Recherche d'information sociale : exploitation des signaux sociaux pour améliorer la recherche d'information

Ismail Badache

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *05/02/2016* **par :**

Ismail BADACHE

Recherche d'Information Sociale
Exploitation des Signaux Sociaux pour Améliorer la Recherche d'Information

---

JURY

Pr. Lynda TAMINE	Université Paul Sabatier	Présidente du Jury
Pr. Gabriella PASI	Universitá di Milano Bicocca	Rapporteur
Pr. Philippe MULHEM	Laboratoire d'Informatique de Grenoble	Rapporteur
Pr. Patrick GALLINARI	Université Pierre et Marie Curie	Examinateur
Pr. Mohand BOUGHANEM	Université Paul Sabatier	Directeur de thèse

---

École doctorale et spécialité :
MITT : Image, Information, Hypermedia
Unité de Recherche :
Institut de Recherche en Informatique de Toulouse (UMR 5505)
Directeur(s) de Thèse :
Mohand BOUGHANEM
Rapporteurs :
Gabriella PASI et *Philippe MULHEM*

Recherche d'Information Sociale

Exploitation Des Signaux Sociaux Pour Améliorer La Recherche d'Information

Ismail Badache

Février 2016

Manuscrit soumis pour le diplôme de Docteur ès Sciences
Doctorant : Ismail Badache
Directeur de thèse : Professeur Mohand Boughanem
IRIT - Université de Toulouse III Paul Sabatier

Commentaires, corrections, et autres remarques sont les bienvenus à :
Ismail.Badache@irit.fr
Institut de Recherche en Informatique de Toulouse, UMR 5505 CNRS,
Université Toulouse III Paul Sabatier,
118 route de Narbonne,
F-31062 Toulouse CEDEX 9

Proverbe Amazigh : Almmud ar timtelt.

Traduction : L'apprentissage doit durer jusqu'au tombeau.

Ohana signifie famille.
Famille signifie que personne ne doit être abandonné, ni oublié.

— Lilo & Stitch

Du plus profond de mon cœur, je dédie ce travail, À mes parents *Hamid* et *Houria* pour lesquels j'exprime mon amour et ma gratitude pour leur sacrifice, leur soutien moral et leurs prières. Ils n'ont eu de cesse de m'encourager et de m'offrir des conditions confortables durant la période de mes études. Qu'Allah leur préserve une agréable vie.

À mon frère *Youcef* et mes sœurs *Sara* et *Yasmine* qu'ils trouvent dans ce travail l'expression de ma reconnaissance en leur souhaitant un avenir plein de succès et de bonheur.

À toute personne qui m'a soutenu moralement durant la réalisation de ce mémoire, En témoignage de ma fidélité et mon attachement en leur souhaitant toute la joie et le bonheur du monde...

« Le meilleur homme parmi vous est le meilleur envers sa famille. Et je suis véritablement le meilleur envers ma famille. »

We have seen that computer programming is an art, because it applies accumulated knowledge to the world, because it requires skill and ingenuity, and especially because it produces objects of beauty.

— Donald E. Knuth [116]

REMERCIEMENT

Il me sera très difficile de remercier tout le monde car c'est grâce à l'aide de nombreuses personnes que j'ai pu mener cette thèse à son terme.

Je voudrais tout d'abord remercier grandement et exprimer ma très profonde gratitude à mon directeur de thèse, *Mohand Boughanem* pour la confiance qu'il m'a accordée en acceptant de diriger ma thèse, alors que j'ignorais tout de la Recherche d'Information. Je suis ravi et fier d'avoir travaillé en sa compagnie car outre son appui scientifique, il a toujours été là pour me soutenir et me conseiller au cours de l'élaboration de cette thèse. J'ai appris à ses côtés la rigueur scientifique et la pédagogie pour présenter et rédiger les travaux pendant ces années de thèse. Je ne le remercierai jamais assez et je lui serai toujours reconnaissant.

Je souhaite également remercier les rapporteurs de ma thèse, Professeur *Philippe Mulhem* (Laboratoire d'Informatique de Grenoble) et Professeur *Gabriella Pasi* (Université de Milan), pour avoir accordé du temps à une lecture attentive et détaillée de mon manuscrit ainsi que pour leurs remarques encourageantes et constructives. Je tiens également à remercier les examinateurs, Professeur *Lynda Tamine* (Université Paul Sabatier), Professeur *Patrick Gallinari* (Université Pierre et Marie Curie) d'avoir accepté de participer à mon jury de thèse et échangé leurs points de vue.

Je remercie toute l'équipe IRIS et la direction du laboratoire IRIT de m'avoir accueilli chaleureusement durant cette thèse. Des remerciements tout particuliers à *Guillaume Cabanac*, *Max Chevallier*, *Marc Boyer* et *Vincent Dugat* de m'avoir accordé leur confiance lors des enseignements que j'ai réalisés. Je ne peux pas oublier *Chantal Morand*, *Jean-Pierre Baritaud* et tout le personnel de l'IRIT et de l'école doctorale pour leur aide durant ces années au laboratoire. Merci également à l'ensemble des doctorants, les échanges professionnels ou moins professionnels permettent d'égayer la vie d'un doctorant.

Special thanks to *Melanie* for accepting to collaborate during CLEF 2015 competition (Social Book Search) which we have been ranked first *SBS Winners*. Hope to colaborate with you for other work!

La gaieté de mes amis et mes collègues, leur présence, leur soutien, en particulier *M'hamed*, *Darine*, *Mahdi* et *Adel*. Un remerciement spécial pour *Amine Aminus* qui a pris le temps de relire ma thèse. Je remercie également : *Makram, Baby, Adel Z, Rabah, Messaoud, Farouk, Mojdeh, ChaChou, Rahim, Ahmed, Mohamed H & Mohamed E, Hamid, Imen, Chiraz, Sabrina, Rafik A & Rafik B, Thibaut, Amir, Liana, Laure, Diep, Eya, Amjad, Ghada, Ameni, Meriem, Manel, Firas, Faiza, Thomas, Djamel, Baptiste, Ophélie, Bilel...*, et tous ceux qui ont fait partie de ma vie estudiantine.

Plus important que tous, *Ma Famille*, rien de tout cela aurait été possible sans leur amour, leur soutien et leur patience. *Maman* et *Papa* qui ont été ma force durant toutes ces années : je voudrais particulièrement et chaleureusement vous dire *MERCI* avec amour éternel. Je remercie également ma sœur *Yasmine* pour les nombreuses heures qu'elle a consacrées pour la relecture de mes articles. Des remerciements éternels et parfumés au Jasmin paradisiaque à Madame *F. Amiar* et mon merveilleux Tonton *Hocine*.

ABSTRACT

Our work is in the context of social information retrieval (SIR) and focuses on the exploitation of user-generated content in the process of seeking information. The User-generated content, or UGC, refers to a set of data (eg. social signals) whose content is mainly produced either directly influenced by end users. It is opposed to the traditional content produced, sold or distributed by professionals. The term became popular since 2005, in the Web 2.0 environments and in new social media. This movement reflects the democratization of the means of production and interaction in the web thanks to new technologies. Among these means more and more accessible to a wide public, we can cite social networks, blogs, microblogs, wikis, etc.

The majority of information retrieval (IR) systems exploit two classes of features to rank documents in response to user's query. The first class, the most used one, is query-dependent, which includes features corresponding to particular statistics of query terms such as term frequency, and term distribution within a document or in the collection of documents. The second class, referred to as documents prior, corresponds to query-independent features such as the number of incoming links to a document, PageRank, topical locality, presence of URL, document authors, etc.

One of the important sources which can also be used to measure the a priori interest of Web resources is social data (signals) associated with Web resource resulting from user interaction with this resource. These interactions representing annotations, comments or votes, produce useful and interesting social information that characterizes a resource in terms of popularity and reputation. Major search engines integrate social signals (e.g. Google, Bing). Searchmetrics1 showed that it exists a high correlation between social signals and the rankings provided by search engines such Google. We propose an approach that exploits social signals generated by users on the resources to estimate a priori relevance of a resource. This a priori knowledge is combined with topical relevance modeled by a language modeling (LM) approach.

We also hypothesize that signals are time-dependent, the date when the user action has happened is important to distinguish between recent and old signals. Therefore, we assume that the recency of signals may indicate some recent interests to the resource, which may improve the a priori relevance of document. Secondly, number of signals of a resource depends on the resource age. Generally, an old resource may have much more signals than a recent one. We introduce the time-aware social approach that incorporates temporal characteristics of users' actions as prior in the retrieval model. Precisely, instead of assuming uniform document priors in this retrieval model, we assign document priors based on the signals associated to that document biased by both the creation date of the signals and the age of the document.

Keywords: Information retrieval, Social Networks, User generated content, Social Signals, Social properties, Time, Diversity.

RÉSUMÉ

Notre travail se situe dans le contexte de recherche d'information sociale (RIS) et s'intéresse plus particulièrement à l'exploitation du contenu généré par les utilisateurs dans le processus de la recherche d'information. Le contenu généré par les utilisateurs (en anglais User-generated content, ou UGC) se réfère à un ensemble de données (ex. signaux sociaux) dont le contenu est principalement, soit produit, soit directement influencé par les utilisateurs finaux. Il est opposé au contenu traditionnel produit, vendu ou diffusé par les professionnels. Le terme devient populaire depuis l'année 2005, dans les milieux du Web 2.0, ainsi que dans les nouveaux médias sociaux. Ce mouvement reflète la démocratisation des moyens de production et d'interaction dans le Web grâce aux nouvelles technologies. Parmi ces moyens de plus en plus accessibles à un large public, on peut citer les réseaux sociaux, les blogs, les microblogs, les Wikis, etc.

Les systèmes de recherche d'information exploitent dans leur majorité deux classes de sources d'évidence pour trier les documents répondant à une requête. La première, la plus exploitée, est dépendante de la requête, elle concerne toutes les caractéristiques relatives à la distribution des termes de la requête dans le document et dans la collection (tf-idf). La seconde classe concerne des facteurs indépendants de la requête, elle mesure une sorte de qualité ou d'importance a priori du document. Parmi ces facteurs, on en distingue le PageRank, la localité thématique du document, la présence d'URL dans le document, ses auteurs, etc.

Une des sources importantes que l'on peut également exploiter pour mesurer l'intérêt d'une page Web ou de manière générale une ressource, est le Web social. En effet, grâce aux outils proposés par le Web 2.0 les utilisateurs interagissent de plus en plus entre eux et/ou avec les ressources. Ces interactions (signaux sociaux), traduites par des annotations, des commentaires ou des votes associés aux ressources, peuvent être considérés comme une information additionnelle qui peut jouer un rôle pour mesurer une importance a priori de la ressource en termes de popularité et de réputation, indépendamment de la requête.

Nous supposons également que l'impact d'un signal social dépend aussi du temps, c'est-à-dire la date à laquelle l'action de l'utilisateur est réalisée. Nous considérons que les signaux récents devraient avoir un impact supérieur vis-à-vis des signaux anciens dans le calcul de l'importance d'une ressource. La récence des signaux peut indiquer certains intérêts récents à la ressource. Ensuite, nous considérons que le nombre de signaux d'une ressource doit être pris en compte au regard de l'âge (date de publication) de cette ressource. En général, une ressource ancienne en termes de durée d'existence a de fortes chances d'avoir beaucoup plus de signaux qu'une ressource récente. Ceci conduit donc à pénaliser les ressources récentes vis-à-vis de celles qui sont anciennes. Enfin, nous proposons également de prendre en compte la diversité des signaux sociaux au sein d'une ressource.

Mots clés : Recherche d'information, Réseaux sociaux, Contenu généré par l'utilisateur, Signaux sociaux, Propriétés sociales, Temps, Diversité.

PUBLICATIONS

Nos idées et contributions ont déjà paru dans les publications scientifiques suivantes :

Articles de conférences internationales

1. Mélanie Imhof, Ismail Badache, Mohand Boughanem. Multimodal Social Book Search. Dans : Conference on Multilingual and Multimodal Information Access Evaluation (CLEF), Toulouse, France, CEUR Workshop Proceedings, 9 pages, septembre 2015.

2. Ismail Badache, Mohand Boughanem. A Priori Relevance Based On Quality and Diversity Of Social Signals. Dans : ACM SIGIR Special Interest Group on Information Retrieval (SIGIR), Santiago, Chile, ACM, p. 731-734, août 2015.

3. Ismail Badache, Mohand Boughanem. Document Priors Based On Time-Sensitive Social Signals. Dans : European Conference on Information Retrieval (ECIR), Vienna, Austria, Springer-Verlag, p. 617-622, mars 2015.

4. Ismail Badache, Mohand Boughanem. Social Priors to Estimate Relevance of a Resource. Dans : ACM Information Interaction in context, Regensburg (IIiX), Germany, ACM, p. 106-114, août 2014.

5. Ismail Badache, Mohand Boughanem. Harnessing Social Signals to Enhance a Search. Dans : IEEE/WIC/ACM International Conference on Web Intelligence (WIC), Warsaw, Poland, Vol. 1, IEEE Computer Society, p. 303-309, août 2014.

Articles de conférences nationales

1. Ismail Badache, Mohand Boughanem. Pertinence a Priori Basée sur la Diversité et la Temporalité des Signaux Sociaux. Dans : Conférence en Recherche d'Information et Applications (CORIA), Paris, France, ARIA, p. 23-38, mars 2015.

2. Ismail Badache, Mohand Boughanem. Exploitation de signaux sociaux pour estimer la pertinence a priori d'une ressource. Dans : Conférence en Recherche d'Information et Applications (CORIA), Nancy, France, ARIA, p. 163-178, mars 2014.

3. Ismail Badache. RI sociale: intégration de propriétés sociales dans un modèle de recherche. Dans : Conférence en Recherche d'Information et Applications (CORIA), Neuchâtel, Suisse, ARIA, p. 305-310, avril 2013.

TABLE DES MATIÈRES

TABLE DES FIGURES

TABLE DES TABLEAUX

ACRONYMES

API Application Programming Interface

BA Bayesian Average

CLEF Conference and Labs of the Evaluation Forum

IDF Inverse Document Frequency

IMDb Internet Movies Database

LSI Latent Semantic Indexing

LT Library Thing

ML Modèle de Lanque

MAP Mean Average Precision

nDCG normalized Discounted Cumulative Gain

RI Recherche d'Information

RIS Recherche d'Information Sociale

SSR Social SimRank

SPR Social PageRank

SBS Social Book Search

TF Term Frequency

TREC Text REtrieval Conference

UGC User Generated Content

URL Uniform Resource Locator

XML eXtensible Markup Language

Partie I

INTRODUCTION

Cookie : Anciennement petit gâteau sucré, qu'on acceptait avec plaisir. Aujourd'hui, petit fichier informatique drôlement salé, qu'il faut refuser avec véhémence.

— Luc Fayard

1 INTRODUCTION

1.1 Web social et recherche d'information sociale

Le WWW (World Wide Web), créé au début des années 1990 initialement composé de pages HTML[1] statiques reliées entre elles par des hyperliens, a changé de façon spectaculaire vers un modèle plus collaboratif, dans lequel tous les utilisateurs peuvent être à la fois producteurs et consommateurs de l'information. Le Web social, Web 2.0, a complètement changé la façon dont les personnes communiquent et partagent des informations. Il permet, en effet, aux utilisateurs d'interagir, de produire et de partager des masses importantes de contenus sociaux grâce à une multitude d'outils sociaux (ex. Wikis, réseaux sociaux, blogs, etc).

Ceci a conduit à l'émergence des contenus sociaux générés par les utilisateurs dans les services sociaux sur Internet. Ces contenus sociaux sont généralement éphémères, subjectifs, évolutifs et de nature différente : des *annotations* sociales, des *clics*, des *tweets*, des *commentaires*, des *relations* sociales, des *actions* relevant d'activités sociales telles que le *j'aime*, le *partage*, le *+1*, le *rating*, etc. En outre, l'information sociale peut être caractérisée par plusieurs propriétés tacites (implicites) quantifiables telles que la popularité (l'intérêt que suscite une ressource), la confiance, la réputation d'une ressource, l'engagement des utilisateurs à travers leurs actions sociales. La gestion et l'exploitation de ces contenus dans le domaine de la RI, a conduit à l'émergence de ce que l'on nomme la recherche d'information sociale (RIS). La RIS se trouve au carrefour de la RI et des réseaux sociaux. Elle est appréhendée selon deux axes :

1. le premier porte sur la définition des approches et de modèles de RI spécifiques pour rechercher de nouveaux types de contenus. De même ces contenus ont conduit à l'émergence de nouveaux besoins. Par exemple, la recherche dans les sites des médias sociaux [67], trouver des informations sociales qui répondent à un besoin spécifique [146], poser des questions à des utilisateurs [59], la recherche d'opinion [154], etc.

2. le second concerne l'exploitation des contenus sociaux pour améliorer la RI. Les réseaux sociaux et les contenus générés par l'utilisateur (UGC) pourraient être intégrés au sein du processus de recherche en tant que source d'information additionnelle pour améliorer la pertinence des résultats. Par exemple, les requêtes des utilisateurs peuvent être étendues en utilisant les Wikis [120], les annotations sociales [86]. De nouvelles pages Web publiés pourraient être détectées instantanément grâce à des blogs et à des flux de microblogging [174, 63]. Les données telles que les clics, les tags, les signaux sociaux peuvent être utilisées pour classer des ressources Web [104, 206, 13, 37, 44].

1 https://fr.wikipedia.org/wiki/Hypertext_Markup_Language

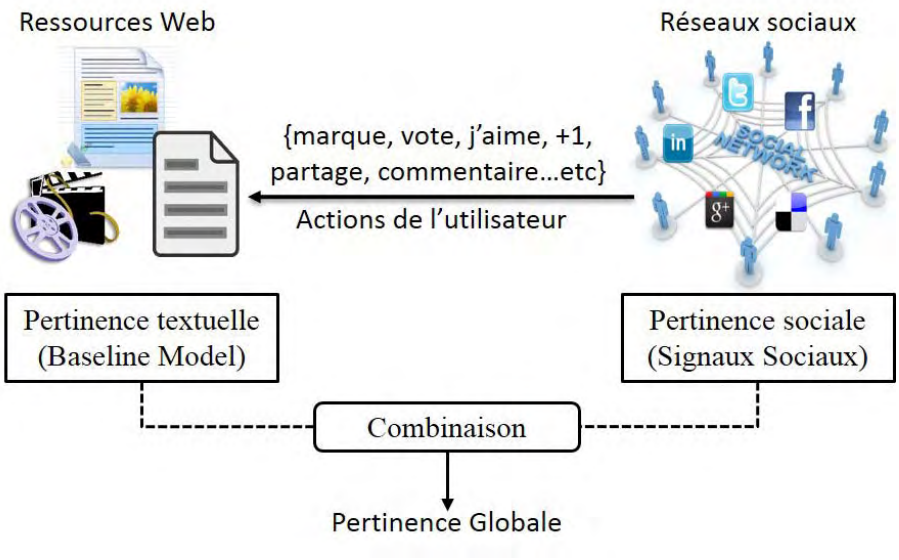

Figure 1.1: Utilisation des signaux sociaux pour améliorer la RI

La motivation derrière l'exploitation de ces contenus, en particulier les signaux (ex. +1, j'aime, etc), sur la performance des systèmes de recherche d'information (SRI) est d'essayer de tirer profit de ces traces provenant des actions collectives des utilisateurs "Wisdom of Crowds" pour améliorer la RI par rapport à un besoin en information. Le concept de "Wisdom of Crowds", présenté par *Surowiecki* [190], se réfère à l'intelligence collective élaborée par les utilisateurs d'Internet qui collaborent pour commenter, tagger ou noter des ressources Web (documents) par l'intermédiaire des Wikis, des blogs et des réseaux sociaux. Comme le montre la figure 1.1, ces interactions sont utiles pour accéder aux ressources Web les plus attractives socialement. Elles peuvent être exploitées à différents niveaux, à savoir au niveau de l'utilisateur (profilage) pour mieux comprendre ses besoins, ou bien du coté de la ressource pour mieux la décrire et mesurer une certaine importance a priori.

Nos travaux se situent dans la seconde classe, l'exploitation des signaux sociaux pour améliorer la RI.

1.2 Défis et enjeux de la recherche d'information sociale

Ces masses de données générées par les utilisateurs réactualisent les problématiques de la recherche d'information (RI) à tous les niveaux, en particulier, au niveau de la définition des modèles de pertinence pour intégrer ces contenus, et en termes d'approches pour l'exploitation efficace de ces contenus dans des tâches de RI. En effet, les modèles de recherche d'information classiques permettent d'évaluer l'intérêt, vue comme la pertinence, d'une ressource en s'appuyant sur des propriétés généralement extraites de son contenu (mots clés pour les pages Web, ou annotation pour les images). Dans notre cas, la problématique principale porte sur la manière de transformer ces contenus hétérogènes en propriétés permettant de les intégrer dans des modèles d'évaluation de pertinence. Cette problématique s'accentue sur d'autres facteurs tels que :

- **Volume** : l'émergence du Web social a conduit à la disponibilité d'énorme quantité de données générées par l'utilisateur. De toute évidence, ces données sociales peuvent améliorer l'efficacité des systèmes de recherche d'information. Cependant, cela demande des études rigoureuses. En effet, les systèmes de RI devraient être en mesure de traiter cette quantité de données et de le rendre utilisable et exploitable. Le défi concerne l'aspect technologique de traitement de l'information (indexation et recherche) ainsi que les aspects conceptuels et méthodologiques. La question porte sur le stockage, l'accès et l'analyse à grande échelle de ces quantités massives d'informations sociales (*Big Data*) [220, 58].

