couramment expérimentée. Habituellement, l'évaluation des profils à court terme est faite sur les sessions de recherches des utilisateurs. Les sessions de recherches sont identifiées et construites par une approche d'identification d'activité délimitée par un espace de temps d'inactivité d'un utilisateur qui est généralement égal à 30 minutes [140-142].

Dans notre cas, nous n'avons pas de données liées aux temps d'activités des utilisateurs. Donc nous ne pouvons pas construire les sessions de recherches. Pour ce faire, nous simulons ces sessions de recherches en nous appuyant sur l'hypothèse que dans une session de recherche, l'utilisateur n'effectue des recherches que sur un seul sujet. Ceci revient à faire une expansion de requête par des termes du profil. Nous sommes conscients que cette méthode n'est pas vraiment identique à l'évaluation court-terme. Par contre, elle nous permet d'évaluer la capacité de nos modèles à générer des profils précis. Partant de là, nous proposons de classifier le profil de l'utilisateur sur les différentes requêtes.

Un exemple est illustré dans la figure 7.3.

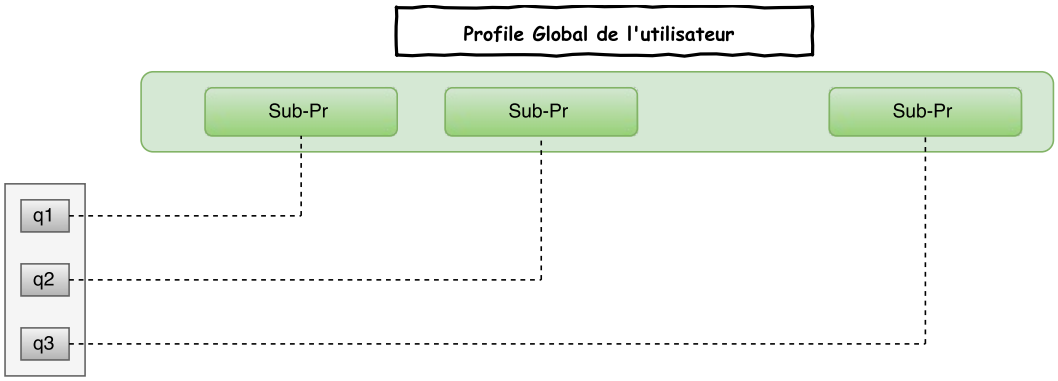



FIGURE 7.3 – Extraction des sous profils de l'utilisateur

Donc pour une requête de l'utilisateur, nous extrayons le sous-profil du profil global de l'utilisateur qui couvre le sujet de la requête. Pour extraire cette partie, nous utilisons des mesures de similarités qui estiment la similarité entre le sujet de la requête et les termes du profil de l'utilisateur. Plusieurs méthodes peuvent être utilisées. Par exemple, nous pouvons utiliser les modèles thématiques (LDA) afin de sélectionner les termes du thème de la requête du profil utilisateur, ou simplement par exemple, utiliser des fonctions de similarités permettant de sélectionner les termes qui sont proches de la requête du profil utilisateur. Nous proposons de nous baser sur la deuxième méthode, et nous suivons ce processus :

— Pour chaque requête de l'utilisateur, nous sélectionnons un ensemble de termes reliés à la requête de l'utilisateur.
— La sélection de ces termes repose sur une fonction de similarité entre les termes de la requête et le profil de l'utilisateur.

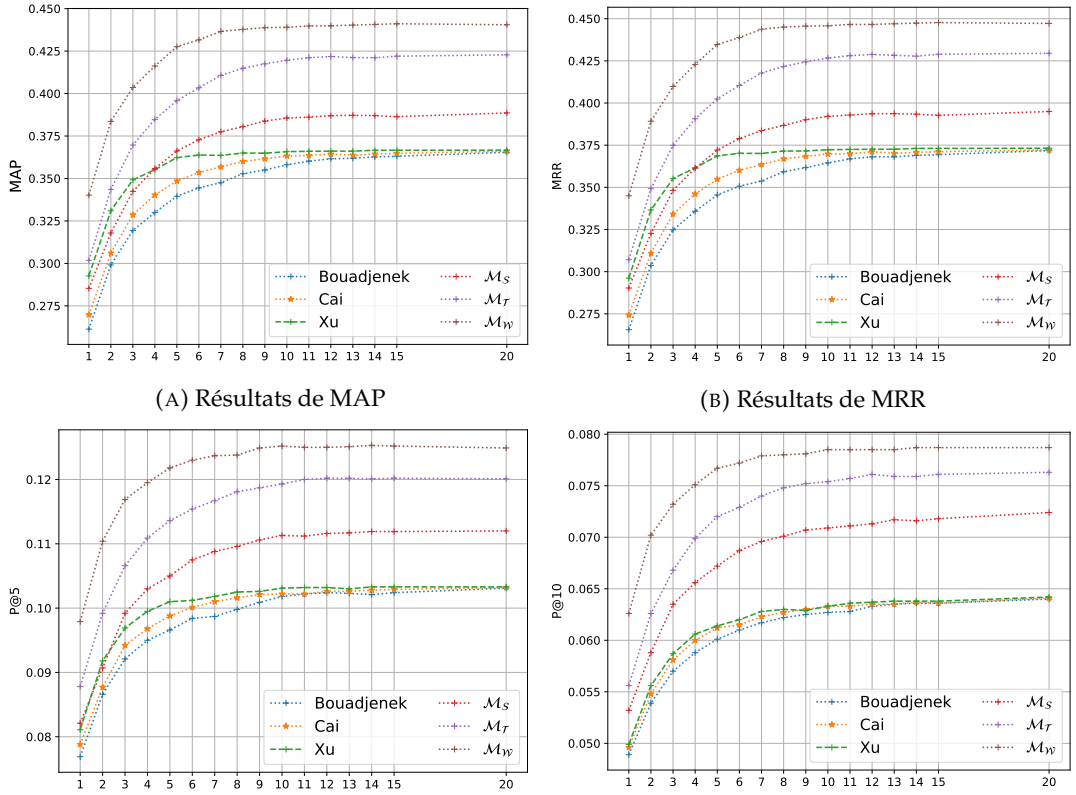
Formellement, soit le profil global $P^u_{T\text{-}X}$ de l'utilisateur $u$, et l'ensemble des requêtes de l'utilisateur $Q^u = \{q_{1_u}, q_{2_u}, ..., q_{N_u}\}$, nous construisons des profils à court termes notés $P^u_{T\text{-}X_q}$ comme suit :

— Pour chaque requête utilisateur $q_{i_u}$, nous sélectionnons un top k tags du profil de l'utilisateur $P^u_{T\text{-}X}$ en calculant la similarité entre la requête et chaque tags $tg_j \in P^u_{T\text{-}X}$ comme suit :

$$sim(q_{i_u}, tg_j) = \delta(q_{i_u}, tg_j) \tag{7.6}$$

— Nous sélectionnons un top k termes proches de la requête avec un seuil de similarité fixé à 0.65.

Nous appliquons la même méthode pour chaque modèle de l'état de l'art. Ainsi, nous évaluons tous les modèles suivant la taille du profil de l'utilisateur (nombre de tags). Les résultats obtenus sur les différentes mesures d'évaluations sont présentés dans la figure 7.4




(A) Résultats de MAP

(B) Résultats de MRR

(C) Résultats de Précisions @5

(D) Résultats de Précisions @10

FIGURE 7.4 – Évaluation suivant la taille du profil utilisateur basé sur les tags

Les résultats obtenus montrent que nos systèmes obtiennent de meilleurs résultats que les modèles de l'état de l'art et cela quelles que soient les mesures d'évaluation.

Plus en détail, les performances des tous les systèmes sont positivement corrélées avec la taille du profil (nombre de tags). Nous constatons aussi que nos modèles $MT^u_T$, $MT^u_W$ et $MT^u_U$ obtiennent les meilleurs résultats et cela avec un seul terme et les performances augmentent encore plus pour atteindre presque le double en valeur de MAP. Ceci montre l'efficacité de nos modèles et leurs capacités à identifier les meilleurs tags qui représentent l'utilisateur et leur attribuent un meilleur poids. Néanmoins, nous notons que le modèle $MT^u_S$ n'obtient les meilleurs résultats qu'à partir du terme 5 par rapport au modèle Xu.

## 7.6 Conclusion

Dans ce chapitre, nous avons présenté quatre modèles utilisateurs construits à partir des tags assignés aux documents permettant d'identifier les tags importants pour décrire les centres d'intérêts de l'utilisateur.

L'estimation des modèles repose sur une hypothèse globale qui précise que les tags d'un utilisateur doivent décrire les sujets des documents auxquels ils sont associés.

Pour valider nos propositions, nous avons conduit une série d'expérimentation qui a validé et montré les performances de nos modèles et leurs capacités à identifier les tags importants d'un document et ainsi construit des profils utilisateurs précis. Nous avons également analysé la spécificité de chaque modèle, ses apports et ses limites. Les résultats obtenus répondent donc positivement à nos questions de recherche.

Les profils des utilisateurs que nous avons proposés reposent seulement sur les tags qu'ils ont employés pour annoter les documents. Le contenu des documents n'est pris en compte que dans l'estimation des tags importants. La prise en compte du contenu des documents est importante car ils permettent d'enrichir les profils des utilisateurs basés seulement sur les tags et sont plus à même de décrire les sujets d'intérêt de l'utilisateur. Le chapitre suivant fera l'objet d'un modèle utilisateur basé sur les documents, il tirera également avantage des modèles de tags.

# Chapitre 8

# Évaluation du profil utilisateur basé sur les documents

## 8.1 Introduction

Ce chapitre est consacré à l'évaluation de notre deuxième contribution présentée au chapitre 5, dans lequel nous avons proposé une modélisation de l'utilisateur au travers des documents qu'il a annoté. Nous avons proposé 4 variantes du modèle utilisateur basé sur les documents.

En plus de la comparaison des nos modèles avec ceux de l'état de l'art, nous avons évalué nos modèles suivant les objectifs suivants :

— Étudier dans quelle mesure les tags d'un document sont capables d'estimer les termes les plus importants d'un document et ainsi proposer un profil de l'utilisateur précis où les termes qui le représentent sont les termes des documents et qui ont été estimés par l'intermédiaire des tags.
— Étudier l'impact du modèle de tags de document dans le modèle de document. Plus précisément, évaluer l'importance des tags dans la pondération des termes du document. Nous notons que cette évaluation combine les modèles de tags que nous avons présenté dans notre première contribution avec le modèle de document présentés dans notre seconde contribution.