- **Structure des réseaux sociaux** : chaque réseau social propose une structure propre à son réseau qui le différencie de ses concurrents. Par exemple, les associations d'amitié sur Facebook relient des *amis* de façon bidirectionnelle. Twitter propose des relations unidirectionnelles appelées *Followers*. Google+ adopte cependant une autre approche où les liens sociaux sont classés dans des *cercles* de confiance (ex. famille, collègues, amis et connaissances). En outre, le réseau social peut impliquer différents types d'entités en fonction des activités au sein du réseau social. Dans les réseaux Wiki, deux types d'entités sont impliqués : les auteurs et les articles. Les réseaux de bookmarking social impliquent plusieurs entités, y compris les utilisateurs, les documents et les tags. Cette diversité des structures de réseaux sociaux apporte des difficultés supplémentaires.

- **Acteurs sociaux** : l'évaluation des acteurs sociaux consiste à identifier les utilisateurs influents dans le réseau social. La pertinence sociale d'un acteur dépend cependant de la structure du réseau social. Par exemple, les acteurs importants dans les Wikis sont définis comme des experts caractérisés par des contributions précieuses sur certains sujets et qui ont reçu moins de critiques. Dans le cas des réseaux de partage de médias, la pertinence sociale est assimilée à la popularité de l'utilisateur. A côté de ces deux propriétés, la pertinence sociale peut être définie par l'autorité, la confiance et l'influence des personnes sur le réseau social.

1.3 Questions de recherche

Cette thèse porte sur le problème de la définition de la pertinence en exploitant les signaux sociaux, en particulier l'évaluation de l'importance sociale d'une ressource. Les questions de recherche auxquelles nous avons répondu durant notre thèse sont les suivantes :

1. Est-ce que les signaux sociaux peuvent être des critères de pertinence ?
2. Comment traduire les signaux sociaux en propriétés sociales ?
3. Quelles sont les propriétés sociales utiles pour évaluer la pertinence a priori d'une ressource ?
4. Comment prendre en compte les signaux sociaux et leur temporalité pour estimer l'importance d'une ressource ?
5. Est-ce que la diversité des signaux sociaux impacte les résultats ?

6. Quel modèle théorique pour combiner la pertinence a priori d'une ressource et sa pertinence thématique ?

7. Est-ce que la qualité du signal est influencée par son réseau social ?

1.4 Contributions

Nos travaux visent à améliorer la qualité des résultats de recherche d'information adhoc en exploitant les signaux sociaux. La tâche adhoc consiste à restituer des documents pertinents vis-à-vis d'un besoin d'information exprimé sous forme de mots-clés formant la requête. Nos contributions peuvent être résumées comme suit :

1. **Exploitation individuelle et groupée des signaux sociaux**. Les signaux sociaux associés aux ressources Web peuvent être considérés comme une information additionnelle qui peut jouer un rôle pour mesurer une importance a priori d'une ressource indépendamment de la requête. Nous démontrons comment ces signaux issus de plusieurs réseaux sociaux, qui sont sous forme d'actions relevant d'activités sociales telles que le nombre de *j'aime* et de *partage*, peuvent être combinées (groupées) pour quantifier des propriétés sociales telles que la *popularité* et la *réputation* et peuvent être utiles pour améliorer les références, en termes de pertinence, d'un système de RI. Plus précisément, nous avons tout d'abord montré qu'il y a une corrélation entre la présence des signaux sociaux sur une ressource (document recherché) et sa pertinence a priori. Nous avons ensuite présenté une approche basé sur un modèle de langue, permettant la combinaison de ces signaux, modélisés comme une probabilité d'importance a priori d'une ressource, et la pertinence thématique. Les résultats montrent que la prise en compte des signaux de manière individuelle et groupés améliore les résultats de recherche. Outre, le modèle de langue, nous avons exploité ces signaux dans une approche supervisée en utilisant différentes techniques d'apprentissage.

2. **Temporalité des signaux sociaux**. La première contribution confirme que plus ces signaux sont fréquents sur une ressource plus son importance a priori croît. Cependant, dans les travaux existants les signaux sociaux sont pris en compte indépendamment du moment où l'action (le signal) s'est produite et de la date de publication de la ressource. Ils sont pris en compte uniquement par rapport à leur fréquence dans la ressource. Nous avons étudié l'impact de la temporalité des signaux sur la performance d'un système de RI. Nous avons deux hypothèses. Dans la première, nous considérons que les ressources associées aux signaux frais (récents) devraient être favorisées par rapport à celles qui sont associées à des signaux anciens. Nous proposons de compter les occurrences d'un signal en les pondérant (en les *boostant*) avec sa date d'apparition. Dans la seconde, nous pensons que la date de publication d'une ressource joue un rôle important sur la vie sociale de cette ressource dans les réseaux sociaux. Une vieille ressource a une plus grande chance d'avoir un grand nombre d'interactions par rapport à une ressource publiée récemment. Donc, pour limiter l'impact de l'ancienneté de la ressource, nous proposons de normaliser la distribution des signaux sociaux associés à une ressource par la date de publication de la ressource (âge de la ressource).

Nous avons montré que la prise en compte des facteurs temporels, date de publication de la ressource et du signal, améliore les résultats de recherche.

3. **Qualité et diversité des signaux sociaux**. Nous avons également étudié l'impact de la diversité des signaux au sein d'une ressource sur la performance d'un système de RI. Le terme "diversité" est souvent utilisé dans les approches de l'état de l'art comme diversité liée au contenu textuel restitué vis-à-vis d'un besoin en information [8, 163]. A notre connaissance, nos travaux sont les premiers à avoir introduit cette notion de diversité de signaux en RI. La diversité que nous voulons étudier, désigne la variété et la variabilité des signaux sociaux sous toutes leurs formes au sein du même document. Nous supposons que la multiplicité de ces traces ou signaux sociaux générés par les utilisateurs fait l'objet d'une sorte de témoignage de plusieurs communautés sur la qualité d'un document ou une ressource Web. Cette diversité provenant de multiples sources de créativité interactive avec une ressource serait un atout pour mesurer l'importance sociale de cette ressource. Donc, nous pensons que la diversité peut être considérée comme un facteur de pertinence sociale, qui contribuerait à l'amélioration de la recherche d'information. Les résultats montrent que la diversité améliore la pertinence. Nous avons analysé la qualité de chaque signal par rapport à son réseau social. En effet, afin de mieux comprendre l'effet de ces signaux sur le processus de sélection des documents pertinents, nous avons analysé leur distribution par rapport à la pertinence dans les différents documents renvoyés par l'ensemble des requêtes.

Afin d'évaluer l'apport de nos différentes contributions, nous nous sommes basés sur les deux corpus fournis par la campagne d'évaluation CLEF (Conference and Labs of the Evaluation Forum) : SBS 2015 (Social Book Search) et IMDb 2011 (Internet Movies Database), contenant respectivement 2.8 millions et 167438 documents ainsi que leurs données sociales collectées à partir de plusieurs réseaux sociaux.

1.5 Organisation de la thèse

Cette thèse est structuré en 7 chapitres suivants :

- Le *Chapitre 1* introduit un aperçu du contexte de nos travaux. Les questions de recherche et les principales contributions sont également présentées dans cette section.
- Le *Chapitre 2* présente les concepts généraux de la recherche d'information, ainsi que les questions relatives à l'indexation et les modèles de RI. Enfin, un aperçu sur le protocole et les métriques d'évaluation en RI.
- Le *Chapitre 3* présente la recherche d'information sociale. Nous décrivons d'abord l'information sociale dans le Web. Ensuite, la notion de la RI sociale sera définie. Nous présentons une vue d'ensemble des principales tâches de la RI sociale, ainsi qu'un aperçu sur les travaux liés à l'exploitation des informations sociales dans le processus de RI, en mettant l'accent sur les approches liées à nos contributions. Nous présentons également les principales collections standards utilisées en RI

sociale. Enfin, nous analysons les limites de l'état de l'art en positionnant nos contributions.

- Le *Chapitre 4* présente notre première contribution qui concerne l'exploitation des signaux sociaux pris en compte individuellement et groupés sous forme de propriétés sociales (*popularité* et *réputation*. Ces facteurs sont modélisés comme une probabilité d'importance a priori de la ressource. Ensuite, une étude par les techniques d'apprentissage sera présentée.

- Le *Chapitre 5* s'intéresse particulièrement à la temporalité associée à ces signaux. Nous supposons que l'importance a priori d'un document dépend non seulement de la qualité de ces signaux mais aussi de la date de leur création ainsi que la date de publication de la ressource. De ce fait, plutôt que d'estimer cette importance a priori par un simple comptage des signaux dans le document, nous intégrons également la date de publication de la ressource, pour ne pas pénaliser les nouvelles ressources, et les dates des signaux pour privilégier les signaux récents.

- Le *Chapitre 6* : est consacré à l'évaluation de l'impact de la diversité des signaux sociaux au sein d'un document. Nous analysons également leur distribution dans les différents documents renvoyés par l'ensemble des requêtes.

- Le *Chapitre 7* conclut cette thèse, nous dressons le bilan et la synthèse de nos travaux. Nous introduisons ensuite les limites et les perspectives.

Nos expérimentations sont menées sur deux types de collection, IMDb (Internet Movies Database) et SBS (Social Book Search), contenant respectivement 167438 et 2.8 millions documents ainsi que leurs données sociales collectées à partir de plusieurs réseaux sociaux.

Partie II

SYNTHÈSE DES TRAVAUX DE L'ÉTAT DE L'ART

The best place to hide a dead body is page 2 of Google search results.

— Unknown

2

RECHERCHE D'INFORMATION TEXTUELLE

Introduction

Les origines de la recherche d'information (RI) peuvent revenir à l'époque de la seconde Guerre mondiale où des quantités massives de la documentation et des rapports sur les armes ont été produites [50]. A cette époque, l'indexation des documents était déjà une tâche lourde. L'ampleur de cette tâche a été décrite dans la célèbre publication de Vannevar Bush au sujet de la *memex* (memory extender) [39]. Cette réalité n'a pas changé depuis, mais elle est devenue une tâche encore plus complexe. La croissance d'Internet et le WWW (World Wide Web) a généré d'énormes volumes d'informations. Ces informations sont juste à quelques clics de souris, mais l'accès à ces informations constitue une demande croissante pour créer des outils d'aide à la recherche d'information pour satisfaire les besoins des utilisateurs [122]. Le système qui fournit cette aide est généralement connu sous le nom *moteur de recherche*. Le terme moteur de recherche est considéré comme un synonyme de système de RI, basé sur un algorithme bien défini.

Ce chapitre présente la recherche d'information, qui est nécessaire pour une exploration plus approfondie dans les chapitres suivants. Premièrement, la RI sera définie historiquement. Deuxièmement, les concepts de traitement de l'information seront présentés dans le but de fournir une compréhension sur le processus de la recherche d'information. Troisièmement, les principaux modèles classiques de recherche seront introduits. Enfin, un aperçu sur le protocole et les métriques d'évaluation en RI est proposé.

2.1 Définition

La recherche d'information (RI), l'un des premiers domaines de recherche en informatique, a proposé les premières solutions automatiques pour le stockage de texte et leur recherche [129, 38]. L'une des premières définition pour les systèmes de RI est proposée par *Salton* en 1968 :

> *"Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information."* [176]

En 1980, *Rijsbergen* donne cette définition :

> *"Information retrieval systems address the representation, organization of, and access to large amounts of heterogeneous information encoded in digital format."* [168]

Une définition très générale par *Robertson* en 1981 considère un système de RI comme un système qui mène

> *"the user to those documents that will best enable him/her to satisfy his/her need for information."* [169]

En 1983, *Salton* a proposé une nouvelle définition et a défini le système de RI comme suit :

> *"An information retrieval system is an information system, that is, a system used to store items of information that need to be processed, searched, retrieved, and disseminated to various user populations."* [73]

En 1984, *Belkin* donne une définition similaire avec un cadrage plus spécifique du besoin d'information. Sa définition décrit les systèmes de RI au regard du problème de gestion :

> *"The goal of an information [retrieval] system is for the user to obtain information from the knowledge resource which helps her/him in problem management."* [19]

Kowalski propose une définition plus spécifique et plus détaillée, qui définit également les informations qui sont considérées dans les systèmes de RI :

> *"An Information Retrieval System is a system that is capable of storage, retrieval, and maintenance of information. Information in this context can be composed of text (including numeric and date data), images, audio, video and other multi-media objects."* [122]

2.2 Concepts et processus de RI

Tout processus de recherche d'information est construit autour de 3 fonctions : l'indexation, le requêtage (recherche) et l'appariement. Ces étapes sont plus ou moins complexes en fonction de la tâche de recherche. Ce processus qui permet, à partir d'une requête, d'ordonnancer les documents est illustré par la figure 2.1. Une architecture typique du système de recherche d'information est également présentée selon *Baeza-Yates* et *Ribeiro-Neto* [167] sur la figure 2.1.

Nous détaillons les principales étapes dans un processus de recherche d'informations dans ce qui suit.

2.2.1 *Indexation*

Cette étape consiste à identifier pour chaque document les termes importants, puis à exploiter ces termes comme index pour accéder rapidement aux documents. Un des objectifs de l'indexation est donc de permettre de retrouver rapidement les documents contenant les termes (mots-clés) de la requête. L'indexation se déroule hors ligne et il n'est pas nécessaire de l'effectuer plus d'une fois, à moins que la collection de documents soit modifiée. L'indexation peut être : manuelle, semi-automatique ou automatique.

1. **Indexation manuelle** : chaque document est analysé par un expert du domaine, qui identifie les mots clés appelés descripteurs. L'indexation manuelle fournit une terminologie pour indexer et rechercher des documents, assurant ainsi

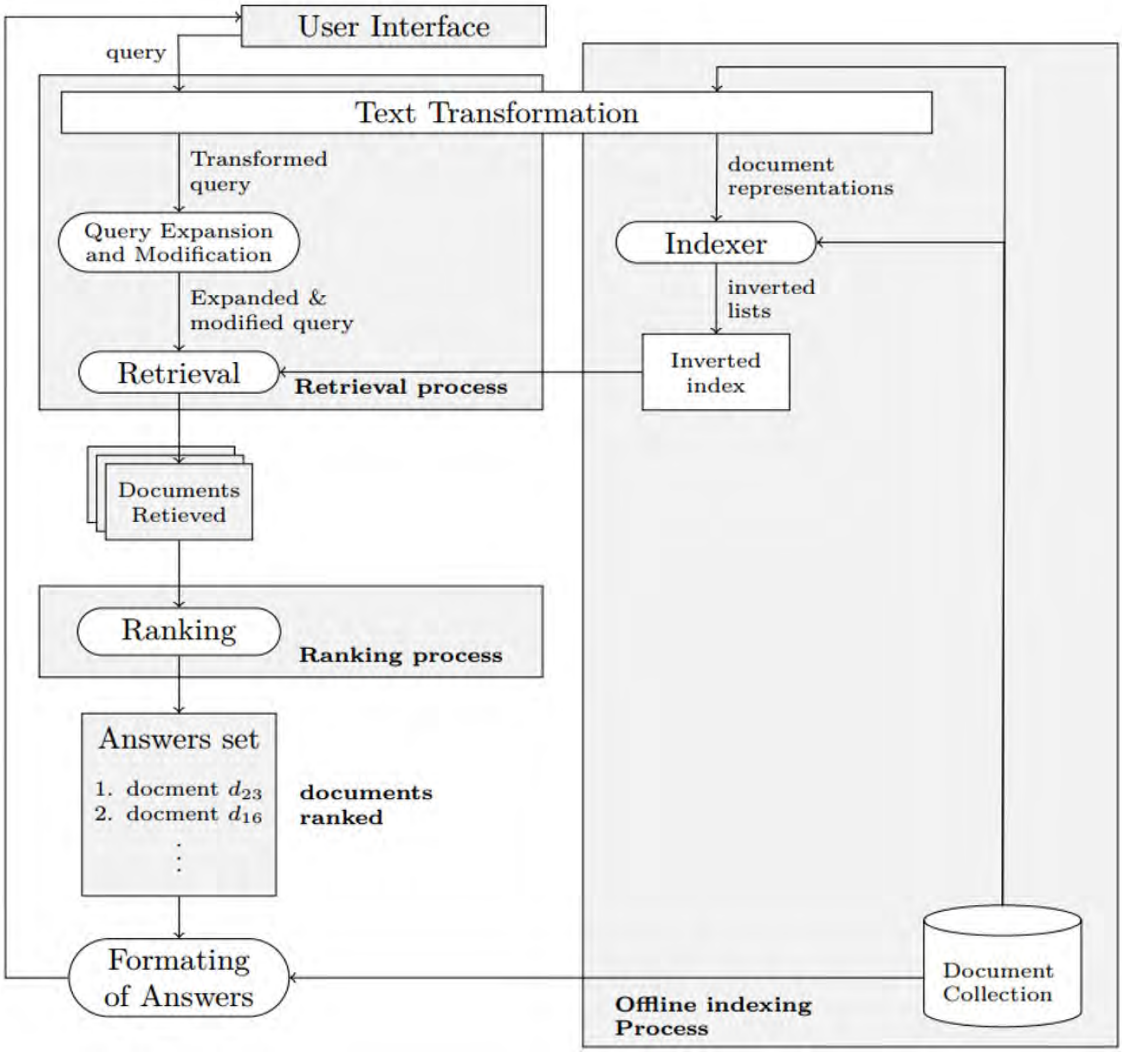

Figure 2.1: Système de recherche d'information selon *Baeza-Yates* et *Ribeiro-Neto* [167]

une meilleure qualité des résultats retournés par le système de RI. Cependant, l'indexation manuelle présente un effort trop coûteux en temps et en besoin humain. De plus, un degré de subjectivité lié au facteur humain fait que le même document peut être indexé de différentes façons par des personnes différentes, et même par la même personne mais à des moments différents [72].

2. **Indexation automatique** : fait appel aux robots d'indexation, ce qui rend le processus d'indexation complètement automatisé. L'indexation automatique, basée essentiellement sur une approche statistique, est adoptée par la majorité des systèmes de RI en raison de son coût réduit par rapport à l'indexation manuelle [129, 140, 176].

3. **Indexation semi-automatique** : est basée sur un processus automatique. En outre, elle fait appel à une intervention humaine (par un expert) pour effectuer la sélection finale des mots clés significatifs et établir les relations sémantiques entre les mots clés en se basant sur un thésaurus ou une base terminologique qui est une liste organisée de descripteurs (mots-clés) en suivant des règles bien définies.

Généralement, l'indexation comprend une série de traitements automatisés. Ils sont appliqués sur les documents et aussi sur les requêtes. On distingue : l'extraction des mots (segmentation), l'élimination des mots vides, la normalisation et la pondération. L'indexation repose sur plusieurs méthodes de transformation de texte et de normalisation sont appliquées aux documents. Premièrement, le texte du document est divisé en *tokens*, ce qui équivaut à des mots. Les mots grammaticaux, appelés mots vides, sont alors enlevés. Un processus de lemmatisation et racinisation [162, 158] est effectué pour transformer les mots ayant un sens similaire dans une forme de base commune. Ensuite, les documents et les termes sont représentés en utilisant une structure de données commune appelée "index inversé" [73, 117, 221]. Cette structure assure un accès rapide à la collecte de documents en mettant en correspondance les termes à l'ensemble des documents où ils apparaissent.

2.2.1.1 *Extraction des mots*

Cette phase consiste à extraire/segmenter le texte du document en mots. La segmentation (*tokenization*) du texte est une première étape importante dans ce processus, elle n'est appliquée aussi bien qu'au texte du document qu'à celui de la requête [52]. Dans de nombreux cas, les "tokens" sont les mêmes que les mots ou les termes. Les "tokens" peuvent être des chaînes de caractères qui sont séparées par des espaces. Cela ne nous dit pas, cependant, la façon de traiter des caractères spéciaux tels que les traits d'union et les ponctuations. Devrions-nous traiter "apple" la même façon que "Apple" ? Est-ce-que "en-ligne" représente deux mots ou un seul mot ? Est-ce-que l'apostrophe dans "O'Connor" doit être traitée de la même que celle de "l'Internet" ? Dans certaines langues, la tokenization devient encore plus complexe [52]. La langue chinoise, par exemple, n'a pas de séparateur de mots clair comme un espace en français. Donc, une analyse lexicale est nécessaire pour identifier les "tokens" en reconnaissant tout ce qui est des séparateurs, des caractères spéciaux, des chiffres, les ponctuations, etc.

2.2.1.2 *Élimination des mots vides*

Les textes contiennent souvent des termes non significatifs appelés mots vides (pronoms personnels, prépositions, etc). Ce traitement a pour but d'enlever les mots grammaticaux, ainsi que rejeter les mots dépassant un certain nombre d'occurrences dans la collection. La suppression de ces termes peut réduire de manière considérable la taille de l'index. Selon le modèle de recherche d'information utilisé, la suppression de ces mots a généralement peu d'impact sur l'efficacité du moteur de recherche, et peut même l'améliorer [52]. Malgré ces avantages potentiels, il peut être difficile de décider du nombre de mot à inclure dans la liste des mots vides. Certaines listes de mots vides utilisées dans la recherche contiennent des centaines de mots. L'élimination de ces mots peut poser des problèmes, par exemple, il devient impossible de rechercher avec des requêtes contenant des entités nommées ou des expressions avec des prépositions, ou encore en éliminant le mot "a" de « vitamine a ». Pour éviter cela, les systèmes de recherche utilisent de très petites listes de mots vides lors du traitement de texte d'un document, mais ensuite elles utilisent des listes plus longues pour le traitement par défaut du texte de la requête [52].

2.2.1.3 *Normalisation*

Cette phase est liée à la lemmatisation (ou racinisation), il s'agit d'un traitement morphologique des mots permettant de regrouper les variantes d'un mot. En effet, dans un texte, il peut y avoir différentes formes d'un mot désignant le même sens. Le but de ce processus est de les représentés par un seul mot qui porte un concept commun (ex. biologie, biologiste, biologique ? par : biologie). Grâce à la lemmatisation, les documents contenant différentes formes d'un même mot auront les mêmes chances d'être restitués. Par conséquent, elle réduit la taille de l'index et améliore le rappel, mais elle peut réduire la précision. Par exemple, l'ensemble des mots « operate operating operates operation operative operatives operational » va devenir « oper » en appliquant Porter [161]), ce qui implique une perte de précision pour des requêtes telles que « operational and research », « operating and system » et « operative and dentistry ».

Nous distinguons quatre principaux types de lemmatisation : a) par analyse grammaticale en utilisant un dictionnaire (ex : Tree-tagger[1] [182]); b) par utilisation de règles de transformation de type condition-action principalement pour l'anglais (ex : l'algorithme de Porter [161]); c) par troncature des suffixes à X caractères (ex. la troncature à 7 caractères); d) ou encore par la méthode des n-grammes utilisée souvent pour la langue chinoise [141].

2.2.1.4 *Pondération des mots*

Cette étape vient après l'identification des termes des documents et leur normalisation. Les termes qui représentent un document n'ont pas la même importance. Donc, la pondération est une phase primordiale puisqu'elle traduit l'importance des termes en indices qui reflètent le poids relatif des mots dans les documents. Estimer l'importance d'un terme n'est pas une tâche facile. Prenons l'exemple d'une collection contenant un million de documents. Un terme qui existe dans tous les documents n'est pas utile dans l'index car il ne peut fournir aucune information discriminante sur le document qui pourrait intéresser un utilisateur. Tandis qu'un terme qui apparaît dans peu de documents peut être de grande valeur vu qu'il permet de discriminer les documents pertinents. De même, un terme fréquent dans un document peut être important dans le document.

De manière générale, les fonctions de pondération des termes, telle que la mesure $TFIDF$, font intervenir deux facteurs : fréquence du terme t dans le document d, notée TF (*Term Frequency*, et la fréquence inverse du document, notée IDF (*Inverse Document Frequency*). La formule du $TFIDF$ est donnée par le produit des deux fonction TF et IDF comme suit :

$$TFIDF_{t_d} = TF_{t,d} \cdot IDF_t \tag{2.1}$$

Le poids est souvent calculé durant le processus de recherche, sachant que certains types de poids exigent des informations liées à la requête, mais la grande partie de calcul se fait pendant le processus d'indexation.

Les facteurs TF et IDF sont définis comme suit :

1. TF **(Term Frequency)** : ce facteur prend en compte le nombre d'occurrence d'un terme dans un document. L'idée derrière cette mesure est que plus un terme est

1 http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

fréquent dans un document plus il est important. Elle représente une pondération locale d'un terme dans un document. On trouve plusieurs variantes de cette mesure. Soit le document d_j et le terme t_i, alors la fréquence TF_{ij} du terme dans le document peut être donnée selon l'une des formulations suivantes :

$$TF_{ij} = 1 + log(td_{ij}) \tag{2.2}$$

Ou bien :

$$TF_{ij} = \frac{td_{ij}}{\sum_k td_{kj}} \tag{2.3}$$

Avec td_{ij} représente le nombre d'occurrences du terme t_i dans le document d_j. Le dénominateur est la somme des occurrences de tous les termes dans le document d_j. La dernière déclinaison permet de normaliser la fréquence du terme pour éviter les biais liés à la longueur du document.

L'inconvénient du TF se situe au niveau de la pertinence globale. Certains termes sont plus significatifs que d'autres, bien qu'apparaissant avec la même fréquence dans un document. Par exemple, dans une collection de documents traitant de la compétition Roland Garros, le terme *Nadal* est plus important que le terme *tennis*, même si ces deux termes apparaissent équitablement dans un document. Pour cette raison le TF est souvent couplé avec la mesure IDF.

2. IDF **(Inverse Document Frequency)** : ce facteur mesure la fréquence d'un terme dans toute la collection, c'est la pondération globale. En effet, un terme fréquent dans la collection, a moins d'importance qu'un terme moins fréquent. Cette mesure est exprimée selon l'une des déclinaisons suivantes :

$$IDF_i = log\left(\frac{N}{n_i}\right) \tag{2.4}$$

Ou bien :

$$IDF_i = log\left(\frac{N}{n_i + 1}\right) \tag{2.5}$$

Avec N représente le nombre de documents de la collection et n_i est le nombre de documents dans lesquels le terme t_i apparaît.

Un autre facteur qui est considéré dans la pondération est la taille du document. Son but est de normaliser les fréquences en fonction de la taille des documents [185, 171].

2.2.2 *Requêtage*

La recherche vise à sélectionner les documents pertinents qui couvrent les besoins d'information de l'utilisateur. Cette phase dépend de la représentation du document, les besoins d'information de l'utilisateur et les préférences de l'utilisateur (par exemple, la langue, la date, le format, etc). Cette étape s'intéresse à l'expression des besoins

de l'utilisateur, souvent à travers une liste de mots-clés représentant la requête [20]. Ainsi, la requête soumise par l'utilisateur subit les mêmes traitements que ceux réalisés précédemment sur les documents au cours de leur indexation. Toutefois, la requête peut être étendue ou reformuler pour renforcer les préférences des utilisateurs et le retour de pertinence [83, 170]. À la fin du processus de recherche, une liste de documents sera retournée.

2.2.3 *Appariement*

Une fois les documents indexés et la requête analysée, le système de RI procède à la mesure de pertinence de chaque document vis-à-vis du besoin d'information (requête) selon une fonction de correspondance relative au modèle de recherche, et à renvoyer ensuite à l'utilisateur une liste de résultats. Cette mise en correspondance génère un score de pertinence reflétant le degré de similarité entre la requête et le document. Ce score est calculé à partir d'une valeur appelée $RSV(q,d)$ (Retrieval Status Value), où q représente une requête et d un document. Cette mesure de pertinence système, que l'on essaye de rapprocher le plus possible du jugement de pertinence de l'utilisateur vis-à-vis du document, prend en compte les poids des termes calculés au moment de l'indexation. Le score final permet d'ordonner les documents retournés. Certains de ces documents parmi les résultats retournés par le système de RI peuvent potentiellement satisfaire les besoins de l'utilisateur. Ces documents sont appelés documents pertinents. Un système parfait ne doit retourner que des documents pertinents, en rejetant les non-pertinents. Cependant, les systèmes de recherche d'information parfaits n'existent pas. Aujourd'hui, les systèmes retournent généralement une liste classée de documents, dans lesquels les documents en tête de liste sont ceux qui sont les plus susceptibles d'intéresser les utilisateurs, ou les plus susceptibles d'être pertinents. En effet, l'utilisateur consulte généralement les premiers documents renvoyés (les 10 ou 20 premiers).

Différents algorithmes de RI ont été proposés dans la littérature dans le but de formaliser la pertinence. Dans la suite, nous présentons les principaux modèles de l'état de l'art.

2.3 Modèles de RI

Après que le terme RI ait été inventé en 1950 [144], un grand nombre d'algorithmes de recherche ont été proposés. Ces algorithmes de recherche peuvent être classés selon au moins un des modèles suivants : les modèles booléens [107, 177], les modèles vectoriels [179], les modèles probabilistes [171, 172], les modèles de langue [160] et l'apprentissage d'ordonnancement en RI [128].

Un modèle de recherche d'information est un fondement théorique qui représente les documents et les requêtes, et définit une stratégie d'ordonnancement des documents retournés vis-à-vis de la requête. En outre, un modèle de recherche d'information est modélisé avec un quadruplé $[D, Q, F, R(q_i, d_j)]$ [167] où :

- D est un ensemble de documents;
- Q est un ensemble de requêtes;

- F est un framework de modélisation pour les documents, les requêtes et leurs relations;
- $R(q_i, d_j)$ est une fonction d'ordonnancement de pertinence avec le $q_i \in Q$ et $d_j \in D$. Un nombre $\in R$, généralement $R : Q \times D \rightarrow [0, 1]$.

Les modèles proposés en recherche d'information dans la littérature comportent trois caractéristiques principales de documents, y compris du texte, des liens et le multimédia. Dans ce qui suit, nous présentons les principaux modèles de recherche d'information sur la base de ces propriétés.

- **les modèles ensemblistes** : ces modèles trouvent leurs fondements théoriques dans la théorie des ensembles. On distingue le modèle booléen pur (*boolean model*), le modèle booléen étendu (*extended boolean model*) et le modèle basé sur les ensembles flous (*fuzzy set model*) [177, 167, 76].
- **les modèles vectoriels** : basés sur l'algèbre, plus précisément le calcul vectoriel. Ils englobent le modèle vectoriel (*vector model*), le modèle vectoriel généralisé (*generalized vector model*), Latent Semantic Indexing (LSI) et le modèle connexionniste [167, 178].
- **les modèles probabilistes** : se basent sur les probabilités. Ils comprennent le modèle probabiliste général, le modèle de réseau de document ou d'inférence (Document Network) et le modèle de langue [167].

Dans le modèle booléen, les documents et les requêtes sont représentés sous forme d'un ensemble de termes. Ainsi, comme préconisé dans [76], il s'agit d'un modèle ensembliste. Dans le modèle vectoriel, les documents et les requêtes sont représentés sous forme de vecteurs dans un espace de N-dimensions. Pour le modèle probabiliste, la modélisation des documents et des requêtes est basée sur la théorie des probabilités. Nous détaillons dans la suite les principaux modèles issus de chacune de ces trois classes. Nous nous référons aux nombreux ouvrages sur la RI [16, 33, 73] pour des présentations exhaustives des modèles de recherche d'information.

2.3.1 *Modèle vectoriel*

Gerard Salton [178] et ses collègues ont proposé un modèle basé sur le critère de similitude de *Luhn* [129] qui a une forte motivation théorique [73]. Ils ont considéré les représentations de l'index et la requête en tant que vecteurs incorporés dans un espace euclidien de M dimensions, ces dimensions étant les termes du vocabulaire d'indexation, où chaque terme est attribué à une dimension indépendante. Chaque document est représenté par un vecteur : $d_j = (w_{1j}, w_{2j}, ..., w_{Mj})$. De même chaque requête est représentée par un vecteur : $q_i = (w_{1i}, w_{2i}, ..., w_{Mi})$. Avec : w correspond au poids d'un terme dans le document d_j ou dans la requête q_i. La pondération des composantes de la requête est soit la même que celle utilisée pour les documents, soit donnée par l'utilisateur lors de sa formulation. La pertinence entre la requête et le document est traduite par la similarité de leurs vecteurs associés. Le mécanisme de recherche consiste donc à retrouver les vecteurs documents qui s'approchent le plus du vecteur requête. Les principales mesures de similarité utilisées sont :

- **le produit scalaire** :

$$RSV(q_i, d_j) = \sum_{k=1}^{M} w_{ki} \cdot w_{kj} \tag{2.6}$$

- **la mesure de Jaccard** :

$$RSV(q_i, d_j) = \frac{\sum_{k=1}^{M} w_{ki} \cdot w_{kj}}{\sum_{k=1}^{M} w_{ki}^2 + \sum_{k=1}^{M} w_{kj}^2 - \sum_{k=1}^{M} w_{ki} \cdot w_{kj}} \tag{2.7}$$

- **la mesure cosinus** :

$$RSV(q_i, d_j) = \frac{\sum_{k=1}^{M} w_{ki} \cdot w_{kj}}{\sqrt{\sum_{k=1}^{M} w_{ki}^2} \cdot \sqrt{\sum_{k=1}^{M} w_{kj}^2}} \tag{2.8}$$

La métaphore des angles entre vecteurs dans un espace multidimensionnel rend facile l'explication des implications du modèle pour les non-experts. Jusqu'à trois dimensions, on peut facilement visualiser les vecteurs du document et de la requête.

La figure 2.2 illustre un exemple de vecteur de document et un exemple de vecteur de requête dans l'espace qui est engendré par les trois termes "social", "economic" et "political". L'interprétation géométrique intuitive rend relativement facile l'application du modèle à de nouveaux scénarios de recherche d'information.

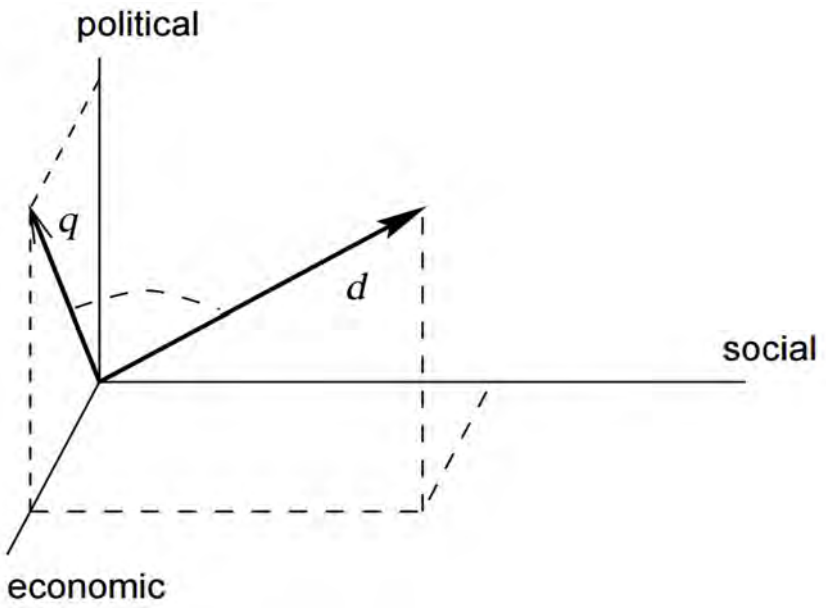

Figure 2.2: Représentation de requête et document dans l'espace des termes à 3 dimensions

Les principaux avantages du modèle vectoriel sont liés tout d'abord, à la pondération non binaire des termes, ce qui offre une meilleure qualité des résultats. Le modèle permet une correspondance partielle ou approximative entre les documents et les requêtes (best match). Les documents sont triés selon leur degré de similarité vis-à-vis de la requête. La longueur des documents est traitée naturellement dans l'appariement, car elle est considérée dans le calcul des poids des termes. Cependant, le modèle vectoriel a l'inconvénient de considérer l'indépendance de tous les termes de l'index. Par contre,

en pratique, la prise en compte globale de la dépendance des termes peut induire à une baisse de qualité des réponses du système [16] car les dépendances sont généralement locales. C'est pour ces raisons que le modèle vectoriel préserve sa popularité en recherche d'information, et reste encore utilisé comme modèle de référence lors d'une évaluation de comparaison de modèles.

2.3.2 *Modèle de langue*

Les modèles de langue ont été appliqués à la recherche d'information par un certain nombre de chercheurs dans la fin des années 1990 [159, 88, 143]. Ils proviennent des modèles probabilistes de génération de langage développés pour les systèmes de reconnaissance automatique de la parole dans le début des années 1980 [165]. Les systèmes de reconnaissance automatique de la parole combinent des probabilités de deux modèles distincts : le modèle acoustique et le modèle de langue. La puissance du modèle acoustique réside dans la production des textes candidats suivants, donnés par ordre de probabilité décroissante : "food born thing", "good corn sing", "mood morning", et "good morning". Ensuite, le modèle de langue a pour but de déterminer l'expression la plus probable, à savoir, dans notre cas "good morning" est la phrase la plus appropriée car elle apparaît plus fréquemment en anglais que les autres phrases. En effet, quand le modèle de langue est combiné avec le modèle acoustique, le système devient capable de prendre des décisions, en augmentant ainsi la performance du système.

Pour la recherche d'information, l'idée de base des modèles de langue est de déterminer la probabilité $P(Q|D)$, la probabilité que la requête Q puisse être générée à partir du document D. Cette formulation est similaire à l'idée derrière les modèles probabilistes formulés pour la première fois dans [140]. Cependant, comme on peut voir plus loin, la façon de calculer $P(Q|D)$ dans les modèles de langue est différente de celle des modèles probabilistes traditionnels [34]. Typiquement, cette probabilité est calculée selon des méthodes paramétriques : on suppose que la distribution des mots suit une certaine norme (par exemple, distribution Poisson) parmi les documents pertinents (et non-pertinents). En fonction des distributions des mots parmi deux ensembles (pertinent et non-pertinent) de documents échantillons, on peut estimer les probabilités des mots pour la pertinence. En suivant cette approche, le modèle de langue du livre que vous lisez en ce moment serait d'attribuer une probabilité exceptionnellement élevée pour les mots "social signals", indiquant que ce livre serait un bon candidat pour les signaux sociaux si la requête contient ces mots.

Le principe des approches utilisant un modèle de langue est différent. On ne tente pas de modéliser directement la notion de pertinence dans le modèle, mais on considère que la pertinence d'un document face à une requête est en rapport avec la probabilité que la requête puisse être générée par le modèle de langue du document. Ainsi, on considère qu'un document D incarne un sous-langage, pour lequel on tente de construire un modèle de langue M_D. Le score du document face à une requête Q est déterminé par la probabilité que son modèle génère la requête :

$$Score(Q, D) = P(Q|M_D) \tag{2.9}$$

On écrira aussi $P(Q|D)$ pour représenter la même probabilité dans les descriptions plus tard.

De façon générale, une requête peut être vue comme une suite de mots : $Q = t_1 t_2 \dots t_n$. Nous avons donc :

$$Score(Q, D) = P(t_1 t_2 \dots t_n | M_D) = \prod_{t_i \in Q} P(t_i | D) \tag{2.10}$$

Cependant, dans cette formulation, les documents longs, et contenant des mots fréquents vont être favorisés. Afin de remédier à ce problème, nous pouvons utiliser la loi de Bayes :

$$Score(Q, D) = P(D|Q) = \frac{P(D) \cdot P(Q|D)}{P(Q)} \tag{2.11}$$

En supposant que l'ordre des documents est indépendant de $P(Q)$ et les termes sont indépendants les uns des autres, la formule 2.11 peut s'écrire comme suit :

$$P(D|Q) \overset{\text{rank}}{=} P(D) \cdot P(Q|D) = P(D) \cdot \prod_{t_i \in Q} P(t_i | D) \tag{2.12}$$

Avec t_i représente les mots de la requête Q.

$P(D)$ représente la probabilité a priori du document D, son utilité est de modéliser et intégrer d'autres sources d'évidence indépendantes de la requête (ex. longueur de document) dans le processus de la recherche d'information. L'estimation de $P(t_i|D)$ peut être effectuée en utilisant différents modèles (ex. Jelineck Mercer, Dirichlet) [215].

Cette probabilité $P(t_i|D)$ s'appuie sur une estimation de la fréquence des termes t_i de la requête Q dans le document D (estimation par maximum de vraisemblance). Ceci peut conduire à assigner une probabilité nulle pour les documents ne contenant pas 1 terme de la requête. Dans ce cas particulier, le score de similarité du document est nul alors que le document pourrait partiellement répondre au besoin en information formulé par la requête. Pour remédier à cet inconvénient, les modèles de langues font appel à des techniques de lissage [103, 135]. Le lissage permet d'assigner une probabilité non nulle à des événements absents. Les méthodes les plus utilisées en RI sont celles basées sur l'interpolation. Elles consistent à estimer la probabilité d'un terme en fonction du document et d'une collection de référence, souvent la collection de document même. Dans la littérature, il y a une série de méthodes proposées. Ci-dessous nous présentons quelques unes classiques.

Le lissage par interpolation, par exemple de *Jelinek-Mercer* [103], consiste à combiner un modèle avec un ou des modèles d'ordre inférieur systématiquement comme suit :

$$P(t_1 t_2 \dots t_n | D) = \prod_{t_i \in Q} (\lambda \cdot P(t_i|D) + (1 - \lambda) \cdot P(t_i)) \tag{2.13}$$

Le modèle de base $P(t_i)$ peut être défini par la probabilité d'occurrence de terme dans la collection estimée selon un maximum de vraisemblance. Dans l'équation, λ est un paramètre inconnu qui doit être fixé de façon empirique, ce qui représente un inconvénient pour cette technique.

Une autre technique de lissage souvent utilisée en recherche d'information est appelée lissage de *Dirichlet*, elle est définie comme suit [215] :

$$P(t_1 t_2 \dots t_n | D) = \prod_{t_i \in Q} \left(\frac{tf(t_i, D) + \mu P(t_i|C)}{|D| + \mu} \right) \tag{2.14}$$

Avec $|D|$ représente la taille du document (le nombre total d'occurrences de mots), et $tf(t_i, D)$ est la fréquence du mot t_i dans D.

Il existe plusieurs autres méthodes et techniques de lissage, le document de *Chen* et *Goodman* les présente soigneusement [47].

2.4 Évaluation

L'évaluation des approches de RI est nécessaire pour mesurer leur efficacité, leur performance et pour pouvoir les comparer en étudiant l'impact des différents facteurs employés dans ces approches.

Un système de RI efficace doit répondre de façon satisfaisante aux besoins d'information de l'utilisateur en termes de qualité des résultats retournés, de rapidité du système ainsi que la facilité d'utilisation du système qui représentent les principaux facteurs à évaluer pour un système de RI [138]. Dans notre cas et de manière plus générale en RI, on s'intéresse particulièrement à : la capacité d'un système à sélectionner des documents pertinents que l'on nomme efficacité (*effectiveness*). Le mode d'évaluation généralement utilisé de nos jours est basé sur celui développé dans le projet Cranfield [51] communément appelé le paradigme de Cranfield. Ce paradigme définit la méthodologie d'évaluation des systèmes de RI en se basant sur trois éléments : une collection de documents sur laquelle les recherches sont effectuées, un ensemble de requêtes de test (besoins des utilisateurs) et la liste des documents pertinents pour chacune des requêtes (jugements de pertinence). L'idée générale de ce paradigme est de créer un environnement unique afin de pouvoir comparer les systèmes équitablement. Cet environnement est appelé la collection de test.

2.4.1 *Collection de test*

La collection ou corpus de test est un aspect fondamental qui constitue le contexte d'évaluation, c'est-à-dire les éléments qui vont servir à tester le processus de recherche d'information. Généralement, chaque collection de test est caractérisée par une collection de documents, une collection de requêtes, et des jugements de pertinence des documents par rapport à ces requêtes.