Les évaluations de ce chapitre portent sur les éléments suivants : dans un premier lieu, nous présentons les paramètres des différents modèles de l'état de l'art et de nos modèles, ensuite nous comparons nos modèles avec les modèles de l'état de l'art.

## 8.2 Paramètres des modèles

L'objectif de cette expérimentation est d'évaluer nos modèles en comparant leurs performances à celles des modèles de l'état de l'art pour la tâche d'ordonnancement de documents personnalisé.

Les modèles que nous évaluons dans cette section intègrent les modèles de tags présentés dans le chapitre 8. Nous avons choisi de prendre les mêmes valeurs des paramètres de ces modèles de tags. Donc, nous n'avons conduit une évaluation à deux plis que pour les paramètres $\alpha$ (cf. formule 5.2) et le paramètre $\beta_c$. De plus, comme présenté précédemment, la valeur du seuil de similarité qui donne de bons résultats est égale à 0.65 et nous avons choisi de garder cette valeur dans les expérimentations conduites dans ce chapitre.

Dans ce qui suit, nous présentons en détail les modèles que nous évaluons et les valeurs de leurs paramètres :

— Le modèle $\mathbf{MC^u_B}$ où le profil de l'utilisateur est estimé avec le modèle $P^u_{C\text{-}B}$.

$$RSV(q,d,u) = \beta_c.RSV(q,d) + (1-\beta_c).RSV(d,P^u_{C\text{-}B}) \tag{8.1}$$

Où la valeur du paramètre $\beta_c$ est à 0.78. Pour la partie $RSV(d,P^u_{C\text{-}B})$, le modèle $P^u_{C\text{-}B}$ présenté dans la formule 5.10, présente deux paramètres. Le seuil de similarité entre le tag et le terme du document est fixé à la valeur de 0.65 et la valeur du paramètre $\alpha$ (cf. formule 5.2) est fixée à 0.72.

— Le modèle $\mathbf{MC^u_S}$ où le profil de l'utilisateur est estimé avec le modèle $P^u_{C\text{-}S}$.

$$RSV(q,d,u) = \beta_c RSV(q,d) + (1-\beta_c) RSV(d,P^u_{C\text{-}S}) \tag{8.2}$$

Où la valeur du paramètre $\beta_c$ est égale à 0.69. Pour la partie $RSV(d,P^u_{C\text{-}B})$, le modèle $P^u_{C\text{-}B}$ présenté dans la formule 5.13, présente deux paramètres. Le seuil de similarité entre le tag et le terme du document est fixé à la valeur de 0.65 et la valeur du paramètre $\alpha$ (cf. formule 5.2) est fixée à 0.72.

— Le modèle $\mathbf{MC^u_T}$ où le profil de l'utilisateur est défini en employant le modèle $P^u_{C\text{-}T}$.

$$RSV(q,d,u) = \beta_c RSV(q,d) + (1-\beta_c) RSV(d,P^u_{C\text{-}T}) \tag{8.3}$$

Où la valeur du paramètre $\beta_c$ est égale à 0.51. Pour la partie $RSV(d,P^u_{C\text{-}T})$, le modèle $P^u_{C\text{-}T}$ présenté dans la formule 5.16, présente trois paramètres. Le seuil de similarité entre le tag et le terme du document fixé à la valeur de 0.65, la valeur du paramètre $\alpha$ (cf. formule 5.2) est fixée à 0.49 et le nombre de thèmes pour le LDA est fixé à 400.

— Le modèle $\mathbf{MC^u_W}$ où le profil de l'utilisateur est modélisé par le modèle $P^u_{C\text{-}W}$.

$$RSV(q,d,u) = \beta_c RSV(q,d) + (1-\beta_c) RSV(d,P^u_{C\text{-}W}) \tag{8.4}$$

Où la valeur du paramètre $\beta_c$ est égale à 0.49. Pour la partie $RSV(d,P^u_{C\text{-}W})$, le modèle $P^u_{C\text{-}W}$ présenté dans la formule 5.19, présente deux paramètres. Le seuil de similarité entre le tag et le terme du document est fixé à 0.65 et la valeur du paramètre $\alpha$ (cf. formule 5.2) est fixée à 0.42.

— Nous proposons de conduire une expérimentation préliminaire pour un modèle unifié $\mathbf{MC^u_U}$, qui prend en compte les trois principaux profils $P^u_{C\text{-}S}$, $P^u_{C\text{-}T}$, et $P^u_{C\text{-}S}$. Nous avons choisi de ne pas prendre en compte le modèle $P^u_{C\text{-}B}$, car d'une part, le modèle Basique prends en compte tous les tags de l'utilisateur et d'autre part, nous voulions évaluer l'impact des modèles de tags que nous avons proposés dans notre première contribution, qui filtrent les tags des utilisateurs.
Nous n'avons pas conduit d'évaluation à deux plis pour ce modèle, et aussi nous avons gardé les mêmes valeurs des paramètres pour chaque modèle qui le compose (Les valeurs des paramètres présentés dans les points précédents). Le modèle est décrit comme suit :

$$RSV(q,d,u) = \beta_c RSV(q,d) + (1-\beta_c) RSV(d,P^u_{C\text{-}U}) \tag{8.5}$$

où $P^u_{C\text{-}U}$ représente le profil utilisateur basé sur le modèle de contenu unifié décrit par la formule suivante :

$$P(t_j, P^u_{C\text{-}U}) = \frac{\gamma_S P(t_j, P^u_{C\text{-}S}) + \gamma_T P(t_j, P^u_{C\text{-}T}) + \gamma_W P(t_j, P^u_{C\text{-}W})}{\sum_{t_k \in T^u_{d_i}} P(t_k, P^u_{C\text{-}U})} \tag{8.6}$$

— Pour les paramètres $\lambda_S$, $\lambda_T$, et $\lambda_W$, nous avons choisi d'attribuer le même apport pour chaque modèle. Les valeurs de ces paramètres sont : $\lambda_S = 0.33$, $\lambda_T = 0.33$, et $\lambda_W = 0.33$.
— Pour le paramètre $\beta_C$, nous avons choisi arbitrairement la valeur $\beta_C = 0.5$.

## 8.3 Évaluation globale des modèles

L'objectif de cette expérimentation est d'évaluer nos modèles en comparant leurs performances à celles des modèles de l'état de l'art pour la tâche d'ordonnancement de documents personnalisé.

Nous avons comparé nos modèles aux modèles de l'état de l'art et les performances sont mesurées en utilisant la MAP et le MRR, et les résultats sont présentés dans le tableau 8.1.

| Systèmes | MAP | MRR |
|---|---|---|
| Carman | 0.121 | 0.118 |
| $MC^u_B$ | 0.126 | 0.128 |
| $MC^u_S$ | 0.129 | 0.132 |
| $MC^u_T$ | 0.149 | 0.158 |
| $MC^u_W$ | **0.152** | **0.165** |
| $MC^u_U$ | 0.141 | 0.149 |

TABLE 8.1 – Évaluation globale des modèles $MC^u_B$, $MC^u_S$, $MC^u_T$, $MC^u_W$, $MC^u_U$, Carman. Les meilleurs résultats obtenus sont présentés en gras.

Les résultats de l'évaluation des modèles montrent que globalement nos modèles $MC^u_B$, $MC^u_S$, $MC^u_T$, $MC^u_W$ et $MC^u_U$ obtiennent les meilleures performances en valeur de MAP et MRR. Ces résultats démontrent que les tags peuvent être utilisés pour estimer les termes importants des documents et qui reflètent les centres d'intérêts de l'utilisateur et donc impactent positivement les résultats.

Globalement, les modèles permettent de filtrer ou d'attribuer des poids faibles pour tous les termes des documents qui ne représentent pas l'utilisateur. Les performances de chaque modèle dépend du modèle de Tag employé. Ces résultats montrent que les tags ont une influence sur la sélection des termes et donnent plus d'importance aux termes du document qui intéressent et reflètent les centres d'intérêts de l'utilisateur et donc mieux le modéliser.

Nous notons que le modèle $MC^u_W$ obtient les meilleurs de tous les modèles en valeur de MAP et de MRR. Par contre, nous remarquons que les résultats sont différents par rapport à ceux du modèle unifié des tags $MT^u_U$ présenté dans la section 8. En effet, le modèle $MC^u_U$ obtient des performances inférieures par rapport au modèle de contenu sémantique $MC^u_W$ et le modèle de contenu thématique $MC^u_T$, $MC^u_U$. Nous pensons que les raisons pour ces résultats sont dues d'une part à la nature des requêtes, et d'autre part à la prise en compte du profil global de l'utilisateur (tous ses documents) et donc ceci peut engendrer du bruit.

## 8.4 Évaluation des gains personnalisés des modèles

Dans cette partie, nous mesurons les gains personnalisés obtenus par chaque système par rapport à un autre système.

Nous mesurons les valeurs de gain personnalisé P-Gain [1] et les résultats obtenus sont présentés dans le tableau 8.2.

| Systèmes | Carman | $MC_B^u$ | $MC_S^u$ | $MC_T^u$ | $MC_W^u$ |
|---|---|---|---|---|---|
| $MC_B^u$ | 6.4% | - | - | - | - |
| $MC_S^u$ | 17.8% | 13% | - | - | - |
| $MC_T^u$ | 21% | 16.3% | 13.7% | - | - |
| $MC_W^u$ | 57% | 32% | 25% | 22.6% | - |
| $MC_U^u$ | 32% | 25.3% | 12.7% | -11.4% | - 8.6% |

TABLE 8.2 – Gains personnalisés obtenus par les modèles $MC_B^u$, $MC_S^u$, $MC_T^u$, $MC_W^u$, $MC_U^u$, Carman. Les meilleurs résultats obtenus sont présentés en gras, les valeurs (%x) : Taux d'accroissement de notre modèle par rapport au modèle

Les gains personnalisés obtenus par nos modèles sont au minimum 6,4% et au maximum 57%, ce qui montre la robustesse et la stabilité de nos modèles. Ces résultats démontrent que les tags de l'utilisateur sont un bon indicateur pour sélectionner les termes du document annotés.