Dans une tâche de construction d'une collection de test, les jugements de pertinence constituent la tâche la plus complexe. Les jugements de pertinence indiquent pour chaque document du corpus s'il est pertinent, et parfois même son degré de pertinence, pour chaque requête. Afin de construire ces listes de jugements des documents pour toutes les requêtes, les utilisateurs (ou un groupe d'évaluateurs) doivent examiner le contenu de chaque document, et juger s'il est pertinent par rapport à la requête. Dans les campagnes d'évaluation tels que TREC, les collections de documents contiennent plusieurs millions de documents, ce qui rend impossible le jugement exhaustif de pertinence. Donc, dans le cas de grandes collections, les jugements de pertinence sont construits en se basant sur la technique de *pooling* [106], effectuée à partir des 1000 premiers documents retrouvés par les systèmes participants à l'évaluation. La figure 2.3 illustre le protocole adopté par les campagnes d'évaluation officielles.

Pour tenir compte de cela, les collections de documents ont commencé à exister dès les années 1960, afin de permettre une comparaison des systèmes tête-à-tête dans la

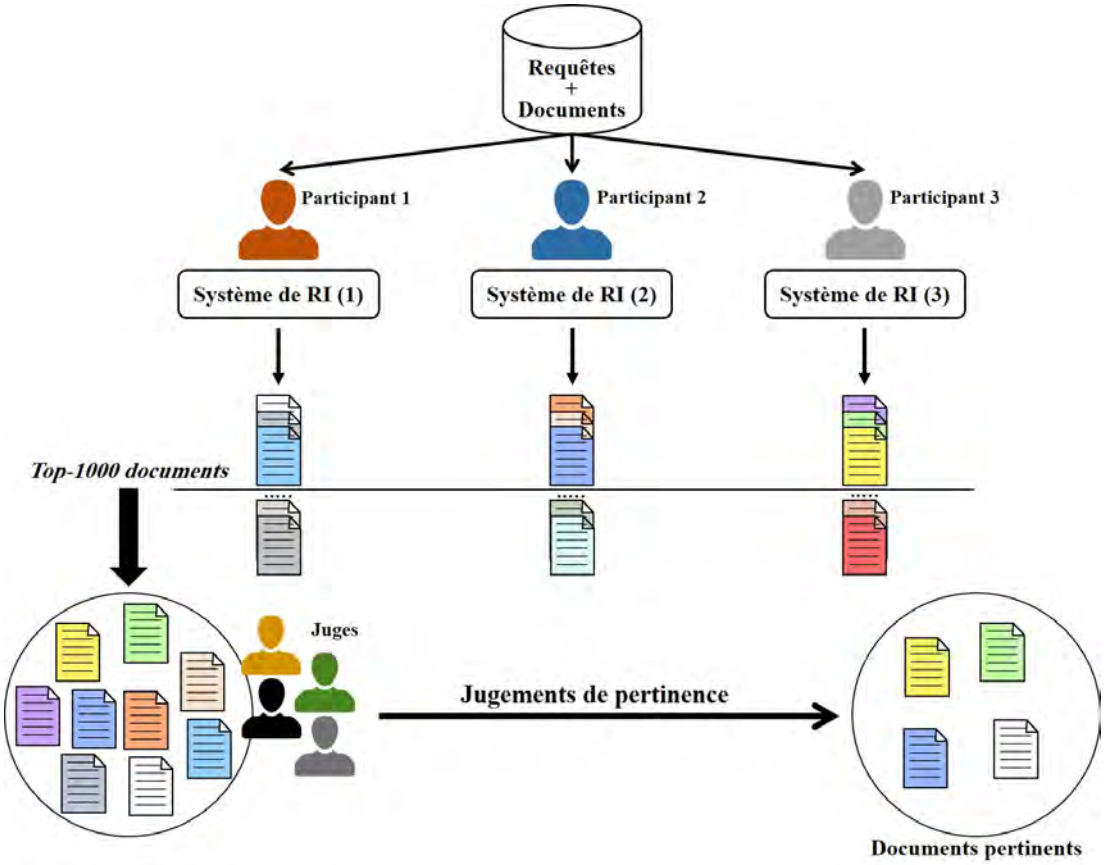

Figure 2.3: Protocole pour les campagnes d'évaluation officielles

communauté du RI. L'une d'elles était la collection Cranfield [74], composée de 1398 résumés d'articles de journaux de l'aérodynamique, un ensemble de 225 requêtes, et un ensemble exhaustif de jugements de pertinence.

Dans les années 1990, le US National Institute of Standards and Technology (NIST) a recueilli une grande quantité de données à travers la compagne de recherche TREC Ad-hoc[2]. Au total, cela a abouti à une collection de test contenant 1.89 millions de documents, principalement constituée d'articles de presse issus de l'agence de presse américaine *Newswire*, ces derniers sont accompagnés des jugements de pertinence pour 450 "tâches de recherche" présentées sous forme de requêtes crées par des experts.

Depuis l'an 2000, Reuters a mis à disposition une large quantité de ressources adoptée pour la classification de texte, le "Reuters Corpus Volume 1". Il est composé de 810.000 articles d'actualité en langue anglaise[3]. Par la suite, un second volume est apparu contenant des données en 13 langues (néerlandais, français, allemand, chinois, japonais, russe, portugais, espagnol, espagnol latino-américain, italien, danois, norvégien et suédois). Pour faciliter la recherche sur les collections de données massives tels que les blogs, la collection Thomson Reuters Text Research Collection (TRC2) a été réalisée, avec plus de 1.8 million de documents[4].

2 http://trec.nist.gov/
3 http://trec.nist.gov/data/reuters/reuters.html
4 http://trec.nist.gov/data/reuters/reuters.html

Les tâches d'évaluation Cross-language ont été menées au sein de Conference and Labs of the Evaluation Forum (CLEF)[5], traitant principalement des langues européennes. La référence pour les langues d'Asie orientale et la recherche multilingue est la NII Test Collection for IR Systems (NTCIR), lancée par la société japonaise pour la promotion des sciences[6].

On trouvera plus de détails sur l'évaluation à base de collections de test dans [180].

2.4.2 *Mesures d'évaluation*

Les mesures d'évaluation permettent d'estimer quantitativement l'efficacité d'un système de RI. L'objectif principal est de quantifier, pour chaque requête la capacité du système à retourner des documents pertinents. La Figure 2.4 illustre les différents ensembles manipulés lors de l'évaluation d'un système de RI, à savoir les ensembles des documents pertinents et des documents retournés par le système. Les documents pertinents non retournés par le système représentent l'ensemble de documents *silence* tandis que les documents non-pertinents retournés par le système génèrent du *bruit*. Un bon système retourne le maximum de documents pertinents (*minimiser le silence*) sans augmenter le nombre de documents non pertinents retournés (*minimiser le bruit*). Nous détaillons dans ce qui suit les principales mesures d'évaluation.

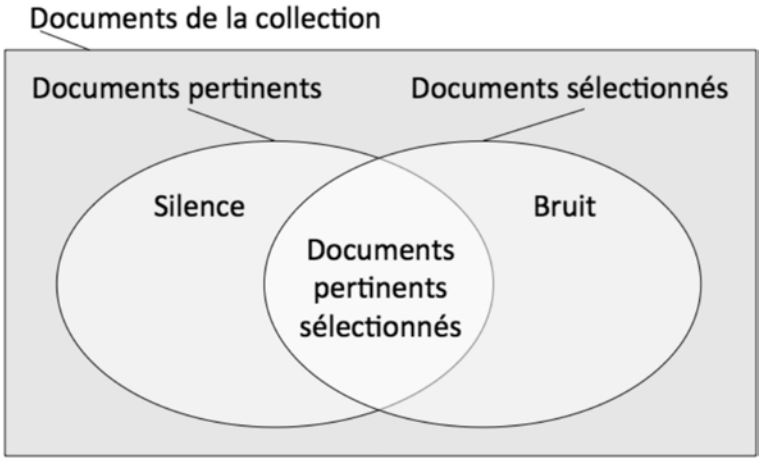

Figure 2.4: Ensembles de documents utilisés pour l'évaluation d'un système de RI

2.4.2.1 *Rappel et précision*

- *Rappel@k* : le rappel évalue la capacité d'un système de recherche d'information à retourner les documents pertinents au rang k dans l'ensemble des documents retournés, et par conséquent, sa capacité à minimiser le *silence*, illustrée dans la figure 2.4. La mesure de rappel est définie par la fraction des documents pertinents sélectionnés sur l'ensemble des documents pertinents au rang k dans la collection. Soit Q un ensemble de $|Q|$ requêtes. La valeur de rappel est en moyenne sur l'ensemble des requêtes $q_h \in Q$ comme suit :

5 http://www.clef-campaign.org/
6 http://research.nii.ac.jp/ntcir/index-en.html

$$\begin{aligned} Rappel@k &= \frac{1}{|Q|} \sum_{q_h \in Q} Rappel(q_h)@k \\ &= \frac{1}{|Q|} \sum_{q_h \in Q} \frac{|S_{q_h}@k \cap R_{q_h}|}{|R_{q_h}|} \end{aligned} \tag{2.15}$$

Avec $S_{q_h}@k$ regroupe l'ensemble des documents sélectionnés par le système de RI pour la requête q_h au rang k. R_{q_h} représente l'ensemble des documents pertinents pour la requête q_h.

- *Précision@r* : la mesure de précision évalue la capacité d'un système de recherche d'informations de ne retourner que des documents pertinents en tête de liste de l'ensemble des documents retournés, à savoir sa capacité à minimiser le *bruit*, illustrée dans la figure 2.4. La précision est définie comme la fraction des documents pertinents dans l'ensemble des documents sélectionnés. Étant donné un ensemble de requêtes Q, la précision d'un système de recherche d'informations est définie par la formule suivante :

$$\begin{aligned} Precision@k &= \frac{1}{|Q|} \sum_{q_h \in Q} Precision(q_h)@k \\ &= \frac{1}{|Q|} \sum_{q_h \in Q} \frac{|S_{q_h}@k \cap R_{q_h}|}{|S_{q_h}@k|} \end{aligned} \tag{2.16}$$

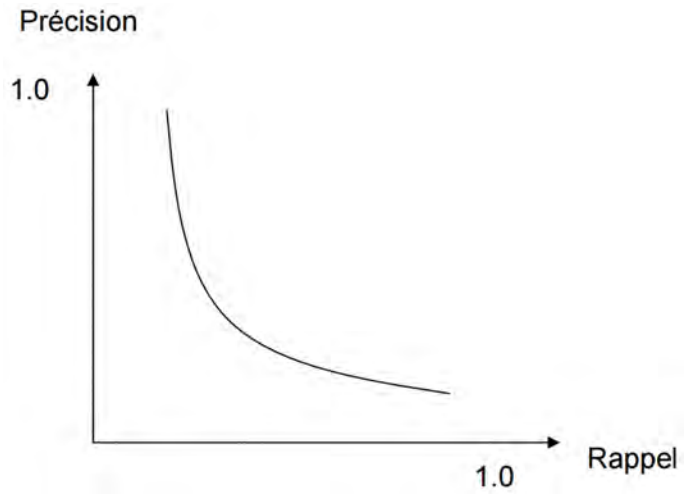

Figure 2.5: Forme générale de la courbe de précision-rappel d'un système de RI

- *Moyenne des précisions moyennes* : Mean Average Precision MAP est obtenue sur l'ensemble des requêtes :

$$MAP@k = \frac{1}{|Q|} \sum_{q_h \in Q} \frac{1}{k} \sum_{R=1}^{k} Precision(q_h)@R \tag{2.17}$$

Il existe plusieurs autres métriques et mesures qui peuvent servir à évaluer la précision d'un système de RI. Nous pouvons citer à titre d'exemple la F-mesure, la R-précision qui sont détaillées dans [180].

2.4.2.2 *Mesure orientée rang nDCG*

Le nDCG (normalized Discounted Cumulative Gain) est l'une des métriques d'évaluation les plus populaires utilisées pour mesurer l'efficacité d'ordonnancement des documents retournés par un système de RI [101]. Le nDCG repose sur des jugements de pertinence graduels de documents, ce qui le rend différent des métriques précédentes. Le nDCG peut être estimé à partir de la mesure de DCG (Discounted Cumulative Gain) appliqué à la liste l_h de résultats retournés normalisée par la même mesure appliquée à l'ordonnancement idéal (meilleur) des jugements de pertinence en fonction de leur degré, notée $IDCG(q_h)$:

$$nDCG@k = \frac{\sum_{q_h \in Q} DCG(l_h)@k}{\sum_{q_h \in Q} IDCG(l_h)@k} \tag{2.18}$$

Où :

$$DCG(l_h)@k = rel_1 + \sum_{R=2}^{k} \frac{rel_R}{log_2 R} \tag{2.19}$$

Avec rel_R qui correspond au jugement effectué par l'utilisateur au document situé au rang R de la liste L_h de résultats.

2.4.2.3 *Test de signification statistique*

Dans le contexte de la recherche d'information, il est important de savoir s'il y a une amélioration d'un système de recherche par rapport à un autre et si cette amélioration est due à une différence réelle entre les deux systèmes ou la différence vient juste du hasard. Parmi les autres raisons, par exemple, il y a du bruit inhérent à une évaluation. Certains sujets sont plus difficiles que d'autres, et les évaluateurs engagés pour juger de la pertinence des documents sont des êtres humains et donc ouverts à la variabilité dans leur comportement [186]. Cette différence entre les améliorations des systèmes est souvent mesurée à l'aide des tests de signification statistique.

Quand un test statistique est utilisé pour comparer les performances de deux systèmes de recherche (soit système X et système Y), un niveau de confiance typique de 95% est utilisé. Cette valeur signifie que dans 95% des choix de X et Y le rendement de X ira au-dessus de celle de Y. En d'autres termes, si la probabilité de la différence observée entre le système X et le système Y, connue en tant que valeur de signification, est assez petite, c-à-d., inférieure à 0.05, alors cette différence est considérée comme statistiquement significative car il y a une probabilité de 5% d'être faussement positifs. Étant donné que la valeur de signification représente la probabilité d'erreur en admettant que le résultat est correct, la valeur 0.05 est considérée comme un niveau d'erreur acceptable.

Les tests de signification les plus couramment utilisés en recherche d'information sont le *t-test Student* [75] et le *Wilcoxon signed-rank test* [202]. Cependant, malgré le fait que le *t-test Student* adopte une distribution paramétrique, beaucoup d'études, par

exemple, *Sanderson* et *Zobel* [181], ont montré qu'il peut correctement distinguer entre les améliorations des deux systèmes.

Jusqu'à maintenant, nous avons introduit le domaine de la recherche d'information dont cette thèse fait partie. Dans le chapitre suivant, nous allons commencer notre investigation spécifique sur l'état de l'art lié à l'implication du Web social dans la recherche d'information.

3

RECHERCHE D'INFORMATION SOCIALE

Introduction

Des millions d'utilisateurs à travers le monde ont intégré les sites de réseaux sociaux dans leurs routines quotidiennes. Les réseaux sociaux représentent des liens entre des personnes qui partagent des intérêts communs. Les comportements individuels et collectifs peuvent être extraits à partir des réseaux sociaux.

Ces dernières années, et particulièrement depuis 2005, les chercheurs ont pris conscience que ces réseaux sociaux peuvent être une source fructueuse pour contribuer au développement de plusieurs tâches en recherche d'information. Par exemple, la recherche de ce type d'information pour satisfaire un besoin en information de l'utilisateur, ou l'intégration de ces contenus sociaux comme une nouvelle source d'évidence dans le modèle de recherche afin d'améliorer la qualité des résultats de recherche.

Ce chapitre présente la recherche d'information sociale. Nous donnons tout d'abord un panorama du type d'information sociale présente dans le Web. Ensuite, nous définissons la notion de la RI sociale, en mettant en exergue les principales tâches de la RI sociale. Ensuite, nous présentons un aperçu sur des travaux liés à l'exploitation des informations sociales dans le processus de la RI. Enfin, nous analysons les limites de l'état de l'art en positionnant nos contributions.

3.1 Information sociale dans le Web

L'information sociale concerne toutes les informations générées par les utilisateurs sur la toile. Elle est le résultat des services du Web 2.0 (ex. Facebook, Twitter, etc), qui permettent à l'utilisateur d'annoter, de collaborer et de contribuer au développement de la démocratisation de la production du contenu dans le Web. En effet, les internautes sont passés de simples consommateurs à des producteurs d'information (voir la figure 3.1). Leurs contributions peuvent être de différentes natures : les contenus publiés dans les plate-formes sociales (ex. Facebook, Twitter, etc), les relations (ex. amis, followers, etc), les réactions/interactions, les annotations, les commentaires, l'opinion, etc. L'ensemble de ces informations est appelé contenus générés par des utilisateurs, (en anglais UGC : User Generated Content), encadrés particulièrement par les réseaux sociaux.

3.1.1 *Réseaux sociaux*

Selon *Wasserman* et *Faust* [199], un réseau social peut être défini comme suit :

> *"a finite set or sets of actors and the relation or relations defined on them. The presence of relational information is a critical and defining feature of a social network."*

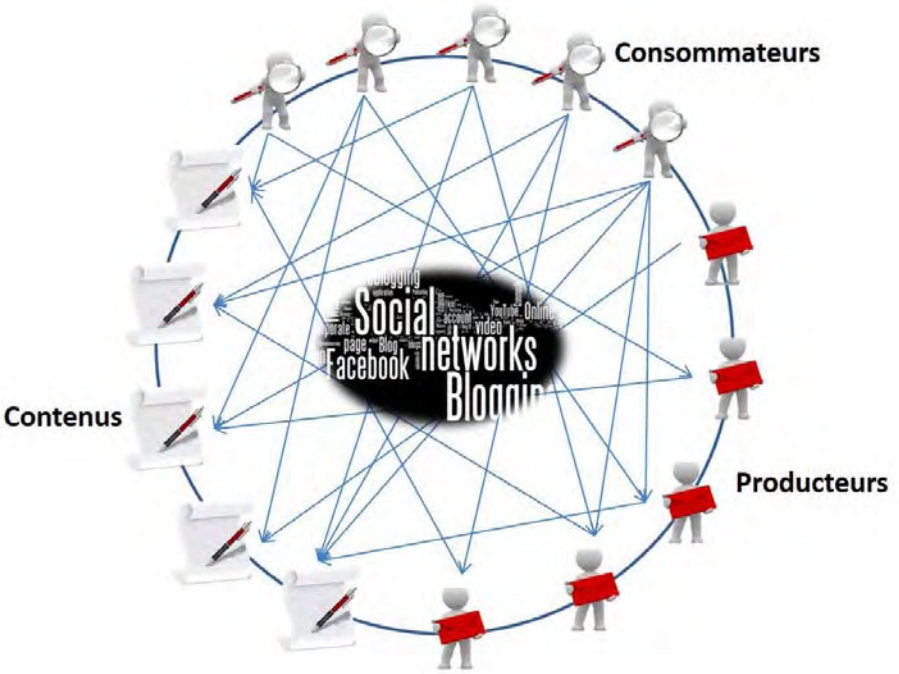

Figure 3.1: Le Web social (modèle producteur-consommateur)

D'un point de vue théorique, un réseau social est essentiellement un grand graphe où les nœuds représentent les utilisateurs et les arêtes représentent les relations entre les utilisateurs [64]. Il est représenté par un graphe $G = (V, E)$ où l'ensemble des nœuds V représente les utilisateurs et l'ensemble des arêtes $E = V \times V$ représente les relations entre eux. Dans le cas d'un réseau social non-orienté, une arête (v_i, v_j) représente une relation symétrique, associant des utilisateurs v_i et v_j. L'amitié est un exemple typique de relation non-orientée dans le réseau social. Dans le cas d'un réseau social orienté, une arête (v_i, v_j) représente une relation orientée de v_i à v_j. Par exemple, la communication électronique est représentée par une bordure directe (v_i, v_j) où v_i, v_j représente l'expéditeur et le destinataire, respectivement. Pour indiquer l'importance d'un utilisateur dans le réseau ou les points forts d'une relation sociale, les poids du réseau sont associés à des nœuds et des arêtes, respectivement.

La figure 3.2 illustre un réseau social de 12 personnes. Une arête bidirectionnelle dénote une relation réciproque entre les acteurs. Dans cet exemple, tous les acteurs des réseaux sociaux sont reliés à au moins un nœud dans le graphe.

D'un point de vue pratique, un réseau social est défini par un ensemble d'acteurs qui partagent plusieurs relations avec d'autres [82]. Les acteurs réfèrent principalement à des personnes, mais représentent, dans un autre contexte, les institutions, les communautés, les éléments d'information, etc. Un réseau social peut impliquer un ou plusieurs types d'acteurs, comme indiqué par les réseaux sociaux professionnels où deux types d'acteurs sont présents : les travailleurs et les entreprises. Les relations sociales impliquent deux ou plusieurs acteurs dans une amitié, partenariat ou encore un simple échange des contenus sociaux. Les deux acteurs et les relations sociales évoluent avec le temps. De nouveaux acteurs et relations peuvent apparaître dans le réseau social, d'autres peuvent disparaître.