Même si le modèle de Contenu Basique $MC_B^u$ ne filtre aucun tag, il obtient quand même des résultats meilleurs que le modèle de Carman. Les gains obtenus sont de 6.4% par rapport à Carman. Bien que ces résultats ne soient pas très élevés, néanmoins cela donne une bonne indication que les tags sont en mesure de sélectionner les termes importants d'un document. En effet, le poids d'un terme du document dépend de sa fréquence dans le document (normalisé par la taille du document) et de sa proximité avec le modèle de Tag (la similarité moyenne entre le terme et tous les tags du document). Et comme, le modèle de Tag $T_{B_{d_i}^u}$ ne filtre aucun tag et les prend tous en compte et si ces tags n'ont aucun lien avec le sujet du document, alors la proximité entre le modèle de Tag et le terme va être faible voir nulle et ceci va faire diminuer le poids final du terme. Ce qui explique les performances faibles.

En ce qui concerne le modèle de Contenu Standard $MC_B^u$, les résultats obtenus montrent que le modèle de Tags standard influe positivement sur la sélection des termes importants du document. En effet, nous relevons un gain de 13% et de 17.8% par rapport au modèle de Contenu basique et au modèle de Carman, respectivement. Concernant requêtes pour lesquelles le modèle n'a pas pu avoir de bons résultats, nous pensons que cela est dû au fait que le modèle de Tags élimine certains tags car ils ne sont pas présent dans le document (limite relevée du modèle de Tag standard présenté dans les expérimentations de la première contribution) et donc ces termes vont avoir un poids très faible même s'ils sont importants.

Le modèle de Contenu Thématique $MC_T^u$ obtient des taux de gains encore plus importants que les modèles $MC_B^u$ et $MC_S^u$. Les gains obtenus de 21%, 16.3% et de 13.7% par rapport au modèle de Carman, au modèle de Contenu basique $MC_B^u$, et au modèle de Contenu Standard $MC_S^u$, respectivement, démontrent encore une fois que

1. Le P-Gain est mesuré par la formule $P\text{-}Gain = \frac{\#Q_+ - \#Q_-}{\#Q_+ + \#Q_-}$

les tags arrivent à filtrer les documents et à ne garder que ceux qui sont importants et décrivent mieux les centres d'intérêts de l'utilisateur. De plus ce modèle obtient de meilleurs résultats que le modèle de contenu unifié $MC^u_U$ de 11.4%

Les résultats obtenus montrent que les gains obtenus par le modèle de Contenu Sémantique $MC^u_W$ obtiennent les meilleurs gains par rapport à tous les modèles. Effectivement, nous constatons une amélioration de 57%, de 32%, de 25%, de 22,6% et de 8.6% par rapport au modèle de Carman, au modèle de Contenu basique $MC^u_B$, au modèle de Contenu Standard $MC^u_S$, au modèle de Contenu Thématique $MC^u_S$ et au modèle de Contenu sémantique $MC^u_U$ , respectivement.

Les gains obtenu par le modèle de Contenu unifié $MC^u_U$ sont sont de 32%, 25.3%, 12.7%, -11.4%, et -8.6% par rapport au modèle de Carman, au modèle de Contenu basique $MC^u_B$, au modèle de Contenu Standard $MC^u_S$, au modèle de Contenu Thématique $MC^u_T$, et au modèle de Contenu Sémantique $MC^u_S$, respectivement.

En résumé, les résultats de ces expérimentations confirment nos hypothèses et démontrent que les tags sont un moyen prometteur et important pour sélectionner les termes importants d'un document et reflètent les centres d'intérêts de l'utilisateur.

## 8.5 Évaluation de l'impact de la taille du profil de l'utilisateur

Le profil de l'utilisateur est construit au travers des documents qu'il a annotés. Une fois qu'on a estimé le modèle de chaque document annoté, nous estimons pour chaque terme du vocabulaire le poids final qui est la moyenne de chaque poids dans le modèle de document.

Dans cette expérimentation, nous souhaitons réduire la taille des termes sélectionnés d'un document, d'une part pour avoir des modèles de documents plus précis et d'autre part réduire le vocabulaire inutiles et qui peut apporter du bruit et donc dégrader les performances des systèmes. Nous prenons les mêmes valeurs des paramètres des modèles que la section précédente.

Premièrement, pour chaque document $d^u_i$ nous estimons les modèles de contenu $MC^u_B$, $MC^u_S$, $MC^u_T$, $MC^u_W$, $MC^u_U$. Cette étape va retourner une distribution des termes du document suivant chaque modèle.

Ensuite, nous sélectionnons les 100 meilleurs termes pour chaque document pour chaque modèle, puis nous construisons le modèle de l'utilisateur en fusionnant tous les modèles de contenu des documents.

Ensuite, pour une requête utilisateur, nous extrayons le sous profil du profil global de l'utilisateur qui couvre le sujet de la requête. Pour extraire cette partie, nous procédons à la même évaluation que dans le chapitre précédent comme suit :

— Pour chaque requête de l'utilisateur, nous sélectionnons un ensemble de termes reliés à la requête de l'utilisateur.
— La sélection de ces termes repose sur une fonction de similarité entre les termes de la requête et le profil de l'utilisateur.

Ensuite, nous évaluons les modèles suivants différentes valeurs de taille du profil. Nous choisissons des profils de taille 10, 20, 40, 50, 60, 100, 200 et 300 termes. Les résultats en MAP obtenus sont présentés dans la figure 8.1

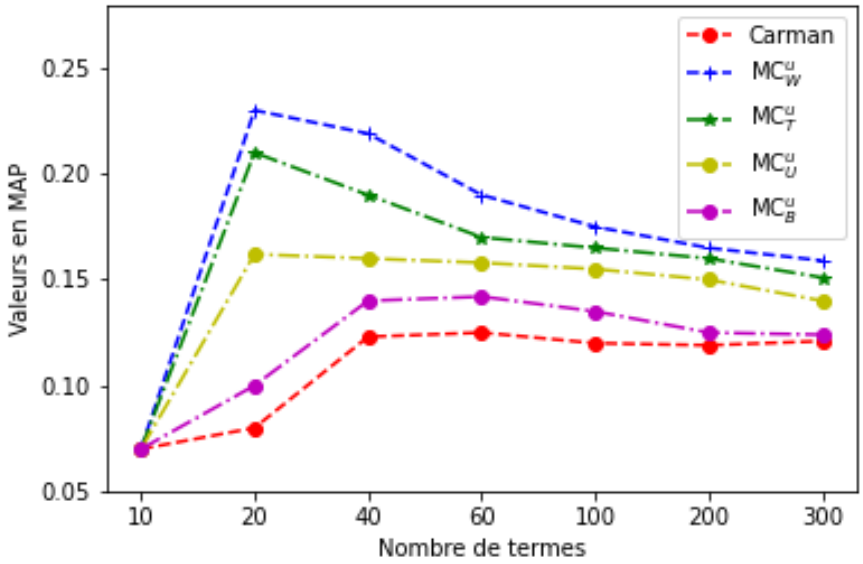



FIGURE 8.1 – Évaluation suivant la taille du profil utilisateur basé sur les documents

Les résultats obtenus montrent que globalement les modèles obtiennent les meilleures performances pour une sélection de termes entre 20 et 60 termes.

Plus précisément, les modèles $MC^u_S$, $MC^u_T$, $MC^u_W$, $MC^u_U$ obtiennent de meilleurs résultats pour un profil de taille de 20 termes. Nous remarquons que les résultats se dégradent au fur et à mesure que la taille du profil de l'utilisateur augmente. Les scores se stabilisent globalement après 200 termes.

Nous remarquons aussi que d'après les valeurs des paramètres des modèles, l'apport des modèles peut être très faible. D'ailleurs, l'apport du modèle de contenu standard est très faible par rapport au score du RSV(q,d) (comme présenté dans la formule 8.2 et $\beta_C = 0.69$). De plus, le paramètres $\alpha$ qui permet de contrôler la combinaison du score d'importance du terme dans le document et de son importance dans le modèle de Tag est très faible aussi (5.2 et $\alpha = 0.72$). Ceci démontre la faible capacité du modèle à estimer les meilleurs termes pour représenter un utilisateur.

Nous remarquons aussi, que cet apport est important pour le modèle de contenu thématique $MC^u_T$ et le modèle de contenu sémantique $MC^u_W$. D'ailleurs, pour le modèle $MC^u_T$ à presque le même apport que le score RSV(q,d) ($\beta_C = 0.51$). La valeur du paramètre $\alpha$ qui permet de contrôler la combinaison du score d'importance du terme dans le document et de son importance dans le modèle de Tag est plus importante et sa valeur est égale à 0.42. Ce résultat montre que particulièrement, que les modèles de tags ont un impact sur les modèles de contenu.

Nous notons aussi, que le modèle de contenu unifié $MC^u_U$ n'obtient pas les meilleurs résultats comme observé dans le modèle de Tag unifié $MT^u_U$. Nous soulignons, que nous avons présenté qu'une expérimentation préliminaire du modèle (les valeurs du modèle sont choisies arbitrairement).

Le modèle $MC_B^u$ et le modèle se comportent de la même manière. Leurs scores sont très proches. Les deux modèles obtiennent les meilleurs résultats pour une taille du profil entre 40 et 60 termes à la différence des autres modèles (meilleures valeurs pour un profil de taille 20).

Les deux modèles sont très différents. Le modèle de Carmen n'exploite aucun tag et se base exclusivement sur les termes des documents. Par contre, le modèle de contenu basique exploite les tags de d'utilisateur mais sans aucun filtre (tous les tags sont pris en compte). Les résultats de ces deux derniers modèles restent relativement stables et on remarque une légère dégradation au fur à mesure que la taille des profils augmente.

Les modèles $MC_S^u$, $MC_T^u$, $MC_W^u$, $MC_U^u$ exploitent les modèles de tags que nous avons présentés dans le chapitre précédent. Les limites des modèles de tags seront présentes aussi dans les modèles de contenu $MC_S^u$, $MC_T^u$, $MC_W^u$, $MC_U^u$.

Une raison des résultats des modèles de tags est due à la nature des documents : les documents apportent beaucoup d'information pour représenter les utilisateurs, la prise en compte des tous les documents peut néanmoins apporter du bruit dans le profil des utilisateurs.