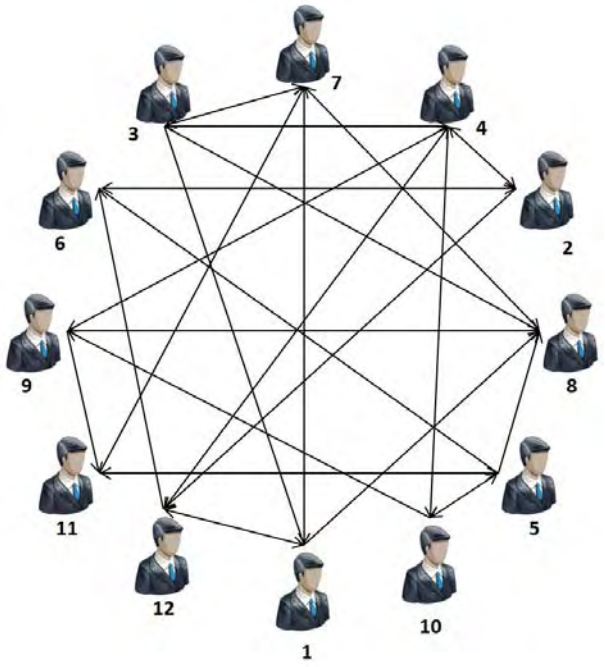

Figure 3.2: Exemple d'un réseau social

Les motivations des utilisateurs pour utiliser les réseaux sociaux sont diverses [100, 102, 123, 217]. En effet, ces réseaux sociaux offrent relativement un nouveau service de communication aux personnes désirant diffuser des informations qu'ils n'auraient probablement pas partagé autrement en utilisant les services existants (ex. messagerie instantanée, téléphone, courriel, etc). Les utilisateurs proposent des mises à jour biographiques et des publications sur des évènements d'actualité [61] ou des situations de crise et catastrophes [175, 195]. Les réseaux sociaux ont joué un rôle important durant les révolutions du printemps arabe, ce qui a permis aux gens de communiquer dans un cercle sociale relativement restreint.

Nous listons ci-dessous quelques exemples de réseaux sociaux :

- **Facebook**[1] : Créé en 2004 par *Mark Zuckerberg* à l'université Harvard. C'est le réseau social le plus populaire au monde avec 1.49 milliards d'utilisateurs actifs mensuellement. En moyenne, les utilisateurs comptent 130 amis et passent 6 heures et 45 minutes par mois. Facebook est un réseau social sur Internet permettant à toute personne possédant un compte de créer son profil et de publier des informations, dont elle peut contrôler la visibilité. L'usage de ce réseau s'étend du simple partage d'informations (statuts, photos, liens, vidéos, etc) à la constitution de pages et de groupes visant à faire connaitre des institutions, des entreprises ou des causes variées.

- **Twitter**[2] : Créé en mars 2006 par *Jack Dorsey* à San Francisco, et lancé en juillet de la même année. Le service est rapidement devenu populaire, jusqu'à réunir plus de 300 millions d'utilisateurs actifs, qui publient 500 millions de tweets chaque

1 https://www.facebook.com/
2 https://twitter.com/

jour dont plus de la moitié "tweetent" depuis leur téléphone mobile. Chaque utilisateur passe en moyenne une durée de 15 minutes par visite. Twitter est un outil de réseau social et de microblogging qui permet d'envoyer gratuitement des messages courts, appelés tweets (gazouillis) sur Internet. Ces messages sont limités à 140 caractères.

- **Delicious**[3] : Créé fin 2003 par *Joshua Schachter* et acheté par Yahoo en 2005, et racheté par *Chad Hurley* et *Steve Chen*, les fondateurs de Youtube. Delicious est un service de bookmarking, ou peu importe sur quel ordinateur on travaille on peut sauvegarder (marquer) et partager des marque-pages Internet au sein du réseau, et de les classer selon le principe de folksonomie par des mots-clés (ou tags). On peut créer aussi des "tag clouds" ou nuages de mots clefs spécifiques à l'ensemble des signets.

- **GooglePlus**[4] : Créé en 2011 par l'équipe de *Google* et lancé à la fin de cette année, et il a plus de 300 millions d'utilisateurs actifs. Les utilisateurs de Google+ peuvent voir les mises à jour de leurs contacts grâce à des cercles à travers le Stream, qui est semblable aux flux de nouvelles de Facebook. La zone de saisie permet aux utilisateurs de se mettre à niveau sur les états ou l'utilisation des icônes à télécharger et partager des photos, vidéos, liens, etc.

- **LinkedIn**[5] : Créé en mai 2003 par *Reid Hoffman* et *Allen Blue*. Selon les dernières statistiques plus de 259 millions de professionnels dans le monde sont inscrits, ainsi que plus de 150 secteurs d'activité dans 200 pays sont présents sur LinkedIn. LinkedIn est un réseau social à utiliser dans un contexte d'affaire. Les pages des utilisateurs exposent leurs carrières professionnelle et leur permettent de préciser leurs intérêts en matière de débouchés professionnels, d'emplois, loisirs et autre vie sociale, et cela en partageant des liens, textes, vidéos, etc.

- **Pinterest**[6] : Créé en 2010 par *Paul Sciarra*, *Evan Sharp* et *Ben Silbermann*. Approximativement plus de 100 millions d'utilisateurs actifs par mois, et le temps moyen passé sur le site est de 16 minutes et 20 secondes par visite. Pinterest est un site web américain mélangeant les concepts de réseautage social et de partage de photographies. Il permet à ses utilisateurs de partager leurs centres d'intérêt, passions, hobbies, sachant que le nom du site est un mot-valise des mots anglais "pin" et "interest" signifiant respectivement "épingler" et "intérêt".

- **StumbleUpon**[7] : créé en 2002 par *Garret Camp* comme un Plugin Firefox, il a été acheté par Ebay en 2007 avant d'être racheté par ses fondateurs en 2009 avec l'aide de *Sherpalo Ventures*, *Accel Partners* et *August Capital*. StumbleUpon est un moteur de recommandation et de découverte de sites et de contenus sur le Web annonce avoir dépassé le Cap de 30 millions d'utilisateurs dont 10 millions ajoutés en l'espace d'un an. Pendant cet intervalle, le nombre de "Stumbles" personnalisés est passé de 400 millions à plus d'un milliard par mois.

3 https://delicious.com/
4 https://plus.google.com/
5 https://www.linkedin.com/
6 https://www.pinterest.com/
7 https://www.stumbleupon.com/

Les intérêts de ces réseaux sociaux sont multiples, par exemple, le milieu professionnel les voit comme des outils indispensables pour améliorer la visibilité des entreprises sur le Web. En recherche, les réseaux sociaux suscitent beaucoup d'intérêts en termes de contenu social généré par leurs utilisateurs, qui peut être utile dans les tâches de recherche d'information.

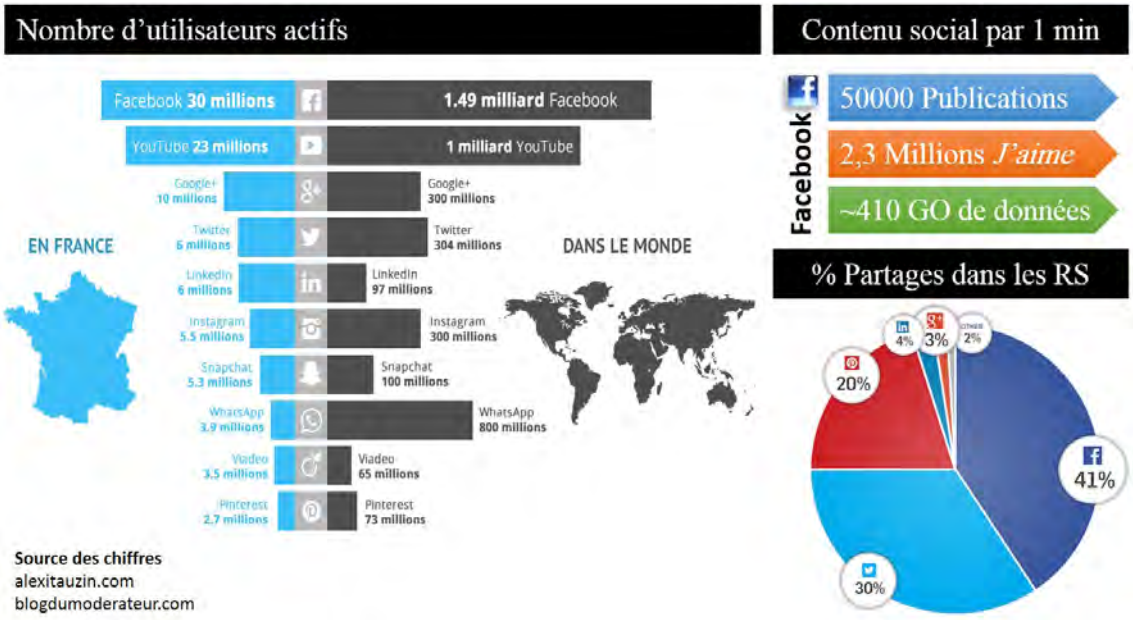

Figure 3.3: Statistiques sur les réseaux sociaux (septembre 2015)

3.1.2 *Contenus générés par les utilisateurs*

Un changement radical dans l'interaction des utilisateurs dans le Web est arrivé dans le début des années 2000. Cette décennie a vu le développement et la maturation des médias sociaux, les environnements en ligne qui permettent aux personnes d'interagir de façon novatrice. En ce qui concerne la science de l'information, une évolution importante apportée par ces médias a été la prolifération des contenus générés par l'utilisateur.

3.1.2.1 *Définition*

Les contenus générés par l'utilisateur, connus sous l'abréviation anglaise UGC (User Generated Content), ont vu ces dernières années une croissance importante et prennent de plus en plus d'intérêt dans le Web. Le terme UGC est cité plusieurs fois et avec différentes définitions selon sa nature informationnelle (tag, j'aime, commentaire, vidéo, etc). Selon *Baeza-Yates* [15] l'UGC est défini comme suit :

> *"User Generated Content is one of the main current trends in the Web. This trend has allowed all people that can access the Internet to publish content in different media, such as text (e.g. blogs), photos or video."* [15]

Dans le même contexte, *Volkovich* et *Kaltenbrunner* [196] voient que l'UGC peut être sous forme d'interactions utilisateur-ressource :

> *"Social news websites have gained significant popularity over the last few years. The participants of such websites are not only allowed to share news links but also to annotate, to evaluate and to comment them."* [196]

Selon toutes ces définitions, le contenu généré par l'utilisateur n'est pas uniquement un document, une image ou une vidéo partagée ou créée par l'utilisateur. D'autres types de contenu y compris la fourniture de méta-données supplémentaires, pour les ressources en ligne telles que des descriptions, ou des termes créés par un ensemble d'utilisateurs afin d'enrichir une ressource par des tags, ou encore un commentaire, un avis. Aujourd'hui, beaucoup de site de médias sociaux offrent aux utilisateurs la possibilité de partager publiquement leurs idées et opinions avec d'autres personnes. La blogosphère est considérée comme étant la source la plus notable pour l'UGC sur le Web. Les services de microblogging, comme Twitter, permettent aux utilisateurs de publier de courts commentaires. Aussi, d'autres sources telles que les services de bookmarking (ex. Delicious et Digg) permettent aux utilisateurs d'annoter les contenus publiés sur le Web. L'UGC peut aussi être sous forme d'une mention ou toute réaction envers une ressource à travers des boutons de j'aime, partage, commentaire (ex, Facebook, voir la figure 3.4), marquer un lien (ex, Delicious, Digg). En effet, c'est à ce type d'UGC auquel nous nous intéressons dans notre travail de recherche.

Figure 3.4: Exemple de contenus sociaux générés par les utilisateurs de Facebook

Selon *Amer-Yahia* et ses collègues [207], le graphe des contenus sociaux comprend 4 types d'interactions : *content-to-content*, *content-to-person*, *person-to-person* et *person-to-content* (voir la figure 3.5). Ces interactions définissent le contexte de la production sociale et la consommation sociale de l'information.

- *Person-Content* : par rapport à cette catégorie on peut citer les traces des utilisateurs et les signaux sociaux sur les réseaux sociaux tels que le commentaire, le j'aime et le partage sur Facebook. Les contenus générés par l'utilisateur peuvent être aussi des tweets, des tags ainsi que des contenus tels que des images, des vidéos et des documents textuels.

- *Content-Person* : cette catégorie concerne principalement, par exemple les mentions sur Twitter ou les citations des auteurs sur les réseaux académiques comme Researchgate, etc.

- *Person-Person* : cette catégorie concerne les relations sociales telles que l'amitié, l'abonnement sur Facebook et Twitter, la recommandation de compétences comme sur LinkedIn, messages, etc.

- *Content-Content* : dans cette catégorie on peut apercevoir, par exemple des liens hypertextes ou des citations.

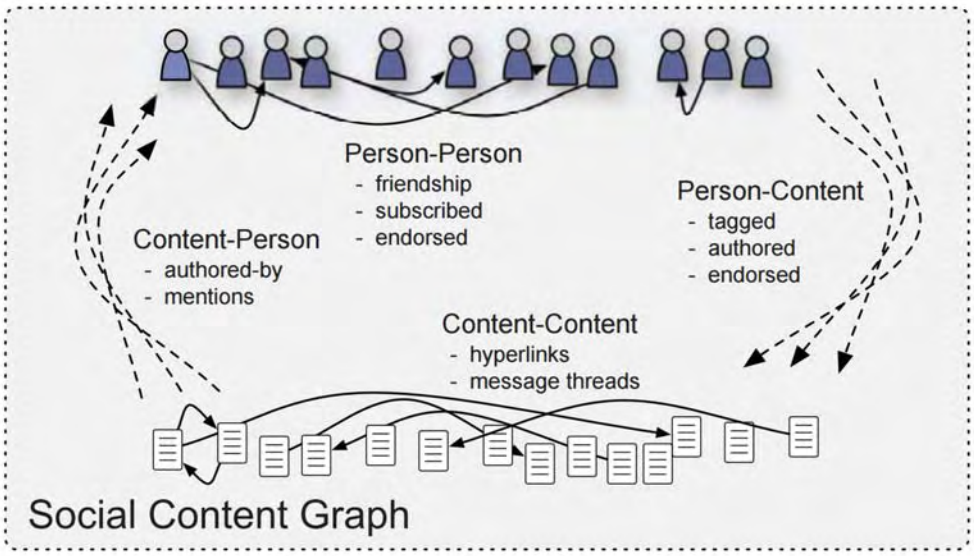

Figure 3.5: Graphe du contenu social selon *Amer-Yahia* [207]

Ces données sociales générées à travers les interactions mutuelles entre les différentes entités dans le Web (utilisateurs, actions, documents, etc) représentent le fruit d'un mécanisme d'intelligence collective que les gens ont développé sur le Web, appelé la "sagesse de la foule" (en anglais *wisdom of crowds*) [190].

L'UGC traduit une tendance de fond, de plus en plus de sites Internet deviennent des plateformes mettant à disposition des outils pour que les internautes puissent créer des contenus. Les réseaux sociaux et les plateformes vidéo/photo ont poussé ce principe très loin. Souvent l'UGC est remis en cause pour son caractère peu fiable, l'information délivrée est au contraire bien plus fiable et étendue que celle qu'on peut retrouver dans les médias traditionnels et permet même d'offrir une source d'information avec une certaine objectivité. A cela plusieurs raisons :

- La disponibilité immédiate sur Internet de plusieurs sources d'informations permet de les comparer, les confronter et en tirer l'information la plus objective possible.
- La sphère Internet regroupe des millions d'utilisateurs apportant avec eux leurs bagages de connaissances, leurs qualifications, leurs diplômes et leurs passions. Ces mêmes personnes sont parfois bien mieux qualifiées que des journalistes par exemple pour parler d'un domaine particulier.
- Les dispositifs de contrôle, notamment sur Wikipédia par exemple, permettent de limiter le vandalisme et la propagation d'informations fausses.

3.1.2.2 *Signaux sociaux*

Aujourd'hui, les signaux sociaux représentent un des types d'UGCs les plus populaires sur le Web. En effet, les pages Web comprennent différents boutons de réseaux sociaux

où les utilisateurs peuvent exprimer s'ils soutiennent, recommandent ou n'aiment pas un contenu (texte, image, vidéo, etc) [4]. Ces boutons qui décrivent des actions relevant d'activités sociales (ex. *j'aime*, *partage*, +1, etc) sont liés à des réseaux sociaux spécifiques (ex. Facebook[8], Google+[9], etc) avec des compteurs indiquant le taux d'interaction avec la ressource Web (voir la figure 3.6).

Figure 3.6: Exemple d'une ressource contenant des signaux sociaux

Un signal social est donc une mesure de l'activité des médias sociaux. C'est une interaction sociale d'une personne réelle avec une ressource sur le Web à travers des fonctionnalités offertes par les réseaux sociaux. Comme avec des backlinks[10], les actions sociales peuvent être interprétées comme approbations pour la ressource, ce qui aide à améliorer son classement dans les moteurs de recherche.

3.1.2.3 *Types des signaux sociaux*

En général, chaque réseau social emploie ses propres signaux sociaux dont les règles de fonctionnement diffèrent et qui n'entraînent pas toutes les mêmes significations et le même impact sur la stratégie Web. Le tableau 3.1 résume les signaux sociaux les plus populaires sur les réseaux sociaux :

8 https://www.facebook.com/
9 https://plus.google.com/
10 Backlink ou lien entrant est un hyperlien pointant vers un site ou une page Web.

Type	Exemple	Réseaux sociaux
Votes	J'aime +1	Facebook, Google+, LinkedIn, StumbleUpon
Message	Tweet Publication	Facebook, Google+, LinkedIn, Twitter
Partages	Partage Re-tweet	Facebook, Google+, LinkedIn, Twitter, Buffer
Signets	Bookmark Épingler	Delicious, Pinterest, Diigo, Digg
Commentaires	Commentaire Répondre	Facebook, Google+, LinkedIn, Twitter
Relations	Abonnés Amis	Facebook, Twitter

Tableau 3.1: Liste des différents types des signaux sociaux

3.1.2.4 *Signaux sociaux et moteurs de recherche*

Malgré l'absence de consensus clair sur la relation exacte entre les signaux sociaux et les célèbres moteurs de recherche, il y a beaucoup de raisons pour lesquelles les signaux ne peuvent pas être ignorés. Plutôt que de considérer les signaux et le classement des résultats par les moteurs de recherche comme deux composants distincts, il est utile de les considérer comme des processus inter-connectés travaillant vers l'objectif global qui est d'augmenter la visibilité en ligne. Dans ce qui suit nous allons découvrir pourquoi les signaux sociaux sont importants.

- **Les premières impressions**

  A titre d'exemple, si une entreprise est active dans l'espace social, il y a de fortes chances que ses pages Web deviennent populaires sur les réseaux sociaux, et les utilisateurs pourront interagir avec le contenu de ces pages ce qui augmente le trafic (et le trafic sur un site est un des critères importants des moteurs de recherche). Par conséquent, même pour les moteurs de recherche qui n'utilisent pas de façon directe les signaux sociaux il y a de fortes chances que les pages concernant cette entreprise apparaîtront parmi les résultats de recherche.

  Une page sur Facebook[11] peut être la première interaction d'un utilisateur avec une entreprise, sachant que ces pages sont indexables par les moteurs de recherche. C'est pour cette raison qu'il est très courant maintenant de bien entretenir les pages Facebook et les mettre à jour régulièrement avec du contenu frais. Un contenu périmé ou une page sociale négligée peut rapidement conduire à une impression négative chez l'utilisateur. Par conséquent, les traces laissées par les utilisateurs sur ces pages peuvent mesurer la qualité, la pertinence et la récence de ces pages.

  Pour résumer, les signaux sociaux sont importants dans une stratégie Web et ils représentent une des manières de détecter la bonne *réputation* de l'information

11 https://www.facebook.com/business/products/pages

sur le Web en les analysant. Ils ont donc un impact indirect sur les résultats des moteurs de recherche actuels.

- **Construction des liens via le social**

  Les sites de médias sociaux sont des espaces Web où les gens partagent des contenus avec leurs amis, famille, etc. Toujours avec le même exemple concernant les entreprises compétitives, elles doivent être là où les conversations se produisent. Les signaux sociaux tels que le *partage* peuvent entraîner des liens vers des pages Web liées à ces entreprises. Sachant que les robots de crawling des moteurs de recherche sont en mesure d'indexer les contenus des médias sociaux comme tout autre contenu Web, donc la propagation des liens via des actions sociales sur des réseaux populaires comme Facebook et Twitter, peut contribuer à la transmission du "jus de lien" (en anglais "link juice")[12] à ces pages Web.

- **Signaux sociaux et Bing**

  Bing, le deuxième plus grand moteur de recherche après Google, est explicite sur son utilisation des signaux tels que le *tweet* et le *j'aime* de Facebook ainsi que d'autres signaux sociaux comme facteurs de classement. Contrairement à Google, les algorithmes de Bing ne mettent pas autant l'accent sur le nouveau contenu d'un site Web. Au lieu de cela, il se focalise sur le contenu des médias sociaux, des liens, la *popularité* sur différents réseaux sociaux qui sont des facteurs importants pris en compte par Bing pour définir le classement des résultats.

  Alors que Google est encore principalement un moteur de recherche basé sur du texte, Bing embrasse le social et le multimédia. Bing est un exemple notable sur l'exploitation des images et des informations issues des médias sociaux pour fournir des résultats plus riches pour les utilisateurs. L'activité des médias sociaux est également présentée sur les pages de résultats de Bing beaucoup plus visible que les autres moteurs de recherche. Des *Tweets* ou des *épingles* (issus de Pinterest) contenant des mots clés pertinents, ainsi que le contenu d'autres formes de médias sociaux comme Facebook, sont souvent intégrés dans les résultats de recherche de Bing. Donc, publier activement des contenus visuels (image, vidéo, etc) sur les réseaux sociaux est un excellent moyen d'accroître la visibilité sur Bing. Bing met aussi en service une fonctionnalité sous forme de barre latérale (social sidebar) qui exploite Facebook [84]. C'est une troisième colonne sur ses pages de résultats qui permet aux utilisateurs connectés de *commenter* et *aimer* les résultats pertinents issus de Facebook sans quitter la page de recherche [26]. Ce service est fonctionnel uniquement aux états unis.

- **Signaux sociaux et Google**

  Google est encore mystérieux sur la façon dont il exploite les signaux sociaux pour déterminer le classement de ses résultats de recherche, mais des études menées chaque année depuis 2013 par Searchmetrics montrent que les signaux sont un facteur important sur le grand moteur de recherche [139].

12 Link juice ou jus de lien est un terme utilisé par les professionnels du référencement pour évoquer les effets bénéfiques sur le référencement naturel que peut transmettre un site à un autre site ou plutôt une page à une autre page par un lien sortant.

Bien que Google n'a pas de partenariat avec Facebook, le moteur de recherche a encore accès à la partie des données de Facebook disponibles publiquement et susceptible d'utiliser certaines d'entre elles pour mieux comprendre la *popularité* des pages Web. Mais cette année, Google a conclu un accord avec Twitter pour indexer les tweets en temps réel, leur permettant d'être plus consultables. L'accès à la base de Twitter signifie que toutes les informations sur Twitter sont disponible pour Google automatiquement sans utilisation des robots de crawling, ce qui rend facile à Google l'affichage instantané des Tweets dans les résultats de recherche. Les algorithmes de Google mettent aussi l'accent sur les profils Twitter bien entretenus qui tweetent et retweetent souvent des contenus, mais le comment reste une boite noire.

En outre, il est pas un secret que Google donne du poids à son propre réseau social, Google+. Le contenu et les interactions sur Google+ sont connus pour avoir un impact positif sur le classement de ses résultats. L'étude réalisée par Searchmetrics [139] a montré que le signal +1 enregistre une corrélation plus élevée avec le classement des résultats de recherche de Google, que d'autres paramètres bien connus tels que les signaux de Facebook et la fréquence des mots-clés .

La popularité des UGCs, en particulier dans le contexte des médias sociaux a donné la naissance à une multitude de nouveaux problèmes en recherche d'information [1]. Précisément, comment mobiliser ces contenus sociaux en faveur de la RI est une question ouverte, qui a donné la naissance à une nouvelle thématique en RI, la Recherche d'Information Sociale (RIS) [21].

3.2 Notion de la RI sociale

L'émergence des réseaux sociaux dans la vie quotidienne des utilisateurs, en produisant des informations qui sont rarement disponibles dans d'autres espaces d'Internet, a contesté les approches traditionnelles de RI qui classent les documents indépendamment de leur contexte social. Pour résoudre ce problème, la RI sociale prévoit une nouvelle génération de modèles de recherche qui font usage de la structure de réseau social et de données sociales afin d'améliorer le processus de RI [21].

La RI sociale est un domaine de recherche innovant qui a émergé au début des années 2000. Elle rassemble deux domaines de recherche, à savoir la recherche d'information et l'analyse des réseaux sociaux.

En 2006, *Kirsch* et ses collègues [115] ont défini la recherche d'information sociale par la prise en compte des données des réseaux sociaux dans le processus de recherche d'information.

> *"Social information retrieval systems are distinguished from other types of information retrieval systems by the incorporation of information about social networks and relationships into the information retrieval process."* [115]

En 2008, *Evans* et ses collègues [68] ont proposé une nouvelle définition qui accentue le terme *interaction sociale* :

> *"Social search is an umbrella term used to describe search acts that make use of social interactions with others. These interactions may be explicit or implicit, co-located or remote, synchronous or asynchronous."* [68]

En 2011, une définition similaire a été proposée par *Karweg* et ses collègues qui ont mené des travaux intéressants concernant le Web social de nos jours.

> *"Social search is a variant of information retrieval where a document or website is considered relevant if individuals from the searcher's social network have interacted with it."* [108]

Une définition générale par *Alonso* en 2011 considère les réseaux sociaux comme une source d'intelligence collective :

> *"Social search is a general term used to describe searches that utilize social networks or involve a collective intelligence process to help the user satisfy an information need."* [5]

En 2012, les chercheurs de *Microsoft*[13], plus précisément, *Teevan* donne une définition plus complète. Sa définition décrit la recherche d'information sociale comme suit :

> *"Social search is an emerging research area that explores how social interactions and social data can enhance existing information-seeking experiences, as well as enable new information retrieval scenarios. This session will showcase different models of social search, including 1) the use of social data to augment search, 2) social data as new information to be searched, and 3) social interaction and collaboration as part of the search process."*

Les systèmes de RI sociale se distinguent des autres types de systèmes de RI par l'exploitation des UGCs et des relations issus des réseaux sociaux dans le processus de recherche d'information. Un système de RI sociale peut être vu aussi comme un système de recherche dans les réseaux sociaux uniquement.

3.3 RI sociale : une vue d'ensemble

Les données sociales ont conduit à l'émergence de nouvelles tâches ainsi qu'à la réactualisation des tâches de RI pour mieux appréhender ces données.

Les approches de RI sociale ont étendu les modèles traditionnels avec différentes caractéristiques sociales afin de satisfaire des motivations sociales derrière les besoins d'information de l'utilisateur. En outre, de nouvelles approches sociales entrent en vue afin de répondre aux nouveaux besoins en information et en RI initiées par les pratiques sociales sur le Web. Une des différentes définitions mentionnées précédemment, nous reprenons celle qui est propre pour *Teevan*. Nous considérons en effet la RI sociale selon 3 axes :

13 http://research.microsoft.com/en-us/events/fs2012/agenda.aspx

1. le premier axe concerne la recherche d'information de nature sociale. Il s'agit de trouver des informations sociales qui répondent à l'utilisateur [146]. On distingue par exemple la recherche d'information dans les blogs, microblogs [54, 130], la recherche de conversations [137], la recherche des experts [134], ou encore des réponses à des questions spécifiques auprès des amis, familles, collègues, ou même des personnes inconnues [84], etc. Nous discutons cette catégorie dans la section 3.3.1.

2. le deuxième porte sur l'exploitation des contenus sociaux pour améliorer la RI, dans laquelle l'information sociale est utilisée afin d'améliorer le processus de recherche d'information, par exemple, les tags dans les folksonomies ont été trouvés utiles pour améliorer la recherche Web et la recherche personnalisée [77, 85], le reclassement (*re-ranking*) des résultats de recherche, la reformulation (expansion) de requête, la personnalisation, etc. Nous discutons cette catégorie de manière générale dans la section 3.3.2, ainsi que les travaux liés à nos contributions dans la section 3.4.

3. le troisième paradigme concerne la recherche d'information effectuée par plusieurs personnes, recherche collaborative. Cette catégorie est loin du cadre que nous traitons, nous ne la décrivons pas [188].

3.3.1 *Recherche d'information dans les contenus sociaux*

Les services sociaux tels que Twitter et Facebook permettent aux utilisateurs de partager et publier des contenus avec un grand public d'utilisateurs (amis, familles, collègues, inconnus, etc). En outre, les utilisateurs les exploitent comme source d'information pour répondre à différents types de besoins, chercher l'opinion des gens sur un sujet donné, chercher un évènement, un expert, des amis ou chercher tout simplement un contenu (blog, micronlog, etc). Nous décrivons ci-après une liste non exhaustive de tâches de RI sur les contenus sociaux.

3.3.1.1 *Recherche dans les services sociaux*

En plus des documents traditionnels, les utilisateurs expriment leurs besoins d'information pour rechercher des contenus générés socialement [63, 67, 131]. Dans ce but, la recherche sociale consiste à assurer la tâche de recherche dans le graphe social tout en tenant compte de la structure du réseau social [183]. En conséquence, les documents sont classés par leur pertinence thématique ainsi que leur importance dans le réseau social [114, 115]. Nous mettons dans cette classe toutes les approches de RI visant les contenus sociaux. On parlera alors de la RI dans les microblogs, la recherche dans les blogs, la recherche dans le réseau social. La spécificité de ces approches réside dans la nature des contenus manipulés, un tweet est différent d'un blog, et de la nature et les propriétés du réseau social exploité.

Dans ce contexte, une des thématiques qui a connu un succès particulier ces dernières années concerne la RI dans les microblogs. La recherche dans les microblogs est devenue un sujet de recherche actif en RI, en particulier après le lancement de la tâche de TREC microblog en 2011 [125]. Des travaux antérieurs, cependant, ont déjà exploré la tâche de recherche des tweets de microblog. *O'Connor* et ses collègues [150] présentent

TweetMotif, une application de recherche exploratoire de Twitter. Contrairement aux approches traditionnelles pour la recherche d'information, qui présentent une simple liste de documents retournés. *TweetMotif* groupe les tweets selon la fréquence des termes significatifs, sous forme d'un ensemble de résultats de sous-thème, qui facilitent la navigation via une interface de recherche de microblog a facettes.

Grâce à la tâche TREC Microblog, de nombreuses approches ont été proposées [6, 40, 99, 142, 200]. Elles exploitent des facteurs temporels (date de publication du tweet), et des facteurs sociaux relatifs au microblogueur. *Metzler* et *Cai* [142] combine un modèle d'apprentissage d'ordonnancement avec le *pseudo-relevance feedback* pour rechercher les 30 tweets les plus pertinents vis-à-vis la requête de l'utilisateur. Leur approche a eu les meilleurs résultats *P*@30 en TREC 2011. *Luo* et ses collègues [130] considère qu'un message de microblog peut être un document structuré, composé non seulement du texte, mais aussi d'autres champs, comme des hashtags, des liens, des retweets et des mentions. L'utilisation de ces informations comme caractéristiques dans un modèle d'apprentissage, apporte une bonne performance de recherche.

La majorité des approches exploitant le réseau social Twitter ont défini des critères de pertinence reflétant l'importance des utilisateurs et des tweets. Ces critères sont : le nombre de tweets d'un auteur, le nombre de fois qu'un utilisateur a été retweeté, le nombre de citations, le nombre d'abonnements, le nombre d'abonnés, etc. Certains travaux ont combiné ces critères linéairement [218, 147, 55], d'autres ont utilisé des techniques d'apprentissage automatique : SVM [105], régression linéaire [66] et RankSVM [48]. *Hong* et ses collègues [90] qualifient le nombre de retweets comme une mesure de popularité de tweet. Ils appliquent des techniques d'apprentissage automatique pour prédire combien de fois les nouveaux messages seront retweetés. Ils ont également exploité d'autres caractéristiques telles que le contenu des messages, des informations temporelles, des méta-données des messages et des utilisateurs et le graphe social de l'utilisateur.

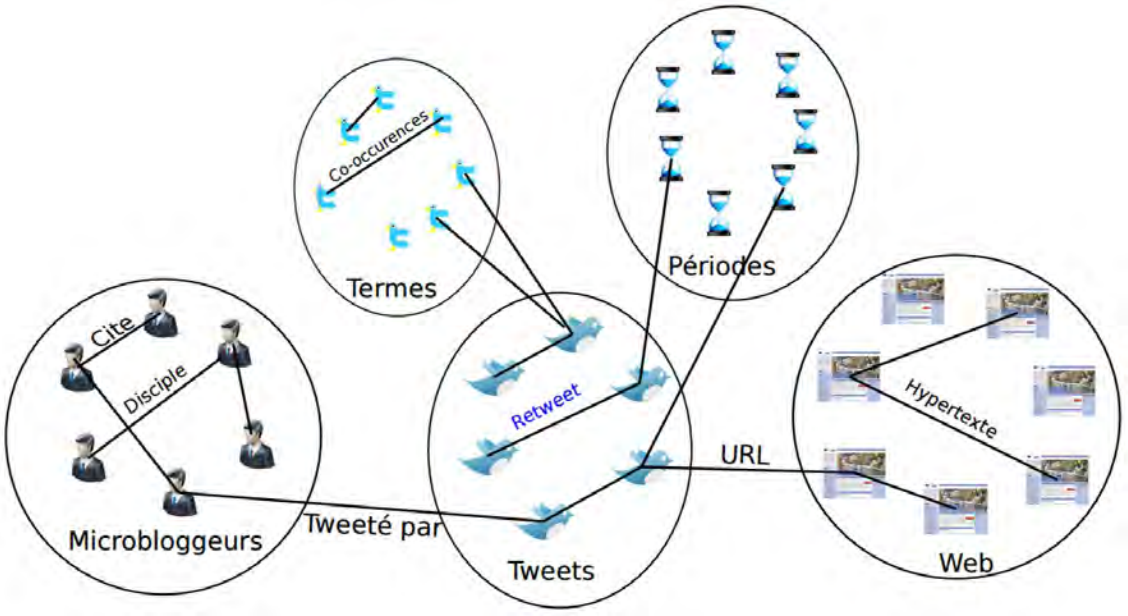

Figure 3.7: Les différents types de liens sur Twitter [54]

D'autres approches liées à des graphes qui représentent les liens sociaux générés au sein des plateformes. Ces graphes représentent différents types de liens comme le montre la figure 3.7 : *utilisateur-utilisateur* dont les liens sont les relations d'amitiés (abonnements ou abonnés ou citation), *utilisateur-tweet* où les liens représentent les statuts des utilisateurs, *tweet-tweet* et dans ce cas les liens représentent les retweets. L'approche présentée par *Yamaguchi* et est ses collègues [209] utilise l'algorithme PageRank pour mesurer l'importance d'un microbloggeur dans un graphe composé par les utilisateurs et les tweets. *Jabeur* et ses collègues [98] utilisent un modèle bayésien pour mesurer la pertinence d'un tweet représenté dans un graphe composé par les termes, les tweets, les utilisateurs et même des périodes temporelles.

3.3.1.2 *Question-Réponse sociale*

Malgré le développement des techniques et des méthodes d'aide à la recherche sur le Web telles que les requêtes de navigation [36] et les requêtes d'auto-complétion, pour aider les utilisateurs à exprimer leurs besoins, de nombreuses requêtes restent sans réponse. *Dror* et ses collègues [65] estiment que cela est dû principalement à deux raisons : (i) l'objectif visé par la requête n'est pas bien exprimé/capturé, et (ii) l'absence de documents pertinents.

Pour remédier à ces problèmes, des systèmes Question-Réponse (*Question/Answering*) ont vu le jour pour connecter les gens et leur permettre de s'entraider pour répondre aux questions. Des exemples de tels systèmes comprennent *Yahoo Answer!*[14], *Answers*[15] et *Aardvark*[16]. En effet, les systèmes Question-Réponse offrent un moyen permettant de répondre à plusieurs types de questions telles que la recommandation, par exemple, la construction d'une nouvelle liste de lecture, *des idées pour de bonnes chansons de fonctionnement ?*, les connaissances factuelles, par exemple, *Est-ce que quelqu'un connait un moyen de mettre les graphiques Excel dans LaTeX ?*, la résolution de problèmes, par exemple, *Comment puis-je résoudre ce problème de distribution de Poisson ?*. Les résultats sont retournés en langage naturel, et sont une réponse immédiate à la question de l'utilisateur, et non pas une liste de sites intéressants ou pertinents. Ces systèmes sont sociaux, parce qu'ils fonctionnent entre un utilisateur qui pose une question et un utilisateur qui pourrait connaître la réponse à cette question, et ils fournissent des fonctionnalités pour interagir et d'évaluer, classer ou de réviser les questions et les réponses. L'utilisateur obtient non seulement des réponses de vrais humains mais a aussi la chance d'entrer en contact avec des utilisateurs ou des experts partageant les mêmes idées dans un certain domaine d'intérêt.

Un des problèmes de base de ces systèmes sociaux réside dans l'absence de garantie sur l'exactitude d'une réponse. Les traces sociales comme le rating des utilisateurs qui ont donné des réponses avant, ainsi que la position de la personne qui donne la réponse dans le cercle social d'un utilisateur, permettent d'évaluer la qualité et la justesse d'une réponse. Un autre inconvénient de ces systèmes asynchrones est qu'on ne sait jamais quand on aura une réponse à notre question ou si on obtiendra une réponse complète. Le fait que les utilisateurs puissent poser des questions et de formuler des réponses en langage naturel n'est pas toujours bénéfique [59]. *Agichtein* [3] et ses collègues tentent

14 https://answers.yahoo.com/
15 http://www.answers.com/Q/
16 En Septembre 2011, Google a annoncé qu'il allait mettre fin à un certain nombre de ses produits, y compris Aardvark.

de résoudre ce problème en proposant un cadre pour identifier le contenu de haute qualité dans les médias sociaux avec leur système de *Yahoo! Answers*. D'autres travaux qui intègrent les réseaux sociaux dans les systèmes Question-Réponse tels que *Hecht* et ses collègues [84] ont développé un système appelé *SearchBuddies* qui répond aux questions publiées sur Facebook avec des résultats de recherche issus des contenus générés par le réseau des amis. Les auteurs notent qu'il y a des défis avec la création de courtes réponses plutôt qu'une liste de résultats, y compris le temps de la répondre.

3.3.1.3 *Recherche de conversations*

Selon *Magnani* [137] une conversation est définie comme étant un arbre où chaque nœud représente un message (par exemple, un tweet posté par un blogueur) à un instant donné en réponse à un nœud parent. Plusieurs travaux proposent de tenir compte des informations sociales offertes par les structures sociales et d'autres informations intrinsèques telles que la longueur de la conversation, pour combiner les messages d'une conversation. *Bhatia* et *Mitra* [24] proposent un modèle probabiliste qui tient compte du contenu thématique des messages, de l'autorité des auteurs, des liens entre conversations et de la taille des conversations. Ces facteurs sont pris en compte comme des connaissances a priori dans le modèle probabiliste. Les résultats montrent que l'incorporation des facteurs sociaux et de la pertinence textuelle améliore les résultats par rapport à la pertinence textuelle seule.

A notre connaissance, seuls les travaux de *Magnani* et *Montesi* [136, 137] et *Ould-Amer* [7] proposent de chercher des conversations dans le microblog Twitter. Les travaux de *Magnani* et *Montesi* proposent une approche de fusion des facteurs suivantes : 1) la pertinence textuelle de la conversation basée sur un modèle vectoriel; 2) la moyenne des tailles des tweets de la conversation; 3) la popularité de la conversation (en fonction du nombre de tweet retweetés); 4) la popularité des auteurs des tweets calculée par la moyenne des nombres de followers de chaque auteur; 5) la densité temporelle des tweets. Une série de fonctions d'agrégation ont été utilisées pour calculer la pertinence des conversations : le maximum, le minimum et la moyenne, sur les différents facteurs. Les auteurs montrent que l'utilisation des facteurs sociaux apporte de bons résultats mais il ne montrent pas le taux d'amélioration de ces facteurs par rapport à la pertinence textuelle seule. *Ould-Amer* et ses collègues [7] proposent un modèle probabiliste permettant d'incorporer à la fois les deux pertinences textuelle et sociale. La pertinence textuelle qu'ils proposent est estimée par un modèle de langue. Ils estiment que les fonctions d'agrégations utilisées par *Magnani* [136, 137] ne fournissent pas une bonne intégration de ces facteurs. Leur contribution majeure réside dans la définition d'un modèle permettant d'étudier la combinaison de ces facteurs. Les résultats obtenus, sur une collection contenant plus de 50000 tweets, avec un ensemble de 15 requêtes tirées pour une partie des campagnes TREC Microblog [125], sont significativement meilleurs par rapport à ceux obtenus par l'approche utilisant le contenu textuel seul ainsi que ceux d'une approche à base de BM25 [7].

3.3.1.4 *Recherche d'opinions*

L'explosion des plates-formes sociales tels que les blogs, les réseaux sociaux, les forums a permis aux utilisateurs d'exprimer leurs opinions sur des événements, produits, services [187, 28]. Ces informations permettent à d'autres personnes d'apprendre d'une

expérience similaire et ensuite prendre une décision précise. Avant de réserver une chambre d'hôtel, par exemple, les utilisateurs souhaitent vérifier les avis sur l'hôtel et le service. Les défis des approches d'extraction d'opinion, est de détecter ces contenus exprimant une opinion sur un sujet donné, puis déterminer la polarité des sentiments associés (sentiment négatif, neutre ou positif). Les propriétés des réseaux sociaux permettent de détecter l'opinion publique et communautaire [154, 91, 56, 100], ainsi que d'identifier l'influence de l'opinion dans le graphe social [149, 191, 204].

3.3.1.5 *Recherche de personnes (experts)*

La recherche dans les réseaux sociaux, permet aux utilisateurs de rechercher d'autres personnes dans le réseau social qui vérifient certains critères de recherche ou possèdent des propriétés sociales particulières dans le graphe social [2]. En cas de réseau social professionnel, par exemple, les utilisateurs ont besoin d'identifier des personnes occupant un poste particulier dans certaines entreprises [134]. De même, les scientifiques expriment leur besoin d'identifier des chercheurs importants dans leur domaine de recherche [60]. D'autres propriétés spécifiques des utilisateurs sont étudiées telles que la popularité [110], l'influence [153] et l'expertise [17, 71], identifier à partir de l'analyse du graphe social. Les approches de recherche de personnes proposent plusieurs modèles basés à la fois sur le descriptive du profil et les relations sociales.

3.3.2 *Exploitation des contenus sociaux pour améliorer la RI*

Cette catégorie consiste à améliorer le processus de la RI classique en utilisant l'information sociale comme une nouvelle source d'évidence qui peut intervenir à différents niveaux. Il existe principalement trois niveaux d'amélioration : (i) l'amélioration de l'index, à savoir la façon dont les documents et les requêtes sont représentés et appariés pour estimer leurs similarités, (ii) la reformulation des requêtes à l'aide de connaissances supplémentaires, à savoir l'expansion de la requête de l'utilisateur, et (iii) le reclassement (*re-ranking*) des documents retournés par un SRI (sur la base du profil d'utilisateur ou d'autres facteurs de pertinence sociale). Dans cette catégorie de RI sociale, nous considérons l'exploitation des contenus sociaux dans ces trois pistes.

3.3.2.1 *Indexation sociale*

Plusieurs travaux de recherche [216, 43, 62, 27] ont indiqué que l'ajout des tags au contenu du document améliore la qualité de la recherche, car ils sont de bons résumés de documents [27], par exemple, l'expansion de document [184, 85]. En particulier, l'information sociale peut être utile pour les documents qui contiennent quelques termes où le processus d'indexation simple ne fournit pas une bonne performance de RI.