Nous soulignons également la limite de la collection de test. Les requêtes sont générées automatiquement à partir des tags des utilisateurs. Et donc, nous rencontrons le même problème soulevé dans le chapitre précédent.

## 8.6 Conclusion

Dans ce chapitre, nous avons présenté un modèle de construction de profil de l'utilisateur en exploitant les tags des documents pour estimer les termes importants des documents annotés par un utilisateur.

Nos propositions sont basées sur une hypothèse globale qui préconise que pour bien représenter les centres d'intérêts et les thématiques auxquelles s'intéresse l'utilisateur, les termes des documents annotés par cet utilisateur sont un bon moyen. Les termes d'un document n'ont pas tous la même importance pour cet utilisateur. Pour intégrer cette importance personnalisée, l'intégration des tags de cet utilisateur employés pour annoter ce document sont un bon indicateur et donc nécessaires à condition qu'ils soient bien pondérés et donc, il faut ne prendre en compte que les tags qui décrivent les centres d'intérêts de l'utilisateur.

Pour évaluer nos modèles et valider nos hypothèses, nous avons réalisé une série d'expérimentations et nous avons montré que nos modèles devancent tous les modèles de l'état de l'art. Nous avons analysé et étudié la spécificité de chaque modèle proposé. Les résultats obtenus soulignent que les tags sont un bon moyen pour sélectionner les termes d'un document.

# Cinquième partie

# Conclusion

# Chapitre 9

# Conclusion et Perspectives

L'extraction, l'analyse et la représentation d'information sur les activités sociales des utilisateurs sur le web jouent un rôle important pour les systèmes de recherche d'information personnalisée et pour les systèmes de recommandation. Ainsi, il est important de créer des modèles d'utilisateurs précis et d'inférer leurs centres d'intérêts à partir de toutes les informations disponibles sur ces plateformes.

Dans cette thèse, nous avons étudié et proposé des modèles utilisateurs exploitant les informations sociales, et plus précisément, les folksonomies. Nous avons proposé, deux contributions principales pour la modélisation de l'utilisateur, spécifiquement, des approches d'estimation des poids des tags et termes pour le profil de l'utilisateur.

## Contribution 1 : Modèle de Tag - Exploitation des documents pour la pondération des tags

Cette première contribution décrite au chapitre 4, réside dans la définition d'un modèle utilisateur représenté par les tags, tel que ces tags couvrent les sujets des documents auxquels ils ont été attribués.

Notre approche se distingue par l'intégration du document dans l'estimation des poids des tags de l'utilisateur. Nous proposons de donner une définition au lien entre le document et les tags et qui est porté par notre hypothèse principale qui est *"Seuls les tags qui décrivent les sujets des documents doivent être pris en compte"*.

Nous avons présenté quatre modèles utilisateurs construits à partir des tags assignés aux documents permettant d'identifier les tags importants pour décrire les centres d'intérêts de l'utilisateur.

- **Le modèle de Tags standard** qui repose sur l'hypothèse *Seuls les tags de l'utilisateur qui sont des termes du document sont des tags pertinents et décrivent le contenu du document d'après l'utilisateur*, présenté en section 4.3
- **Le modèle de Tags thématique** qui repose sur l'hypothèse *Seuls les tags de l'utilisateur qui sont dans le même espace latent que les thèmes du document sont des tags pertinents et décrivent le contenu du document d'après l'utilisateur*, présenté en section 4.3.2
- **Le modèle de Tags sémantique** : qui repose sur l'hypothèse *Seuls les tags de l'utilisateur qui sont dans le même espace sémantique que les termes du document sont des tags pertinents et décrivent le contenu du document d'après l'utilisateur*, présenté en section 4.3.3

Pour valider nos propositions, nous avons conduit une série d'expérimentation présentée dans le chapitre 7 qui a validé et montré les performances de nos modèles

et leurs capacités à identifier les tags importants d'un document et ainsi construit des profils utilisateurs précis.

Globalement, nos modèles ont obtenu de meilleurs résultats par rapport aux modèles de l'état de l'art. Les résultats obtenus répondent donc positivement à nos questions de recherche.

Nous avons également analysé la spécificité de chaque modèle, ses apports et ses limites. Plus en détails, nous avons relevé des performances allant de 1.8% à 51%.

De plus, nous avons étudié l'impact de la taille du profil de l'utilisateur, et nous avons démontré la capacité de nos modèles à identifier les meilleurs tags qui représentent l'utilisateur.

## Contribution 2 : Modèle de Document - Exploitation des tags pour la pondération des termes du document

Dans cette contribution, nous avons présenté une nouvelle approche de modélisation de l'utilisateur basée sur les documents.

La particularité de ce modèle est de faire dépendre les termes du document non seulement du contenu textuel du document mais également des tags attribués par l'utilisateur à ce document.

Le but est de déterminer les termes importants du document qui reflètent les centres d'intérêts de l'utilisateur. Notre hypothèse principale qui est "*Seuls les termes du document qui sont liés aux tags doivent être pris en compte*"

Les termes d'un document n'ont pas tous la même importance pour un utilisateur. Pour intégrer cette importance personnalisée, l'intégration des tags de l'utilisateur employés pour annoter un document sont un bon indicateur et donc nécessaires à condition qu'ils soient bien pondérés et donc, il faut ne prendre en compte que les tags qui décrivent les centres d'intérêts de l'utilisateur. Ainsi, nous intégrons notre contribution précédente dans l'estimation du modèle du document.

Nous avons présenté quatre modèles utilisateurs construits à partir des tags assignés aux documents permettant d'identifier les tags importants pour décrire les centres d'intérêts de l'utilisateur.

- **Le modèle de contenu basique**, présenté en section 5.3.1.
- **Le modèle de contenu standard** qui exploite le modèle de Tags standard, présenté en section 5.3.2
- **Le modèle de contenu thématique** qui exploitent le modèle de Tags thématique, présenté en section 5.3.3
- **Le modèle sémantique** : qui exploitent le modèle de Tags sémantique, présenté en section 5.3.4

Pour valider nos propositions, nous avons conduit une série d'expérimentation présentée dans le chapitre 8 qui a validé et montré les performances de nos modèles et leurs capacités à identifier les termes importants d'un document et ainsi construit des profils utilisateurs précis.

Globalement, nos modèles ont obtenu de meilleurs résultats par rapport aux modèles de l'état de l'art. Les résultats obtenus répondent donc positivement à nos questions de recherche.

Nous avons également analysé la spécificité de chaque modèle, ses apports et ses limites. Plus en détails, nous avons relevé des performances allant de 6.4% à 57%.

De plus, nous avons étudié l'impact de la taille du profil de l'utilisateur, et nous avons démontré la capacité de nos modèles à identifier les meilleurs tags qui représentent l'utilisateur.

## Perspectives

Les modèles que nous avons proposés peuvent être améliorés sur plusieurs aspects.

Premièrement, les limites de nos propositions sont dues à la qualité des données de notre collection de tests.

Un grand pourcentage (37%) (statistiques sur notre collection de test) des tags des utilisateurs ne sont pas présents dans les documents, par conséquent, les trois modèles $MT^u_T$, $MT^u_W$ et $MT^u_U$ affectent négativement les résultats par la non présence des tags dans les thèmes des documents et les représentations vectorielles des termes. En effet, dans l'estimation du modèle de Tag thématique, l'apprentissage des thèmes des documents sont faits indépendamment des tags des utilisateurs.

Nous pouvons employer un modèle LDA étendu où les tags seront intégrés dans l'estimation des thèmes permettant ainsi de résoudre le problème de la non présence de certains tags dans la collection des documents.

Pour le modèle de Tags sémantique, nous proposeront un apprentissage word2vec qui permet d'intégrer les tags dans l'estimation des vecteurs des termes.

Nos propositions peuvent aussi être améliorées du point de vue conceptuel.

Le profil de l'utilisateur est une agrégation des poids de tags pour chaque document (une moyenne). Nous pensons qu'une combinaison simple des scores des trois modèles (le modèle unifié) n'est pas un choix judicieux. Donc, un modèle véritablement unifié pourrait mieux tirer avantage des spécificités de chaque modèle.

Un autre point serait d'adapter nos modèles pour des utilisateurs pauvres (dans le sens ou ils n'ont pas beaucoup de tags) afin qu'ils puissent bénéficier des informations du réseau social (amis) s'ils partagent les mêmes centres d'intérêts.

Les profils utilisateurs basés sur les documents souffrent aussi des limites des modèles de tags que nous avons décrits ci-dessus, comme ces modèles sont intégrés dans l'estimation des termes importants des documents. Ainsi, l'amélioration de ces modèles conduirait à une amélioration significative pour les profils à base de documents.

Un autre élément que nous pensons très important, est le type de la requête de l'utilisateur. En effet, notre collection de test souffre du problème des requêtes générées à partir des tags des utilisateurs. Donc, si nous intégrons une identification des types de requête et adapter le profil utilisateur suivant le type identifié. Ainsi, nous générons des profils dynamiques suivant la requête.

Une autre piste est que nous pouvons explorer les modèles de réseaux de neurones profonds qui considèrent les représentations contextuelles d'incorporation de mots, telles que BERT [143], où chaque occurrence d'un mot dans le corpus peut être représentée différemment. L'avantage de ce type de représentation est que les différentes significations d'un même mot ne sont pas considérées de la même manière.

Le système de RISP utilisé pour évaluer les profils utilisateurs que nous avons proposés, est basé sur une combinaison linéaire entre un score RSV(q,d) et un score RSV (d,u). C'est une première approche basique, simple à implémenter. Cependant, les performances d'un système de RISP ne sont pas seulement par la qualité des profils mais aussi par la qualité des modèles d'appariement et la stratégie d'intégration du profil. Nous envisageons aussi d'étudier cette piste.

# Bibliographie


[1] Robert JÄSCHKE et al. « Tag Recommendations in Folksonomies ». In : *Knowledge Discovery in Databases : PKDD 2007*. Sous la dir. de Joost N. KOK et al. Berlin, Heidelberg : Springer Berlin Heidelberg, 2007, p. 506-514. ISBN : 978-3-540-74976-9.

[2] Michael G. NOLL et Christoph MEINEL. « Web Search Personalization via Social Bookmarking and Tagging ». In : *Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference*. 2007, p. 367-380.

[3] Paul HEYMANN, Daniel RAMAGE et Hector GARCIA-MOLINA. « Social Tag Prediction ». In : *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '08. Singapore, Singapore : Association for Computing Machinery, 2008, p. 531-538. ISBN : 9781605581644. DOI : 10.1145/1390334.1390425. URL : https://doi.org/10.1145/1390334.1390425.

[4] Shenghua BAO et al. « Optimizing Web Search Using Social Annotations ». In : *Proceedings of the 16th International Conference on World Wide Web*. WWW '07. Banff, Alberta, Canada : Association for Computing Machinery, 2007, p. 501-510. ISBN : 9781595936547. DOI : 10.1145/1242572.1242640. URL : https://doi.org/10.1145/1242572.1242640.

[5] Thomas Vander WAL. « Explaining and Showing Broad and Narrow Folksonomies ». In : fév. 2005. URL : http://www.personalinfocloud.com/2005/02/explaining_and_.html.

[6] Mohamed Reda BOUADJENEK et al. « Personalized Social Query Expansion Using Social Bookmarking Systems ». In : *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '11. Beijing, China : Association for Computing Machinery, 2011, p. 1113-1114. ISBN : 9781450307574. DOI : 10.1145/2009916.2010075. URL : https://doi.org/10.1145/2009916.2010075.

[7] Mark J. CARMAN, Mark BAILLIE et Fabio CRESTANI. « Tag Data and Personalized Information Retrieval ». In : *Proceedings of the 2008 ACM Workshop on Search in Social Media*. SSM '08. Napa Valley, California, USA : Association for Computing Machinery, 2008, p. 27-34. ISBN : 9781605582580. DOI : 10.1145/1458583.1458591. URL : https://doi.org/10.1145/1458583.1458591.

[8] H.N. KIM et al. « Collaborative user modeling for enhanced content filtering in recommender systems ». In : t. 51. 4. 2011, p. 772-781. URL : http://dx.doi.org/10.1016/j.dss.2011.01.012.

[9] D. VALLET et al. « Personalized Content Retrieval in Context Using Ontological Knowledge ». In : t. 17. 3. Mar. 2007, p. 336-346. DOI : 10.1109/TCSVT.2007.890633.

[10] Shengliang XU et al. « Exploring Folksonomy for Personalized Search ». In : *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '08. Singapore, Singapore : Association for Computing Machinery, 2008, p. 155-162. ISBN : 9781605581644. DOI : 10.1145/1390334.1390363. URL : https://doi.org/10.1145/1390334.1390363.

[11] Yi CAI et al. « Exploring personalized searches using tag-based user profiles and resource profiles in folksonomy ». In : t. 58. Juin 2014. DOI : 10.1016/j.neunet.2014.05.017.

[12] Mohamed Reda Mr BOUADJENEK, Hakim HACID et Mokrane BOUZEGHOUB. « LAICOS : An Open Source Platform for Personalized Social Web Search ». In : ii. 2013, p. 1446-1449. ISBN : 978-1-4503-2174-7. DOI : 10.1145/2487575.2487705. URL : http://dl.acm.org/citation.cfm?id=2487705%7B%5C%%7D5Cnhttp://doi.acm.org/10.1145/2487575.2487705.

[13] Yi CAI et al. « Personalized Resource Search by Tag-Based User Profile and Resource Profile ». In : *WISE 2010*. 2010, p. 510-523.

[14] David VALLET, Iván CANTADOR et Joemon M. JOSE. « Personalizing Web Search with Folksonomy-Based User and Document Profiles ». In : *Advances in Information Retrieval, 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010. Proceedings*. 2010, p. 420-431.

[15] Dong ZHOU, Séamus LAWLESS et Vincent WADE. « Improving Search via Personalized Query Expansion Using Social Media ». In : t. 15. 3-4. Hingham, MA, USA : Kluwer Academic Publishers, juin 2012, p. 218-242. DOI : 10.1007/s10791-012-9191-2. URL : http://dx.doi.org/10.1007/s10791-012-9191-2.

[16] Nawal OULD-AMER, Philippe MULHEM et Mathias GÉRY. « Personalized Parsimonious Language Models for User Modeling in Social Bookmaking Systems ». In : *Advances in Information Retrieval - 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings*. 2017, p. 582-588. DOI : 10.1007/978-3-319-56608-5_52. URL : https://doi.org/10.1007/978-3-319-56608-5_52.

[17] Chahrazed BOUHINI, Mathias GÉRY et Christine LARGERON. « Integrating user's profile in the query model for Social Information Retrieval ». In : mai 2014, p. 1-2. ISBN : 978-1-4799-2393-9. DOI : 10.1109/RCIS.2014.6861091.

[18] Mohamed Reda BOUADJENEK, Hakim HACID et Mokrane BOUZEGHOUB. « Sopra : A New Social Personalized Ranking Function for Improving Web Search ». In : *ACM SIGIR '13*. 2013, p. 861-864.

[19] Kerstin BISCHOFF et al. « Can All Tags Be Used for Search ? » In : *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. CIKM '08. Napa Valley, California, USA : ACM, 2008, p. 193-202. ISBN : 978-1-59593-991-3. DOI : 10.1145/1458082.1458112. URL : http://doi.acm.org/10.1145/1458082.1458112.

[20] Francesca CARMAGNOLA et al. « Tag-based user modeling for social multi-device adaptive guides ». In : t. 18. 5. Nov. 2008, p. 497-538. DOI : 10.1007/s11257-008-9052-2. URL : https://doi.org/10.1007/s11257-008-9052-2.

[21] Mohamed Reda BOUADJENEK et al. « Evaluation of Personalized Social Ranking Functions of Information Retrieval ». In : t. 7977. Juil. 2013, p. 283-290. ISBN : 9783642391996. DOI : 10.1007/978-3-642-39200-9_24.

[22] Morgan HARVEY, Ian RUTHVEN et Mark CARMAN. « Ranking social bookmarks using topic models ». In : 2010, p. 1401. ISBN : 9781450300995. DOI : 10.1145/1871437.1871632. URL : http://portal.acm.org/citation.cfm?doid=1871437.1871632.

[23] Morgan HARVEY, Bernd LUDWIG et David ELSWEILER. « You are what you eat : Learning user tastes for rating prediction ». In : t. 8214 LNCS. 2013, p. 153-164. ISBN : 9783319024318. DOI : 10.1007/978-3-319-02432-5_19.

[24] David M. BLEI, Andrew Y. NG et Michael I. JORDAN. « Latent Dirichlet Allocation ». In : *J. Mach. Learn. Res.* 3.null (mar. 2003), p. 993-1022. ISSN : 1532-4435.

[25] Tomas MIKOLOV et al. « Efficient Estimation of Word Representations in Vector Space ». In : *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. Sous la dir. d'Yoshua BENGIO et Yann LECUN. 2013. URL : http://arxiv.org/abs/1301.3781.

[26] Nawal OULD-AMER, Philippe MULHEM et Mathias GÉRY. « Modèles de Document Parcimonieux basés sur les annotations et les word embeddings - Application à la personnalisation ». In : *COnférence en Recherche d'Informations et Applications - CORIA 2017, 14th French Information Retrieval Conference. Marseille, France, March 29-31, 2017. Proceedings, Marseille, France, March 29-31, 2017.* 2017, p. 251-264. DOI : 10.24348/coria.2017.21. URL : https://doi.org/10.24348/coria.2017.21.

[27] Nawal Ould AMER, Philippe MULHEM et Mathias GÉRY. « Toward Word Embedding for Personalized Information Retrieval ». In : *CoRR* abs/1606.06991 (2016). arXiv : 1606.06991. URL : http://arxiv.org/abs/1606.06991.

[28] Nawal Ould AMER. « Enhancing Personalized Document Ranking using Social Information ». In : *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, UMAP 2016, Halifax, NS, Canada, July 13 - 17, 2016*. Sous la dir. de Julita VASSILEVA et al. ACM, 2016, p. 345-348. DOI : 10.1145/2930238.2930374. URL : https://doi.org/10.1145/2930238.2930374.

[29] Philippe MULHEM, Nawal Ould AMER et Mathias GÉRY. « Variations axiomatiques pour la recherche d'information personnalisée ». In : *COnférence en Recherche d'Informations et Applications - CORIA 2017, 14th French Information Retrieval Conference, Marseille, France, March 29-31, 2017. Proceedings*. Sous la dir. de Jian-Yun NIE et Sylvain LAMPRIER. ARIA, 2017, p. 1-16. DOI : 10.24348/coria.2017.12. URL : https://doi.org/10.24348/coria.2017.12.

[30] Philippe MULHEM et al. « TimeLine Illustration Based on Microblogs : When Diversification Meets Metadata Re-ranking ». In : *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11-14, 2017, Proceedings*. Sous la dir. de Gareth J. F. JONES et al. T. 10456. Lecture Notes in Computer Science. Springer, 2017, p. 224-235. DOI : 10.1007/978-3-319-65813-1_22. URL : https://doi.org/10.1007/978-3-319-65813-1%5C_22.

[31] Philippe MULHEM, Nawal Ould AMER et Mathias GÉRY. « Axiomatic Term-Based Personalized Query Expansion Using Bookmarking System ». In : *Database and Expert Systems Applications - 27th International Conference, DEXA 2016, Porto, Portugal, September 5-8, 2016, Proceedings, Part II*. 2016, p. 235-243.

DOI : 10.1007/978-3-319-44406-2_17. URL : https://doi.org/10.1007/978-3-319-44406-2_17.

[32] Seyyed Hadi HASHEMI, Jaap KAMPS et Nawal Ould AMER. « Neural Endorsement Based Contextual Suggestion ». In : *Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016*. Sous la dir. d'Ellen M. VOORHEES et Angela ELLIS. T. 500-321. NIST Special Publication. National Institute of Standards et Technology (NIST), 2016. URL : http://trec.nist.gov/pubs/trec25/papers/Uamsterdam-CX.pdf.

[33] Nawal Ould AMER et al. « Word Embedding for Social Book Suggestion ». In : *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016*. Sous la dir. de Krisztian BALOG et al. T. 1609. CEUR Workshop Proceedings. CEUR-WS.org, 2016, p. 1136-1144. URL : http://ceur-ws.org/Vol-1609/16091136.pdf.