Tout au long de notre analyse de l'état de l'art, nous avons remarqué que l'information sociale a été utilisée de deux manières différentes pour l'amélioration de la représentation du document : (i) soit par l'ajout de méta-données sociales au contenu des documents, ou (ii) en personnalisant la représentation des documents, en supposant que chaque utilisateur a sa propre vision sur un document donné.

Par rapport au premier point, certaines approches étudient l'utilisation des méta-données sociales pour enrichir le contenu des documents. Dans [43, 41, 62, 46, 44], les

auteurs indexent le document à la fois avec son contenu textuel et ses contenus sociaux associés (tags et commentaires). Cependant, chaque approche utilise une méthode différente pour pondérer les termes des méta-données sociale, par exemple, TF-IDF pour [43, 46]. Concernant le deuxième point, étant donné un document, chaque utilisateur possède sa propre compréhension de son contenu. Chaque utilisateur emploie son propre vocabulaire pour décrire, commenter et annoter ce document. Par conséquent, la solution est de créer des indexes personnalisés [208, 31].

3.3.2.2 *Reformulation de la requête*

La reformulation de requête est un processus qui consiste à transformer une requête q initiale en une autre requête q'. Cette transformation peut être soit un *raffinement* ou une *expansion*. Le raffinement de requête réduit la requête de telle sorte que l'information inutile soit éliminée, tandis que l'expansion de requête rajoute de nouvelles informations à la requête initiale pour la rendre moins ambiguë et élargir son champ de recherche.

L'information sociale peut ainsi être utilisée pour étendre les requêtes. *Koolen* et ses collègues [120] proposent une approche d'expansion de requêtes utilisant Wikipédia comme collection externe. Ils appliquent ensuite cette approche dans la recherche de livres. D'autres pistes concernant le "Pseudo-Relevance Feedback" à partir de Wikipédia ont été explorées, notamment par l'approche de *Li* et ses collègues [124] qui traitent les requêtes dites "faibles". Ces requêtes ne permettent pas de retourner suffisamment de documents pertinents lors de la première recherche. Cette approche a montré une amélioration de qualité, en particulier sur les premiers documents renvoyés. En outre, *Bao* et ses collègues [18] suggèrent l'utilisation d'un graphe bipartite entre les annotations sociales et les pages Web avec des arêtes indiquant le nombre d'utilisateurs (voir la figure 3.8). En se basant sur cette figure, ils proposent deux algorithmes : le *SocialSimRank (SSR)* et le *SocialPageRank (SPR)*. Le *SocialSimRank* est un algorithme itératif pour évaluer quantitativement la similitude entre deux annotations. Par conséquent, une matrice $(N_A \times N_A; N_A : annotations)$ est créée, elle stocke toutes les pondérations de similarité $S_A = (a_i, a_j)$ entre chaque paire d'annotations. Le *SocialSimRank* est ensuite utilisé comme une forme d'expansion de requête, où des tags similaires sont inclus dans le calcul de similarité entre la requête et le document. D'autres travaux d'expansion de requête exploitent les tags dans l'estimation de la similarité entre la requête et le document [23, 25, 219].

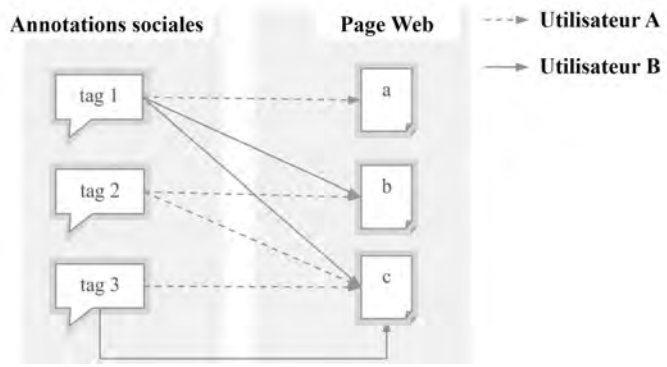

Figure 3.8: Illustration du graphe *SocialSimRank* de *Bao* [18]

3.3.2.3 *Reclassement de résultats*

En RI, le classement des résultats consiste à définir une fonction d'ordonnancement qui permet de quantifier les similarités entre les documents et les requêtes. Nous distinguons deux classes pour le classement des résultats qui diffèrent dans la manière dont elles utilisent l'information sociale. La première classe utilise l'information sociale en intégrant une pertinence sociale au processus de classement, tandis que la seconde utilise l'information sociale pour personnaliser les résultats de recherche.

- **Classement basé sur la pertinence sociale** : plusieurs approches ont été proposées pour améliorer le classement des documents retournés vis-à-vis d'une requête en se basant sur la pertinence sociale. La pertinence sociale se réfère à des facteurs sociaux qui caractérisent un document en termes d'intérêt social, sa popularité, sa réputation, etc.

  En plus de l'algorithme *SocialSimRank* présenté dans las section 3.3.2.2, *Bao* et ses collègues [18] proposent le *SocialPageRank* qui est un algorithme qui calcule la qualité de la page (popularité) par le nombre d'annotations sociales. Pour chaque composant (page Web, l'annotation et l'utilisateur) de la figure 3.8, un PageRank peut être calculé sur la base des liens entre eux. La popularité (PageRank) d'un utilisateur peut être dérivée de la popularité des annotations et la page Web sur laquelle l'utilisateur a effectué une annotation, de même pour la page Web et l'annotation. Ensuite, le PageRank de la page Web est utilisé dans la fonction d'ordonnancement des documents. Enfin, *Yanbe* et ses collègues [210] proposent le *SBRank*, qui indique le nombre d'utilisateurs qui ont marqué une page. Ils utilisent *SBRank* comme une fonction de pertinence dans la recherche Web. D'autres travaux exploitent d'autres types d'informations sociales sont détaillés dans la section 3.4.

- **Classement social personnalisé** : en général, les utilisateurs ont des intérêts différents, des profils différents et des habitudes différentes. Par conséquent, dans un système de RI, en fournissant les mêmes documents classés de la même manière ne convient pas peut être à tous les utilisateurs. Ainsi, une fonction personnalisée pour trier les documents différemment selon chaque utilisateur devrait améliorer les résultats de recherche.

  Plusieurs approches ont été proposées pour personnaliser le classement des résultats de recherche en utilisant l'information sociale [22, 42, 32, 148, 193, 198, 205]. Les travaux se distinguent par l'information sociale qu'ils considèrent pour représenter le profil de l'utilisateur. Une partie dans ces travaux exploitent les tags donnés par les utilisateurs pour construire son profil [148, 205], d'autres les documents qu'il mis en favoris [193] ou les relations sociales [42].

La majorité de ces approches se situent dans le contexte de l'annotation sociale, d'autres critères sociaux, en particulier les signaux sociaux, sont exploités pour améliorer cette tâche. Ces approches sont présentées dans la section 3.4.

3.4 Signaux sociaux pour améliorer la recherche

Dans cette section, nous présentons les travaux connexes exploitant les signaux sociaux comme source d'évidence additionnelle pour mesurer la pertinence a priori de la ressource. Ces travaux peuvent être classés selon deux classes : 1) signaux sociaux indépendants du temps et 2) signaux sociaux dépendants au temps.

3.4.1 *Approches basées sur les signaux sociaux indépendants du temps*

Certaines approches se concentrent sur la façon d'améliorer la recherche d'information en exploitant les actions des utilisateurs et leur réseau social sous-jacent. *Cheng* et ses collègues [49] présentent une étude détaillée sur les caractéristiques des vidéos sur YouTube[17]. Une étude complémentaire destinée à analyser le comportement des utilisateurs dans la recherche de vidéos et les besoins d'information habituels (ex. musique, comédie, sport, etc) est présentée dans [53]. Ces travaux se focalisent sur l'analyse statistique de vidéos sur YouTube en termes de nombre de *vues*, de *j'aime*, de *je n'aime pas*, récence et popularité des vidéos, etc. Cependant, aucun d'entre eux n'a exploité l'apport potentiel de ces caractéristiques dans le processus de recherche.

Ces dernières années, de nouveaux travaux ont vu le jour et se sont concentrés sur l'étude de la richesse et la possibilité d'exploiter ces caractéristiques générées par les utilisateurs. *Chelaru* et ses collègues [44, 45] ont étudié l'impact des signaux sociaux (*aime*, *n'aime pas*, *commentaire*, etc.) sur l'efficacité de la recherche sur YouTube. Ils ont montré, bien que les critères de base basés sur la similarité de la requête avec le titre de la vidéo et les annotations soient efficaces pour la recherche de vidéo, les critères sociaux sont également utiles. Il permettent en effet d'améliorer le classement des résultats de recherche pour 48% des requêtes. Les auteurs ont exploré l'impact des retours sociaux (*social feedback*) sur la recherche des vidéos dans YouTube en utilisant des techniques de sélection d'attribut (*greedy feature selection algorithm*), ainsi que des fonctions d'apprentissage d'ordonnancement issues de l'état de l'art.

D'autres travaux s'intéressent à exploiter les caractéristiques sociales pour améliorer la recherche d'information sur le Web. *Karweg* et ses collègues [108] proposent une approche pour estimer la pertinence d'une ressource Web en combinant un score thématique et un score social basé sur deux facteurs de pertinence sociale : a) l'intensité d'engagement d'un utilisateur pendant une interaction avec un document, mesurée à partir du nombre de *clics*, nombre de *votes*, nombre *d'enregistrement* et de *recommandation*; b) le degré de confiance pour chaque utilisateur estimé à partir de son graphe social en utilisant l'algorithme de PageRank (popularité). Ils montrent que les résultats de l'approche augmentée par les critères sociaux sont plus pertinents pour la majorité des requêtes. En particulier, leur système diminue le temps requis pour le processus de recherche et augmentent la satisfaction des utilisateurs. De façon similaire, *Khodaei* et *Shahabi* [113] proposent une approche de classement basée sur plusieurs facteurs sociaux incluant les relations entre le producteur du document et l'utilisateur qui soumet la requête, ainsi que l'importance des utilisateurs et leurs actions (par exemple, *playcount*: nombre de fois qu'un utilisateur écoute une piste audio sur LastFM[18]) effectuées sur

17 https://www.youtube.com/
18 http://www.lastfm.fr/

les documents. Ils ont mené un ensemble d'expérimentations sur des données issues du site Internet de Radio en ligne last.fm. Les résultats expérimentaux montrent une amélioration significative pour le classement socio-textuel par rapport au classement textuel seul. *Buijs* et *Spruit* [37] proposent une approche sociale appelée *Social Score Method* basée sur plusieurs signaux sociaux issus de différents réseaux sociaux. Le score social est estimé avec un simple comptage des signaux (*partage, bookmark, tweet*), et est combiné avec le TF-IDF. Le score social est utilisé pour déterminer quelles ressources devraient être retournées en premier. Les résultats montrent que l'idée derrière cette approche est une alternative prometteuse aux méthodes existantes (ex. pagerank) afin de déterminer l'importance des pages Web indépendamment de la requête.

Les annotations peuvent être considérées comme une forme de signaux sociaux qui pourraient être pris en compte pour déterminer l'importance relative des pages Web. En général, les signaux sociaux n'attribuent pas des mots ou un *tag* à une ressource. Les signaux sociaux sont peut être moins complexes, comme une simple action *j'aime* sur Facebook. Un *j'aime* indique un vote positif pour une ressource Web alors qu'un *tag* demande plus de traitement pour pouvoir tirer ce genre d'information. En 2007, *Bao* et ses collègues [18] montrent le potentiel des annotations sociales pour déterminer la popularité des pages Web.

Enfin, il existe d'autres études initiées par des chercheurs de Microsoft Bing [145, 166] qui montrent l'utilité de différents contenus sociaux générés par le réseau d'amis de l'utilisateur sur Facebook. *Pantel* et ses collègues [155] ont considéré que les annotations sociales(*j'aime*, *j'aime pas* et *partage*) sont utiles en RI, car ils indiquent qu'une personne dans le réseau social a aimé ou partagé un document qui peut être utile pour d'autres utilisateurs. Ils ont constaté que l'utilisateur peut bénéficier de ces actions sociales de diverses façons, y compris la découverte de recommandations sélectionnées socialement, la recherche personnalisée, estimer la popularité des pages Web, etc. *Kazai* et *Milic-Frayling* [109] ont proposé un modèle de RI sociale qui intègre différents types de vote d'approbation (ex. *ratings*, *reviews*, etc) sur les documents dans une collection de livres. Les approbations reflètent un niveau d'engagement par la communauté liée à la collection (les utilisateurs sur Amazon/LibraryThing), elles peuvent être interprétées comme un indice de confiance, de popularité ou de recommandation. Ces critères sociaux peuvent provenir de différents types d'utilisateurs tels que des experts reconnus et des associations professionnelles, ou à partir d'opinions agrégées d'une communauté plus large représentant une approbation populaire. Ils ont mené des expériences pour intégrer les votes d'approbation dans un modèle de recherche classique, en utilisant une collection contenant 42000 livres et un ensemble de 250 requêtes avec des jugements de pertinence partiellement annotés par des utilisateurs non-experts. Les résultats obtenus en intégrant ces critères sociaux au sein d'un modèle textuel (BM25F), s'avèrent positifs et significatifs.

3.4.2 *Approches basées sur la temporalité des signaux sociaux*

Les travaux présentés dans la section précédente ne prennent pas encompte le moment ou l'action s'est produite et le moment ou la ressource a été publiée. Il existe peu de travaux qui se sont intéressés à ces questions en RI. Ceux qui s'en rapprochent correspondent à ceux réalisés par [93] et [112]. *Inagaki* et ses collègues [93] ont proposé

d'exploiter les caractéristiques de clic (*click through*) en RI. Parmi ces critères, un facteur appelé *ClickBuzz*, qui capte l'intérêt que suscite un document à travers le temps. Ils ont défini le *ClickBuzz* comme une mesure pour déterminer si une page Web reçoit un niveau inhabituel d'intérêt des utilisateurs par rapport au passé. Le *ClickBuzz* est basé sur le nombre de clics sur le document au cours d'un intervalle de temps donné. Cette méthode permet d'exploiter le *feedback* des utilisateurs pour améliorer la qualité des résultats de recherche en favorisant les URL qui ont un intérêt récent pour les utilisateurs. L'utilisation de *ClickBuzz* dans les modèles de classement améliore principalement le *nDCG*@5 avec un taux de 1.57% par rapport aux systèmes de recherche en ligne basés sur le tri par récence des documents retournés vis-à-vis d'une requête.

Les travaux de *Khodaei* et *Alonso* [112] ne se sont pas intéressés à la tâche de recherche d'information. Ils tentent juste d'identifier l'évaluation des intérêts des utilisateurs dans le temps. Ils considèrent, en effet, que la grande masse des contenus générés par les utilisateurs dans les réseaux sociaux offre une occasion d'examiner comment les utilisateurs produisent et consomment ce type de contenu au fil du temps. Ils classent les intérêts sociaux des utilisateurs en cinq classes: "recent", "ongoing", "seasonal", "past" et "random", puis analysent Twitter ainsi que des données de Facebook sur les activités sociales des usagers. Ils discutent également trois solutions différentes où ces signaux sensibles au temps peuvent être appliqués : a) la RI personnalisée; b) la RI basée sur les amis et c) la RI collective.

3.5 Évaluation de la RI Sociale

Comme nous l'avons vu dans le chapitre précédent, l'évaluation en RI se fait principalement à travers les collections de tests, souvent construites dans le cadre de campagnes d'évaluation. La RI sociale ne déroge pas à cette règle, mais le premier défi pour étudier l'impact de l'aspect social sur le processus de recherche d'information reste encore lié à la collection de test.

Malgré la multitude des tentatives pour créer des collections standards[19] permettant d'évaluer les approches sociales, le besoin des collections appropriées augmente de plus en plus. Cependant, le développement de ce genre de collection reste envisageable, surtout avec la mise en place de la tâche Microblog dans la campagne d'évaluation TREC [125], ainsi que la tâche Social Book Search (SBS) dans la campagne d'évaluation d'INEX [29]. D'autres campagnes d'évaluation telles que celles liées la recherche sociale des images a vu le jour avec la campagne d'évaluation de MediaEval [94].

Chaque tâche en RI sociale fait l'objet d'un cadre expérimental relativement particulier. Dans ce qui suit nous allons citer certaines collections standards ainsi que leur tâches correspondantes.

3.5.1 *Les tâches sociales de TREC*

Un atelier TREC est composé d'un ensemble de *Tracks*, ce sont des domaines d'intérêt dans lesquels notamment des tâches de recherches sont définies. Ces tâches servent à démontrer la robustesse et l'efficacité des approches de RI. Comme tâches, nous pouvons citer :

19 http://icwsm.org/2013/datasets/datasets/

- **Microblog Track** : est une campagne d'évaluation pour la recherche de microblog organisée chaque année depuis 2011 en collaboration avec l'atelier TREC. Le but de ce *Track* est de fournir à la communauté de recherche des microblogs un protocole d'évaluation des systèmes de recherche microblog. TREC Microblog comprend une tâche principale ad-hoc, connu comme *real-time ad-hoc search*, et une seconde tâche connue par *filtering track* introduite en 2012. Ces deux tâches sont basées sur le corpus des tweets. En 2011, la collection de text *Tweets2011* contenait environ 16 millions de tweets. L'ensemble de données est construit en utilisant l'API publique *Twitter Stream* qui fournit un échantillon représentatif de 1% du flux des tweets. Ces dernières années, cette collection a évolué et le corpus a été fortement enrichi par de nouveaux tweets. La collection actuelle, connue sous le nom *Tweets2013*, se compose de 243 millions de tweets collectés à partir du flux publique de Twitter entre le 1 Février et le 31 Mars 2013 (inclus) [125].

- **Real-Time Filtering Task** : c'est un nouveau *Track* lancé en 2015 qui consiste à une tâche de filtrage en temps réel visant à contrôler un flux de messages de médias sociaux, conformément au profil d'intérêt d'un utilisateur. La notion de ce *Track* est mise en œuvre en tenant compte de deux modèles de tâches : (i) *Push notifications on a mobile phone*, un contenu qui est identifié comme intéressant par un système basé sur le profil d'intérêt de l'utilisateur peut être affiché à l'utilisateur comme une notification sur son téléphone mobile. L'attente est que ces notifications sont déclenchées dans un temps relativement court après que le contenu soit généré. Il est supposé que les messages de notifications soient relativement courts, (ii) *Periodic email digest*, un contenu qui est identifié comme intéressant par un système basé sur le profil d'intérêt de l'utilisateur peut être agrégé dans un email qui est périodiquement envoyé à un utilisateur. Il est supposé que chaque élément de contenu est relativement court, on pourrait penser des "personalized headlines".

  Dans la tâche *Real-Time Filtering*, les documents sont des tweets. Au cours de la période d'évaluation, les systèmes des participants doivent "listen" le flux de tweets en direct de Twitter et d'identifier les tweets intéressants par rapport aux profils d'intérêt des utilisateurs (l'équivalent de *topics* dans d'autres tâches TREC).

- **Blog Track** : est une campagne d'évaluation pour la detection d'opinion, sa dernière version était en 2010. Ce *Track* utilise une collection appelée *Blogs06*, qui a été créée par l'Université de Glasgow [132]. Les données ont été collectées sur une période de 11 semaines à partir du 6 Décembre 2005 au 21 Février 2006. La collection disponible actuellement représente 148 GB de données [151].

3.5.2 *La tâche sociale de MediaEval*

MediaEval est une campagne d'évaluation consacrée à l'évaluation de recherche d'information multimédias.

La tâche sociale de MediaEval est appelée : *Retrieving Diverse Social Images*. Cette tâche est la suite des éditions des années précédentes [97, 96, 95] et vise à favoriser les nouvelles technologies pour améliorer la pertinence et la diversification des résultats de la recherche en mettant l'accent explicitement sur le contexte des médias sociaux. Cette tâche a été conçue pour les chercheurs travaillant dans l'analyse des médias sociaux, y

compris des domaines tels que : la recherche d'images (texte, visuel, les communautés multimédias), l'ordonnancement, l'apprentissage automatique, le retour de pertinence (*relevance feedback*), le traitement du langage naturel, le crowdsourcing et le géo-tagging automatique.

L'ensemble de données comprend 400 emplacements, allant de ceux très célèbres (par exemple, le *Colosseum of Rome*) au moins connu du grand public (par exemple, le *Palazzo delle Albere*). Pour chaque emplacement, ils fournissent une liste de classement des photos de différentes qualités récupérées en utilisant le nom ou les coordonnées GPS de l'emplacement à travers les plateformes de médias sociaux (par exemple, Flickr : 100-150 résultats par emplacement). Pour servir des informations de référence, chaque emplacement est accompagné d'une photo représentative et une description de l'emplacement de Wikipedia. Pour encourager la participation des groupes de différents domaines de recherche, des ressources supplémentaires telles que les descripteurs visuels et des modèles de localisation textuelle seront fournis pour l'ensemble de la collection. Pour répondre à la tâche, les participants sont libres d'envisager d'utiliser d'autres sources de données externes, telles que les ressources de l'Internet. Au total, l'ensemble de données sera constitué d'environ 44000 photos. L'ensemble de données doit être divisé en un ensemble de développement (par exemple, autour de 50 emplacements à être utiliser pour les méthodes d'apprentissage et un ensemble de test pour l'évaluation finale) [94].

La collection 2015 se compose d'un ensemble de données (*devset*) contenant 153 requêtes de localisation (45375 photos Flickr - collection de 2014 [97]). Un ensemble de 300 locations et 685 utilisateurs (différents que ceux de *devset* et *testset*) et un ensemble de test (*testset*) contenant 139 requêtes : 69 concepts de requêtes liés à la localisation (20700 photos Flickr) et 70 multi-concepts de requêtes liés à des événements et des états associés à des endroits (20694 photos Flickr) [94].