[34] Nayanika DOGRA et al. « LIG at CLEF 2016 Cultural Microblog Contextualization : TimeLine illustration based on Microblogs ». In : *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016*. Sous la dir. de Krisztian BALOG et al. T. 1609. CEUR Workshop Proceedings. CEUR-WS.org, 2016, p. 1201-1206. URL : http://ceur-ws.org/Vol-1609/16091201.pdf.

[35] Nawal Ould AMER et Mathias GÉRY. « LaHC at CLEF 2015 SBS Lab ». In : *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015*. Sous la dir. de Linda CAPPELLATO et al. T. 1391. CEUR Workshop Proceedings. CEUR-WS.org, 2015. URL : http://ceur-ws.org/Vol-1391/11-CR.pdf.

[36] Philippe MULHEM, Nawal Ould AMER et Mathias GÉRY. « LIG at CLEF 2015 SBS Lab ». In : *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015*. Sous la dir. de Linda CAPPELLATO et al. T. 1391. CEUR Workshop Proceedings. CEUR-WS.org, 2015. URL : http://ceur-ws.org/Vol-1391/6-CR.pdf.

[37] Nawal Ould AMER, Philippe MULHEM et Mathias GÉRY. « Recherche de conversations dans les réseaux sociaux : modélisation et expérimentations sur Twitter ». In : *CORIA 2015 - Conférence en Recherche d'Infomations et Applications - 12th French Information Retrieval Conference, Paris, France, March 18-20, 2015*. Sous la dir. de Éric GAUSSIER et Mathias GÉRY. ARIA, 2015, p. 55-70. DOI : 10.24348/coria.2015.6. URL : https://doi.org/10.24348/coria.2015.6.

[38] G. SALTON et M.J. MCGILL. « Introduction to Modern information Retrieval ». In : McGraw-Hill, Inc, 1986.

[39] Gerard SALTON et Christopher BUCKLEY. « Term-weighting approaches in automatic text retrieval ». In : t. 24. 5. 1988, p. 513-523. DOI : https://doi.org/10.1016/0306-4573(88)90021-0. URL : http://www.sciencedirect.com/science/inproceedings/pii/0306457388900210.

[40] Gerard SALTON. « A Comparison Between Manual and Automatic Indexing Methods ». In : USA : Cornell University, 1968.

[41] G. SALTON, A. WONG et C. S. YANG. « A Vector Space Model for Automatic Indexing ». In : t. 18. 11. New York, NY, USA : Association for Computing Machinery, nov. 1975, p. 613-620. DOI : 10.1145/361219.361220. URL : https://doi.org/10.1145/361219.361220.

[42] S. E. ROBERTSON et S. WALKER. « Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval ». In : *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '94. Dublin, Ireland : Springer-Verlag, 1994, p. 232-241. ISBN : 038719889X.

[43] Laure SOULIER. « Définition et évaluation de modèles de recherche d'information collaborative basés sur les compétences de domaine et les rôles des utilisateurs. (Definition and evaluation of collaborative ranking models based on users' domain expertise and roles) ». Thèse de doct. University of Toulouse, France, 2014. URL : https://tel.archives-ouvertes.fr/tel-01110721.

[44] Ricardo A. BAEZA-YATES et Berthier A. RIBEIRO-NETO. « Modern Information Retrieval - the concepts and technology behind search, Second edition ». In : 2011.

[45] Jay M. PONTE et W BRUCE CROFT. « A Language Modeling Approach to Information Retrieval ». In : juin 1998. DOI : 10.1145/290941.291008.

[46] C. T. YU et G. SALTON. « Precision Weighting—An Effective Automatic Indexing Method ». In : t. 23. 1. New York, NY, USA : Association for Computing Machinery, jan. 1976, p. 76-88. DOI : 10.1145/321921.321930. URL : https://doi.org/10.1145/321921.321930.

[47] Thomas HOFMANN. « Probabilistic Latent Semantic Indexing ». In : *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '99. Berkeley, California, USA : ACM, 1999, p. 50-57. ISBN : 1-58113-096-1. DOI : 10.1145/312624.312649. URL : http://doi.acm.org/10.1145/312624.312649.

[48] Thomas GRITHS et Mark STEYVERS. « A Probabilistic Approach to Semantic Representation ». In : *Paper presented at the Proceedings of the 24th annual conference of the Cognitive Science Society* (jan. 2004).

[49] Thomas L. GRIFFITHS et Mark STEYVERS. « Prediction and Semantic Association ». In : *Advances in Neural Information Processing Systems 15*. Sous la dir. de S. BECKER, S. THRUN et K. OBERMAYER. MIT Press, 2003, p. 11-18. URL : http://papers.nips.cc/paper/2153-prediction-and-semantic-association.pdf.

[50] Scott DEERWESTER et al. « Indexing by latent semantic analysis ». In : *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE* 41.6 (1990), p. 391-407.

[51] Chaitanya CHEMUDUGUNTA, Padhraic SMYTH et Mark STEYVERS. « Combining Concept Hierarchies and Statistical Topic Models ». In : *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. CIKM '08. Napa Valley, California, USA : Association for Computing Machinery, 2008, p. 1469-1470. ISBN : 9781595939913. DOI : 10.1145/1458082.1458337. URL : https://doi.org/10.1145/1458082.1458337.

[52] Ioana HULPUS et al. « Unsupervised Graph-Based Topic Labelling Using Dbpedia ». In : *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*. WSDM '13. Rome, Italy : Association for Computing Machinery, 2013, p. 465-474. ISBN : 9781450318693. DOI : 10.1145/2433396.2433454. URL : https://doi.org/10.1145/2433396.2433454.

[53] Xing WEI et W. Bruce CROFT. « LDA-based Document Models for Ad-hoc Retrieval ». In : *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '06. Seattle, Washington, USA : ACM, 2006, p. 178-185. ISBN : 1-59593-369-7. DOI : 10.1145/1148170.1148204. URL : http://doi.acm.org/10.1145/1148170.1148204.

[54] Samuel BRODY et Mirella LAPATA. « Bayesian Word Sense Induction ». In : *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. EACL '09. Athens, Greece : Association for Computational Linguistics, 2009, p. 103-111.

[55] Ivan TITOV et Ryan MCDONALD. « Modeling Online Reviews with Multi-Grain Topic Models ». In : *Proceedings of the 17th International Conference on World Wide Web*. WWW '08. Beijing, China : Association for Computing Machinery, 2008, p. 111-120. ISBN : 9781605580852. DOI : 10.1145/1367497.1367513. URL : https://doi.org/10.1145/1367497.1367513.

[56] Aria HAGHIGHI et Lucy VANDERWENDE. « Exploring Content Models for Multi-Document Summarization ». In : *Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. NAACL '09. Boulder, Colorado : Association for Computational Linguistics, 2009, p. 362-370. ISBN : 9781932432411.

[57] Thomas GRIFFITHS et Mark STEYVERS. « Finding Scientific Topics ». In : t. 101 Suppl 1. Avr. 2004, p. 5228-35. DOI : 10.1073/pnas.0307752101.

[58] Gregor HEINRICH. « Parameter estimation for text analysis ». In : 2004.

[59] Christopher D. MANNING. « Computational Linguistics and Deep Learning ». In : t. 41. 4. Cambridge, MA, USA : MIT Press, déc. 2015, p. 701-707. DOI : 10.1162/COLI_a_00239. URL : http://dx.doi.org/10.1162/COLI_a_00239.

[60] Yoshua BENGIO et al. « A Neural Probabilistic Language Model ». In : *J. Mach. Learn. Res.* 3.null (mar. 2003), p. 1137-1155. ISSN : 1532-4435.

[61] Ronan COLLOBERT et Jason WESTON. « A Unified Architecture for Natural Language Processing : Deep Neural Networks with Multitask Learning ». In : *Proceedings of the 25th International Conference on Machine Learning*. ICML '08. Helsinki, Finland : Association for Computing Machinery, 2008, p. 160-167. ISBN : 9781605582054. DOI : 10.1145/1390156.1390177. URL : https://doi.org/10.1145/1390156.1390177.

[62] Yue LU, Qiaozhu MEI et ChengXiang ZHAI. « Investigating task performance of probabilistic topic models : an empirical study of PLSA and LDA ». In : t. 14. 2. Avr. 2011, p. 178-203. DOI : 10.1007/s10791-010-9141-9. URL : https://doi.org/10.1007/s10791-010-9141-9.

[63] Xing WEI et W. Bruce CROFT. « LDA-based Document Models for Ad-hoc Retrieval ». In : *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '06. Seattle, Washington, USA : ACM, 2006, p. 178-185. ISBN : 1-59593-369-7. DOI : 10.1145/1148170.1148204. URL : http://doi.acm.org/10.1145/1148170.1148204.

[64] Mark STEYVERS et Tom GRIFFITHS. « Probabilistic Topic Models ». In : *Handbook of Latent Semantic Analysis*. Sous la dir. de T. LANDAUER et al. Lawrence Erlbaum Associates, 2007. ISBN : 1410615340. URL : http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20%5C&path=ASIN/1410615340.

[65] David ANDRZEJEWSKI et David BUTTLER. « Latent Topic Feedback for Information Retrieval ». In : *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '11. San Diego, California, USA : ACM, 2011, p. 600-608. ISBN : 978-1-4503-0813-7. DOI : 10.1145/2020408.2020503. URL : http://doi.acm.org/10.1145/2020408.2020503.

[66] Laurence A. F. PARK et Kotagiri RAMAMOHANARAO. « The Sensitivity of Latent Dirichlet Allocation for Information Retrieval ». In : *Machine Learning and Knowledge Discovery in Databases*. Sous la dir. de Wray BUNTINE et al. Berlin, Heidelberg : Springer Berlin Heidelberg, 2009, p. 176-188. ISBN : 978-3-642-04174-7.

[67] Hamed ZAMANI et W. Bruce CROFT. « Estimating Embedding Vectors for Queries ». In : *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*. ICTIR '16. Newark, Delaware, USA : ACM, 2016, p. 123-132. ISBN : 978-1-4503-4497-5. DOI : 10.1145/2970398.2970403. URL : http://doi.acm.org/10.1145/2970398.2970403.

[68] Hamed ZAMANI et W. Bruce CROFT. « Embedding-based Query Language Models ». In : *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*. ICTIR '16. Newark, Delaware, USA : ACM, 2016, p. 147-156. ISBN : 978-1-4503-4497-5. DOI : 10.1145/2970398.2970405. URL : http://doi.acm.org/10.1145/2970398.2970405.

[69] Qingyao AI et al. « Improving Language Estimation with the Paragraph Vector Model for Ad-hoc Retrieval ». In : *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '16. Pisa, Italy : ACM, 2016, p. 869-872. ISBN : 978-1-4503-4069-4. DOI : 10.1145/2911451.2914688. URL : http://doi.acm.org/10.1145/2911451.2914688.

[70] Qingyao AI et al. « Analysis of the Paragraph Vector Model for Information Retrieval ». In : *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*. ICTIR '16. Newark, Delaware, USA : ACM, 2016, p. 133-142. ISBN : 978-1-4503-4497-5. DOI : 10.1145/2970398.2970409. URL : http://doi.acm.org/10.1145/2970398.2970409.

[71] Mohannad ALMASRI, Catherine BERRUT et Jean-Pierre CHEVALLET. « A Comparison of Deep Learning Based Query Expansion with Pseudo-Relevance Feedback and Mutual Information ». In : *Advances in Information Retrieval : 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings*. Sous la dir. de Nicola FERRO et al. 2016, p. 709-715.

[72] Dion GOH et Schubert FOO. « Social information retrieval systems : Emerging technologies and applications for searching the Web effectively ». In : jan. 2007, p. 1-375. DOI : 10.4018/978-1-59904-543-6.

[73] M. BENDER et al. « Exploiting social relations for query expansion and result ranking ». In : *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on*. 2008, p. 501-506.

[74] Ralf SCHENKEL et al. « Efficient Top-k Querying over Social-tagging Networks ». In : *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '08. Singapore, Singapore, 2008, p. 523-530.

[75] Marin BERTIER et al. « Toward Personalized Query Expansion ». In : SNS '09. Nuremberg, Germany : Association for Computing Machinery, 2009, p. 7-12. ISBN : 9781605584638. DOI : 10.1145/1578002.1578004. URL : https://doi.org/10.1145/1578002.1578004.

[76] Claudio BIANCALANA, Alessandro MICARELLI et Claudio SQUARCELLA. « Nereau : A Social Approach to Query Expansion ». In : WIDM '08 (2008), p. 95-102. DOI : 10.1145/1458502.1458518. URL : http://doi.acm.org/10.1145/1458502.1458518.

[77] Andreas HOTHO et al. « BibSonomy : A Social Bookmark and Publication Sharing System ». In : *Proceedings of the Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures*. Sous la dir. d'Aldo de MOOR, Simon POLOVINA et Harry DELUGACH. Aalborg, Denmark : Aalborg University Press, juil. 2006. ISBN : 87-7307-769-0. URL : http://www.kde.cs.uni-kassel.de/pub/pdf/hotho06bibsonomy.pdf.

[78] Tsubasa TAKAHASHI et Hiroyuki KITAGAWA. « A Ranking Method for Web Search Using Social Bookmarks ». In : *Database Systems for Advanced Applications*. Sous la dir. de Xiaofang ZHOU et al. Berlin, Heidelberg : Springer Berlin Heidelberg, 2009, p. 585-589. ISBN : 978-3-642-00887-0.

[79] Hao MA. « An Experimental Study on Implicit Social Recommendation ». In : *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '13. Dublin, Ireland : Association for Computing Machinery, 2013, p. 73-82. ISBN : 9781450320344. DOI : 10.1145/2484028.2484059. URL : https://doi.org/10.1145/2484028.2484059.

[80] C. LUO et al. « Hete-CF : Social-Based Collaborative Filtering Recommendation Using Heterogeneous Relations ». In : *2014 IEEE International Conference on Data Mining*. 2014, p. 917-922. DOI : 10.1109/ICDM.2014.64.

[81] Jilin CHEN et al. « Make New Friends, but Keep the Old : Recommending People on Social Networking Sites ». In : *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '09. Boston, MA, USA : Association for Computing Machinery, 2009, p. 201-210. ISBN : 9781605582467. DOI : 10.1145/1518701.1518735. URL : https://doi.org/10.1145/1518701.1518735.

[82] P. SYMEONIDIS, A. NANOPOULOS et Y. MANOLOPOULOS. « A Unified Framework for Providing Recommendations in Social Tagging Systems Based on Ternary Semantic Analysis ». In : t. 22. 2. 2010, p. 179-192. DOI : 10.1109/TKDE.2009.85.

[83] S. GAUCH et al. « User profiles for personalized information access ». In : *The adaptive web*. Sous la dir. de B. PETER, K. ALFRED et N. WOLFGANG. Springer-Verlag, 2007, p. 54-89.

[84] Manel MEZGHANI et al. « A User Profile Modelling Using Social Annotations : A Survey ». In : *Proceedings of the 21st International Conference on World Wide Web*. WWW '12 Companion. Lyon, France : Association for Computing Machinery, 2012, p. 969-976. ISBN : 9781450312301. DOI : 10.1145/2187980.2188230. URL : https://doi.org/10.1145/2187980.2188230.

[85] Mohamed Reda BOUADJENEK, Hakim HACID et Mokrane BOUZEGHOUB. « Social networks and information retrieval, how are they converging ? A survey, a taxonomy and an analysis of social information retrieval approaches and platforms ». In : *Inf. Syst.* 56 (2016), p. 1-18.

[86] « Short and tweet : experiments on recommending content from information streams ». In : *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM*. Sous la dir. de J. CHEN et al. Atlanta, Georgia, USA, 2010, p. 1185-1194.

[87] Ceren BUDAK et al. « Inferring User Interests From Microblogs ». In : MSR-TR-2014-68. ACM CoNEXT 2017. Mai 2014. URL : https://www.microsoft.com/en-us/research/publication/inferring-user-interests-from-microblogs/.

[88] Jan VOSECKY, Kenneth Wai-Ting LEUNG et Wilfred NG. « Collaborative Personalized Twitter Search with Topic-language Models ». In : *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*. SIGIR '14. Gold Coast, Queensland, Australia : ACM, 2014, p. 53-62. ISBN : 978-1-4503-2257-7. DOI : 10.1145/2600428.2609584. URL : http://doi.acm.org/10.1145/2600428.2609584.

[89] F. ABEL et al. « Cross-system user modeling and personalization on the social web. User Modeling and User-Adapted Interaction (UMUAI), Special Issue on Personalization in Social Web ». In : t. 22. 3. 2011, p. 1-42.

[90] P. KAPANIPATHI et al. « User Interests Identification on Twitter Using a Hierarchical Knowledge Base. In : The Semantic Web : Trends and Challenges ». In : Crete, Greece : Springer, Anissaras, 2014, p. 99-113.

[91] Morgan HARVEY, Mark CARMAN et David ELSWEILER. « Comparing tweets and tags for URLs ». In : t. 7224 LNCS. 2012, p. 73-84. ISBN : 9783642289965. DOI : 10.1007/978-3-642-28997-2_7.

[92] Christoph BESEL, Jörg SCHLÖTTERER et Michael GRANITZER. « Inferring semantic interest profiles from Twitter followees : does Twitter know better than your friends ? » In : avr. 2016, p. 1152-1157. DOI : 10.1145/2851613.2851819.

[93] Wei GONG, Ee-Peng LIM et Feida ZHU. *Characterizing Silent Users in Social Media Communities*. 2015. URL : https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10462.

[94] Stefano FARALLI, Giovanni STILO et P. VELARDI. « Recommendation of microblog users based on hierarchical interest profiles ». In : *Social Network Analysis and Mining* 5 (2015), p. 1-23.

[95] Y. NECHAEV, F. CORCOGLIONITI et C. GIULIANO. « Concealing Interests of Passive Users in Social Media. In : The Re-coding Black Mirror 2017 Workshop colocated with 16th International Semantic Web Conference ». In : *International Semantic Web Conference*. ISWC, 2017, p. 605-621.

[96] G. PIAO et J.J.G. BRESLIN. « Inferring User Interests for Passive Users on Twitter by Leveraging Followee Biographies, vol 10193 LNCS ». In : Springer, Aberdeen, UK, 2017. DOI : doi:10.1007/978-3-319-56608-510.

[97] Parantapa BHATTACHARYA et al. « Inferring User Interests in the Twitter Social Network ». In : RecSys '14. Foster City, Silicon Valley, California, USA : Association for Computing Machinery, 2014, p. 357-360. ISBN : 9781450326681. DOI : 10.1145/2645710.2645765. URL : https://doi.org/10.1145/2645710.2645765.

[98] Qi GAO et al. « Interweaving Trend and User Modeling for Personalized News Recommendation ». In : *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*. WI-IAT '11. USA : IEEE Computer Society, 2011, p. 100-103. ISBN : 9780769545134. DOI : 10.1109/WI-IAT.2011.74. URL : https://doi.org/10.1109/WI-IAT.2011.74.

[99] Kalina BONTCHEVA et Dominic ROUT. « Making sense of social media streams through semantics : a survey ». In : *Semantic Web* 5.5 (2014), p. 373-403.

[100] Manish GUPTA et al. « Survey on Social Tagging Techniques ». In : t. 12. 1. New York, NY, USA : ACM, nov. 2010, p. 58-72. DOI : 10.1145/1882471.1882480. URL : http://doi.acm.org/10.1145/1882471.1882480.

[101] Fabian ABEL et al. « Leveraging the Semantics of Tweets for Adaptive Faceted Search on Twitter ». In : *Proceedings of the 10th International Conference on The Semantic Web - Volume Part I*. ISWC'11. Bonn, Germany : Springer-Verlag, 2011, p. 1-17. ISBN : 978-3-642-25072-9. URL : http://dl.acm.org/citation.cfm?id=2063016.2063018.

[102] F. ABEL et al. « Twitter-based User Modeling for News Recommendations ». In : *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, AAAI Press*. Beijing, China : IJCAI, 2013.

[103] Fabian ABEL. « Contextualization, user modeling and personalization in the social web : from social tagging via context to cross-system user modeling and personalization. » http ://d-nb.info/1014252423. Thèse de doct. University of Hanover, 2011. URL : http://edok01.tib.uni-hannover.de/edoks/e01dh11/660718537.pdf.

[104] G. PIAO et J.J.G. BRESLIN. « User modeling on twitter with wordnet synsets and dbpedia concepts for personalized recommendations ». In : *International Conference on Information and Knowledge Management, Proceedings, ACM*. T. vol. Indianapolis, Indiana, USA, 2016, p. 24-28. DOI : doi:10.1145/2983323.

[105] J. LIU et al. « What's in a name ? : an unsupervised approach to link users across communities ». In : *Proceedings of the sixth ACM International Conference on Web Search and Data Mining, ACM*. Rome, Italy, 2013, p. 495-504.

[106] F. ABEL et al. « Cross-system user modeling and personalization on the social web. User Modeling and User-Adapted ». In : 2013.

[107] Fabian ABEL et al. « Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web ». In : t. 6643. Mai 2011, p. 375-389. DOI : 10.1007/978-3-642-21064-8_26.

[108] John HANNON et al. « A Multi-Faceted User Model for Twitter ». In : *Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization*. UMAP'12. Montreal, Canada : Springer-Verlag, 2012, p. 303-309. ISBN : 9783642314537. DOI : 10.1007/978-3-642-31454-4_26. URL : https://doi.org/10.1007/978-3-642-31454-4_26.

[109] David CARMEL et al. « Personalized Social Search Based on the User's Social Network ». In : *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. CIKM '09. 2009, p. 1227-1236.

[110] J. WENG et al. « TwitterRank : Finding Topic-sensitive Influential Twitterers ». In : *Proceedings of the Third ACM International Conference on Web Search and Data Mining, ACM*. New York, NY, USA, WSDM'10, 2010, pp.

[111] Morgan HARVEY, Ian RUTHVEN et Mark J. CARMAN. « Improving Social Bookmark Search Using Personalised Latent Variable Language Models ». In : *ACM WSDM '11*. 2011, p. 485-494.

[112] Hao-Ran XIE, Qing LI et Yi CAI. « Community-Aware Resource Profiling for Personalized Search in Folksonomy ». In : t. 27. 3. Jan. 2012, p. 599-610. DOI : 10.1007/s11390-012-1247-7. URL : https://doi.org/10.1007/s11390-012-1247-7.

[113] Fei CAI, Shuaiqiang WANG et Maarten de RIJKE. « Behavior-based personalization in web search ». In : t. 68. 4. 2017, p. 855-868. DOI : 10.1002/asi.23735. URL : https://doi.org/10.1002/asi.23735.

[114] C. LU, W. LAM et Y. ZHANG. « Twitter User Modeling and Tweets Recommendation Based on Wikipedia Concept Graph. Paper presented at ». In : *the Twenty-Sixth Conference on Artificial Intelligence Workshops (AAAI)*. 2012.

[115] J. KANG et H. LEE. « Modeling User Interest in Social Media using News Media and Wikipedia ». In : t. 65. 2016, p. 52-64.

[116] G. GROSSE-BÖLTING, C. NISHIOKA et A. SCHERP. « Generic process for extracting user profiles from social media using hierarchical knowledge bases ». In : 2015. DOI : doi:10.1109/ICOSC.2015.7050806.

[117] C. NISHIOKA. « Scherp A ». In : *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, ACM*. New, 2016.

[118] C. NISHIOKA, G. GROSSE-BÖLTING et A. SCHERP. « Influence of Time on User Profiling and Recommending Researchers ». In : *Social Media. In : Proceedings of the 15th*. 2015.

[119] M. MICHELSON et S.A. MACSKASSY. « Discovering Users' Topics of Interest on Twitter : A First Look. In : Proceedings of the 4th Workshop on Analytics for Noisy Unstructured Text Data, ACM ». In : Toronto, ON, Canada, 2010, p. 73-80.

[120] F. ZARRINKALAM. « Semantics-Enabled User Interest Mining ». In : *Lecture Notes in Computer Science*. T. 9088. In : Gandon, 2015, p. 817-828.

[121] F. ZARRINKALAM et al. « Inferring Implicit Topical Interests on Twitter ». In : *European Conference on Information Retrieval*. Padua, Italy : Springer, 2016, p. 479-491.

[122] S. FARALLI, G. STILO et P. VELARDI. « Automatic acquisition of a taxonomy of microblogs users' interests. Web Semantics : Science, Services and Agents on the World Wide Web ». In : 2017. DOI : doi:https://doi.org/10.1016/j.websem.2017.05.004.

[123] Y. NECHAEV, F. CORCOGLIONITI et C. GIULIANO. « Concealing Interests of Passive Users in Social Media. In : The Re-coding Black Mirror 2017 Workshop colocated with 16th International Semantic Web Conference ». In : *International Semantic Web Conference*. ISWC, 2017, p. 605-621.

[124] T. FLATI et al. « Two is bigger (and better) than one : the Wikipedia bitaxonomy project ». In : *52nd Annual Meeting of the Association for Computational Linguistics*. ACL, Association for Computational, 2014.

[125] A. RITTER, S. CLARK et O. ETZIONI. « Named Entity Recognition in Tweets : an Experimental Study ». In : *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*. Edinburgh, United Kingdom, 2011, p. 1524-1534.

[126] C. NISHIOKA. « Scherp A ». In : *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, ACM*. New, 2016.

[127] Claudio BIANCALANA et al. « Social Semantic Query Expansion ». In : *ACM Trans. Intell. Syst. Technol.* 4.4 (oct. 2013). ISSN : 2157-6904. DOI : 10.1145/2508037.2508041. URL : https://doi.org/10.1145/2508037.2508041.

[128] J. WANG et D. SOERGEL. « A User Study of Relevance Judgments for e-Discovery ». In : American Society for Information Science, 2010, p. 74-1.

[129] Dominik BENZ et al. « The Social Bookmark and Publication Management System Bibsonomy ». In : t. 19. 6. Secaucus, NJ, USA : Springer-Verlag New York, Inc., déc. 2010, p. 849-875. DOI : 10.1007/s00778-010-0208-4. URL : http://dx.doi.org/10.1007/s00778-010-0208-4.

[130] Beate KRAUSE, Andreas HOTHO et Gerd STUMME. « A Comparison of Social Bookmarking with Traditional Search ». In : *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval*. ECIR'08. Glasgow, UK : Springer-Verlag, 2008, p. 101-113. ISBN : 3-540-78645-7, 978-3-540-78645-0. URL : http://dl.acm.org/citation.cfm?id=1793274.1793290.

[131] Seyyed Hadi HASHEMI, Jaap KAMPS et Nawal Ould AMER. « Neural Endorsement Based Contextual Suggestion ». In : *Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016*. 2016. URL : http://trec.nist.gov/pubs/trec25/papers/Uamsterdam-CX.pdf.

[132] Marijn KOOLEN et al. « Overview of the CLEF 2015 Social Book Search Lab ». In : *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Sous la dir. de Josanne MOTHE et al. Cham : Springer International Publishing, 2015, p. 545-564. ISBN : 978-3-319-24027-5.

[133] Maryam KARIMZADEHGAN et ChengXiang ZHAI. « Estimation of Statistical Translation Models Based on Mutual Information for Ad Hoc Information Retrieval ». In : *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '10. Geneva, Switzerland : ACM, 2010, p. 323-330. ISBN : 978-1-4503-0153-4. DOI : 10.1145/1835449.1835505. URL : http://doi.acm.org/10.1145/1835449.1835505.

[134] Maryam KARIMZADEHGAN et ChengXiang ZHAI. « Axiomatic Analysis of Translation Language Model for Information Retrieval ». In : *Advances in Information Retrieval*. Sous la dir. de Ricardo BAEZA-YATES et al. Berlin, Heidelberg : Springer Berlin Heidelberg, 2012, p. 268-280. ISBN : 978-3-642-28997-2.

[135] Navid REKABSAZ et al. « Generalizing Translation Models in the Probabilistic Relevance Framework ». In : *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. CIKM '16. Indianapolis, Indiana, USA : ACM, 2016, p. 711-720. ISBN : 978-1-4503-4073-1. DOI : 10.1145/2983323.2983833. URL : http://doi.acm.org/10.1145/2983323.2983833.

[136] Omer LEVY, Yoav GOLDBERG et Ido DAGAN. « Improving Distributional Similarity with Lessons Learned from Word Embeddings ». In : t. 3. 2015, p. 211-225. URL : https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/inproceedings/view/570.

[137] Andrew Kachites MCCALLUM. « MALLET : A Machine Learning for Language Toolkit ». In : http://mallet.cs.umass.edu. 2002.

[138] Morgan HARVEY, Claudia HAUFF et David ELSWEILER. « Learning by Example : Training Users with High-quality Query Suggestions ». In : 2015, p. 133-142. ISBN : 9781450336215. DOI : 10.1145/2766462.2767731.

[139] Navid REKABSAZ, Mihai LUPU et Allan HANBURY. « Exploration of a Threshold for Similarity Based on Uncertainty in Word Embedding ». In : avr. 2017, p. 396-409. ISBN : 978-3-319-56607-8. DOI : 10.1007/978-3-319-56608-5_31.

[140] Alexander KOTOV et al. « Modeling and Analysis of Cross-session Search Tasks ». In : *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '11. Beijing, China : ACM, 2011, p. 5-14. ISBN : 978-1-4503-0757-4. DOI : 10.1145/2009916.2009922. URL : http://doi.acm.org/10.1145/2009916.2009922.

[141] Paul N. BENNETT et al. « Modeling the Impact of Short- and Long-term Behavior on Search Personalization ». In : *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '12. Portland, Oregon, USA : ACM, 2012, p. 185-194. ISBN : 978-1-4503-1472-5. DOI : 10.1145/2348283.2348312. URL : http://doi.acm.org/10.1145/2348283.2348312.

[142] Karthik RAMAN, Paul N. BENNETT et Kevyn COLLINS-THOMPSON. « Toward Whole-session Relevance : Exploring Intrinsic Diversity in Web Search ». In : *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '13. Dublin, Ireland : ACM, 2013, p. 463-472. ISBN : 978-1-4503-2034-4. DOI : 10.1145/2484028.2484089. URL : http://doi.acm.org/10.1145/2484028.2484089.

[143] Jacob DEVLIN et al. « BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding ». In : *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota : Association for Computational Linguistics, juin 2019, p. 4171-4186. DOI : 10.18653/v1/N19-1423. URL : https://www.aclweb.org/anthology/N19-1423.