3.5.3 *La tâche de Social Book Search*

La tâche de Social Book Search (SBS) étudie la recherche de livres dans des scénarios de recherche via une requête de l'utilisateur ou de recommandation. L'objectif est de modéliser et développer des techniques pour aider les utilisateurs dans les tâches de recherche de livres. La tâche de Social Book Search se compose de deux principaux *tracks* : *Interactive Track* et *suggestion track*.

Dans notre cas, nous nous intéressons à *suggestion track* qui vise à explorer des techniques pour faire face aux besoins d'informations complexes, qui vont au-delà de la pertinence thématique et peuvent inclure des aspects comme : genre, récence, engagement, utilité et la qualité de la rédaction, et aux sources d'information complexes qui incluent les profils d'utilisateur, les catalogues personnels et les descriptions du livre contenant les méta-données professionnelles et le contenu généré par l'utilisateur [119, 29].

La collection INEX SBS se compose de 2.8 millions de documents. Chaque document décrit un livre d'Amazon, étendu avec des méta-données sociales de LibraryThing. Chaque livre est un fichier XML représenté avec des champs comme *isbn*, *title*, *review*, *summary*, *rating* and *tag*. La liste complète des champs est indiquée dans le tableau 3.2.

La collection SBS fournie 208 requêtes ainsi que leurs jugements de pertinence fournies par INEX[20] (voir chapitre 3.5). Chaque requête a été dénichée sur LibraryThing

20 http://social-book-search.humanities.uva.nl/#/data

nom du champ			
book	similarproducts	title	imagecategory
dimensions	tags	edition	name
reviews	isbn	dewey	role
editorialreviews	ean	creator	blurber
images	binding	review	dedication
creators	label	rating	epigraph
blurbers	listprice	authorid	firstwordsitem
dedications	manufacturer	totalvotes	lastwordsitem
epigraphs	numberofpages	helpfulvotes	quotation
firstwords	publisher	date	seriesitem
lastwords	height	summary	award
quotations	width	editorialreview	browseNode
series	length	content	character
awards	weight	source	place

Tableau 3.2: Liste des différents champs d'un document SBS.

pour une liste de livres et se compose de cinq champs : *title*, *mediated_query*, *narrative*, *example* and *group*. Par la présente, le champ *narrative* est la description textuelle de la requête dont le champ *mediated_query* est dérivé manuellement. En outre, le champ *example* contient une liste de livres que l'utilisateur a mentionnés comme des exemples positifs ou négatifs [121]. Les évaluations de pertinence sont basées sur les suggestions réelles à la requête originale sur le forum LibraryThing. Les valeurs de pertinence sont pondérées à l'aide d'un arbre de décision qui inclut des informations fiables délivrées par un utilisateur ayant lu le livre. Les requêtes SBS 2015 sont un sous ensemble des requêtes utilisées en 2014. Cependant, les évaluations de pertinence ont été étendus avec des suggestions de livres supplémentaires qui ne figurent pas en SBS 2014 [121].

3.6 Limites et positionnement

La majorité des travaux de recherche cités précédemment se sont focalisés sur l'exploitation des annotations sociales (tags). Mais peu de travaux se sont intéressés aux signaux sociaux tels que *j'aime*, *partage*, *commentaire*, *+1*, etc. Un des problèmes des annotations sociales est que le niveau d'activité des utilisateurs semble suivre une distribution en loi de puissance. Par conséquent, la plupart des annotations sont créées par une petite quantité d'utilisateurs. On ignore encore comment gérer ce genre de ressources car il peut présenter des données biaisées [219]. Nous avons également constaté que même les travaux sur les signaux sociaux s'intéressent uniquement à l'exploitation locale de ses signaux (utiliser les UGCs de Twitter pour la RI dans Twitter, utiliser les signaux de Youtube pour la recherche dans Youtube, etc).

Nos travaux se différencient de l'état de l'art sur les points suivants. Nous proposons plusieurs approches avec différentes intuitions : a) Exploitation des signaux sociaux pris en compte individuellement et groupés sous formes de propriétés (ex. réputation et popularité) dans un modèle de RI; b) Nous étudions également la temporalité de ses signaux par rapport à la date de création de chaque action sociale ainsi que l'âge de la ressource (document); c) Nous introduisons un facteur supplémentaire appelé *diversité* qui estime la diversité des signaux au sein d'un document. Nous notons que dans les travaux de l'état de l'art la diversité a été appliquée uniquement au contenu thématique du document [8] [163].

Dans les chapitres suivants, nous allons aborder le cœur de notre travail en présentant l'ensemble de nos contributions dont chacune fait l'objet d'un chapitre.

Partie III

EXPLOITATION DES SIGNAUX SOCIAUX

Le contenu est roi, mais l'engagement est reine et maître.

— Mari Smith

4

EXPLOITATION INDIVIDUELLE ET GROUPÉE DES SIGNAUX

Introduction

Les systèmes de recherche d'information exploitent dans leur majorité deux classes de sources d'évidences pour trier les documents répondant à une requête. La première, la plus exploitée, est dépendante de la requête, elle concerne toutes les caractéristiques relatives à la distribution des termes de la requête dans le document et dans la collection. La seconde classe concerne des facteurs indépendants de la requête, elle mesure une sorte de qualité ou d'importance a priori du document. Parmi ces facteurs, on en distingue le PageRank [152], la localité thématique du document [57], la présence d'URL dans le document [201], ses auteurs [133], etc.

Il est bien connu aujourd'hui que les informations sociales prolifèrent sur les réseaux sociaux. En 2015 des statistiques montrent que parmi les 3 milliards d'internautes, soit 39% de la population mondiale, 74% sont au moins inscrits sur un réseau social[1]. Ces utilisateurs contribuent activement sur les réseaux sociaux, par exemple, sur Facebook chaque 60 secondes environ 50000 publications sont faites et plus de 2.3 millions de *j'aime* sur différentes ressources, ce qui engendre une masse de données d'environ 410 Go par seconde[2]. En effet, grâce aux outils proposés par le Web 2.0 les utilisateurs interagissent de plus en plus entre eux et/ou avec les ressources. Ces interactions (signaux sociaux), traduites par des *j'aime*, des *+1*, des *partages*, des *tweets*, des *commentaires* ou des *bookmarks* associés aux ressources, peuvent être considérées comme une des sources que l'on peut également exploiter pour mesurer l'intérêt a priori de la ressource en termes de *popularité* et de *réputation*, indépendamment de la requête.

Dans ce chapitre, nous décrivons notre approche pour l'exploitation des signaux sociaux laissés par les utilisateurs sur les ressources pour mesurer la pertinence (l'intérêt) a priori d'une ressource. Cette connaissance a priori est combinée avec la pertinence thématique dans un modèle de recherche qui prend en compte explicitement ces sources d'évidence. Nous évaluons la performance de notre approche sur deux types de collection, IMDb (Internet Movies Database) et SBS (Social Book Search), enrichies de plusieurs données sociales collectées à partir de plusieurs réseaux sociaux.

4.1 Hypothèses et questions de recherche

Notre objectif dans le cadre de nos travaux recherche est de définir des approches permettant l'exploitation des signaux sociaux comme facteurs de pertinence qui peuvent jouer un rôle pour améliorer la recherche d'information. Plus précisément, nous pensons à un schéma de RI sociale structuré de la façon suivante :

1 http://www.blogdumoderateur.com/etude-kpcb-internet-2015/
2 http://www.blogdumoderateur.com/facebook-q2-2015/

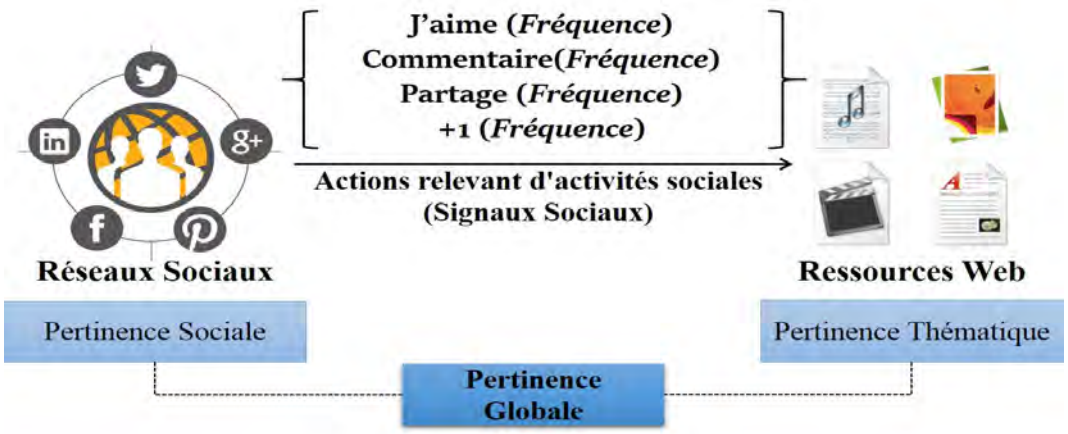

Figure 4.1: Pertinence a priori basée sur les signaux sociaux

Comme le montre la figure 4.1, nous avons un ensemble d'utilisateurs de différents réseaux sociaux qui interagissent avec des ressources Web en les *partageant*, en cliquant sur *j'aime* ou à travers d'autres types de signaux. Ces signaux sont caractérisés par leur fréquence sur le document. Des propriétés sociales telles que la *popularité* et la *réputation* devraient être extraites et quantifiées à partir de ces signaux sociaux. Ces signaux et propriétés peuvent être considérés comme une information additionnelle qui peut jouer un rôle pour mesurer l'importance a priori de la ressource. Donc, la pertinence thématique des ressources Web sera bonifiée par l'importance sociale de ces ressources. Par conséquent, les questions de recherche auxquelles nous souhaitons répondre dans ce chapitre sont les suivantes :

1. Est ce que les signaux sociaux peuvent être des critères de pertinence ?
2. Quelles sont les signaux sociaux utiles pour évaluer la pertinence a priori d'une ressource ?
3. Comment traduire les signaux sociaux en propriétés sociales ?
4. Quel modèle théorique pour combiner la pertinence a priori d'une ressource et sa pertinence thématique ?
5. Impact des signaux sociaux dans les performances d'un système de recherche d'information ?

4.2 Approche de RI exploitant les signaux sociaux

L'approche que nous proposons consiste à estimer l'importance sociale d'une ressource en exploitant ses signaux sociaux associés, soit individuellement où chaque signal représente un facteur de pertinence, soit en regroupant ces signaux en fonction du type d'importance sous-jacent. En effet, certains signaux sont liés à la *popularité* et d'autres liés à la *réputation*.

Afin de prendre en compte ces facteurs sociaux dans l'évaluation de pertinence, nous nous appuyons sur un modèle de langue qui nous permet de combiner de manière

élégante l'importance a priori de la ressource et sa pertinence vis-à-vis de la requête. Avant de décrire le modèle, nous décrivons quelques notations que nous utilisons tout au long de ce manuscrit.

4.2.1 *Préliminaires et notations*

L'information sociale que nous exploitons dans notre approche peut être représentée par le quadruplet <U, R, A, RS> où U, R, A, RS sont des ensembles finis d'instances : *Utilisateurs*, *Ressources*, *Actions* et *Réseaux sociaux*.

4.2.1.1 *Ressources*

Nous considérons une collection $R = \{D_1, D_2, ...D_n\}$ de n ressources. Une ressource D peut être un document traditionnel comme une page Web ou une ressource Web 2.0 comme une vidéo ou toute autre entité similaire. Nous supposons qu'une ressource D peut être représentée à la fois comme un ensemble de mots-clés textuels (bag of words), soit $D_w = \{w_1, w_2, \dots w_z\}$ où w est un terme, et comme un ensemble de caractéristiques sociales réalisées sur cette ressource, $D_a = \{a_1, a_2, \dots a_m\}$ où a est une action relevant d'activité sociale.

4.2.1.2 *Actions*

Cet un ensemble, $A = \{a_1, a_2, ...a_m\}$, représente m actions (signaux sociaux) que les utilisateurs peuvent effectuer sur les ressources. Ces actions représentent la relation entre l'ensemble des utilisateurs $U = \{u_1, u_2, ...u_h\}$ et l'ensemble des ressources R. Par exemple sur Facebook, les utilisateurs peuvent effectuer des actions relevant d'activités sociales comme : *publier*, *aimer*, *partager* ou *commenter*.

4.2.1.3 *Réseaux sociaux*

Il existe un ensemble $RS = \{rs_1, rs_2, ...rs_z\}$ de z réseaux sociaux (graphe). Chaque réseau social spécifique contient un ou plusieurs signaux sociaux spécifiques réalisés sur une ressource D.

4.2.2 *Modèle de langue et probabilité a priori*

Nous nous appuyons sur un modèle de langue [160] pour combiner la pertinence thématique de la ressource vis-à-vis de la requête et son importance sociale, modélisée elle aussi comme une probabilité a priori. La probabilité qu'une ressource D soit pertinente par rapport à une requête Q est estimée selon la formule 4.1 suivante :

$$RSV(Q,D) = P(D|Q) \overset{\text{rank}}{=} P(D) \cdot P(Q|D) = P(D) \cdot \prod_{w_i \in Q} P(w_i|D) \tag{4.1}$$

Où w_i représente les mots de la requêtes Q.

$P(D)$ représente la probabilité a priori du document D, son utilité est de modéliser et intégrer d'autres sources d'évidence indépendantes de la requête dans le processus de

la recherche d'information. L'estimation de $P(w_i|D)$ peut être effectuée en utilisant différents modèles (ex. Jelineck Mercer, Dirichlet) [215]. En effet, la principale contribution de ce chapitre est sur l'estimation de $P(D)$ en exploitant les signaux sociaux.

Pour estimer la probabilité a priori de la ressource $P(D)$, nous avons plusieurs options. Soit nous considérons chaque signal individuellement, dans ce cas, nous considérons autant de probabilités que de signaux. Chaque $P(D)$ mesure l'impact d'un signal donné. Soit nous cumulons et calculons l'effet conjoint de l'ensemble des signaux observés sur une ressource. Une troisième option consiste à regrouper les signaux selon la propriété sous-jacente, nous pensons que les signaux sociaux indiquent un certain engagement des utilisateurs et ont une signification différente. Un signal de type *j'aime* n'a pas le même impact qu'un signal de type *commentaire*. Par conséquent, nous proposons de combiner les signaux en fonction des propriétés qu'ils pourraient représenter. Nous estimons alors la probabilité a priori du document selon le groupe signaux de cette propriété.

4.2.2.1 *Propriétés sociales*

Nous avons analysé la nature de différents types d'actions à travers plusieurs réseaux tels que Facebook, Twitter, etc. Cette analyse nous a permis de définir deux propriétés sociales : a) la *popularité* de la ressource et b) la *réputation* de la ressource. Nous pensons qu'il existe des signaux qui reflètent et mesurent d'avantage la *réputation* de la ressource ainsi que d'autres mesurent la *popularité*. Nous définissons les deux propriétés comme suit :

- *Popularité* : la popularité désigne la chose la plus connue (populaire) par le public. Grâce à l'influence des pairs, des ressources cibles peuvent rapidement monter dans la façon dont elles sont omniprésentes dans la société. Ainsi, la *popularité* d'une ressource peut être estimée en fonction de l'intensité de *partage* de cette ressource entre les utilisateurs à travers les réseaux sociaux. Nous supposons qu'une ressource est populaire si elle a été partagée et commentée par plusieurs utilisateurs dans plusieurs réseaux sociaux.

- *Réputation* : nous supposons que si l'indice de *popularité* d'une ressource est important, cette ressource est relativement intéressante. Cependant, la *popularité* d'une ressource ne reflète pas forcément sa bonne ou mauvaise *réputation*. La *réputation* d'une ressource est une opinion que l'on a sur cette ressource. Nous pensons que l'estimation de cette propriété peut être calculée à partir des actions relevant d'activités sociales qui portent un sens positif telles que le *j'aime* de Facebook ou le *bookmark* d'une ressource comme favoris sur Delicious. En effet, la *réputation* d'une ressource dépend du degré d'appréciation des utilisateurs sur les réseaux sociaux.

4.2.2.2 *Estimation des probabilités a priori*

Une manière simple d'estimer la probabilité a priori est d'effectuer un simple comptage du nombre d'actions spécifiques effectuées sur une ressource. En supposant que les actions sont indépendantes les unes des autres, la formule générale est la suivante :

$$P_x(D) = \prod_{a_i^x \in A} P_x(a_i^x) \tag{4.2}$$

$P_x(a_i^x)$ est estimée en utilisant le maximum de vraisemblance :

$$P_x(a_i^x) = \frac{Count(a_i^x, D)}{Count(a_\bullet^x, D)} \tag{4.3}$$

Pour éviter une probabilité nulle, nous lissons $P_x(a_i^x)$ par la collection R en utilisant Dirichlet [215]. La formule 4.3 devient comme suit :

$$P_x(D) = \prod_{a_i^x \in A} \left(\frac{Count(a_i^x, D) + \mu \cdot P(a_i^x|R)}{Count(a_\bullet^x, D) + \mu} \right) \tag{4.4}$$

$P_x(a_i^x|R)$ est estimée en utilisant le maximum de vraisemblance :

$$P_x(a_i^x|R) = \frac{Count(a_i^x, R)}{Count(a_\bullet^x, R)} \tag{4.5}$$

Avec :

- x se réfère de manière générale à la propriété sociale {*popularité*, *réputation*} estimée à partir d'un ensemble d'actions spécifiques. On peut imaginer d'autres regroupement de signaux, alors dans ce cas x correspondra au regroupement considéré, par exemple des signaux venant d'un réseau social spécifique (*TotalFacebook* qui regroupe le *j'aime*, le *partage* et le *commentaire*). Dans le cas ou les signaux sociaux sont pris en compte de manière individuelle, le x sera ignoré.
- $P_x(D)$ représente la probabilité a priori de D.
- $Count(a_i^x, D)$ représente le nombre d'occurrence de l'action a_i^x dans D.
- $Count(a_i^x, R)$ représente le nombre d'apparition de l'action spécifique a_i^x dans la collection R.
- $Count(a_\bullet^x, Y)$ représente le nombre total de signaux sociaux dans Y (Y est soit le document D ou la collection R).

Prise en compte du rating : le *rating* est un autre signal que nous traitons différemment car il n'est pas un simple comptage d'actions tel que décrit ci-dessus. Le *rating* est une note sur une échelle de 1 à une valeur max de 5, où 3 signifie "moyen" et 5 signifie "excellent", et dépend du nombre des utilisateurs qui notent le document.

Nous proposons d'intégrer le signal *rating* comme une mesure de *réputation* d'un document. A cet effet, nous utilisons la moyenne bayésienne [211] (Bayesian Average (BA)) des notes (*ratings*) pour estimer la moyenne des *rating* dans un document en prenant en compte le nombre d'utilisateurs qui ont noté le document. Plus les utilisateurs notent le même document, la moyenne devient plus fiable et moins sensibles aux valeurs aberrantes. Les documents qui ont de nombreuses évaluations sont amplifiées par rapport aux documents qui ont peu de notes. De même, les documents avec des notes élevées sont amplifiées plus que des documents avec de faibles notes. Donc, la BA d'un document est calculée comme suit :

$$BA(D) = \frac{moy(r) \cdot |r| + \sum_{D' \in R} moy(r') \cdot |r'|}{|r| + \sum_{D' \in R} |r'|}, \tag{4.6}$$

Avec :

- $r = \{\{r_i\}\}$ un multi-ensemble de valeurs des *ratings*, avec $i = 1, nr$. nr est le nombre de *ratings* associé au document D, $nr = |r|$. r_i est le ième *rating* donné par l'utilisateur i au document D.
- moy est la fonction de moyenne des *ratings* r du documents D.
- r' est l'ensemble des *ratings* dans toute la collection R.

Nous notons que la prise en compte du logarithmique du $BA(D)$ permet de compresser le score et réduit, de ce fait, l'impact des critères sur le score global.

$$P(a_i = rating) = \frac{\log(1 + BA(D))}{\log(1 + \sum_{D' \in R} BA(D'))} \tag{4.7}$$

Pour les documents sans notes cela se traduirait par une probabilité nulle. Afin d'éviter une multiplication par zéro ainsi que la pénalisation du score textuel, nous utilisons la méthode de lissage Add-One :

$$P(a_i = rating) = \frac{1 + \log(1 + BA(D))}{1 + \log(1 + \sum_{D' \in R} BA(D'))}. \tag{4.8}$$

4.2.2.3 *Combinaison des probabilités a priori*

Dans notre proposition, nous disposons de diverses sources d'informations sociales qui influencent la probabilité a priori de pertinence. Cette probabilité est calculée par la combinaison de plusieurs propriétés sociales (*popularité* et *réputation*). De manière générale, le problème peut être formalisé comme suit [156] :

$$P_{P \oplus R}(D) = P_P(D) \cdot P_R(D) \tag{4.9}$$

Avec :

- $P_P(D)$, $P_R(D)$ définissent les probabilités a priori relatives à la *popularité* P et la *réputation* R.
- $P_{P \oplus R}(D)$ définit la combinaison des probabilités a priori.

Nous notons que le $P(a_i = rating)$ est un facteur qui quantifie la propriété *réputation*.

4.3 Expérimentations et résultats

Pour évaluer notre approche, nous avons mené une série d'expérimentations sur deux principales collections de test IMDb (Internet Movie Database) et SBS (Social Book Search). Nous avons comparé notre approche qui combine des connaissances a priori du document avec sa pertinence thématique, aux modèles de référence basés uniquement sur le contenu textuel. Nos objectifs expérimentaux sont :

1. d'abord, évaluer si les signaux sociaux issus de différents réseaux sociaux améliorent la RI.
2. ensuite, évaluer l'impact des signaux pris individuellement et groupés pour représenter certaines propriétés (*popularité* et *réputation*).
3. et enfin, mesurer la corrélation entre les signaux et la pertinence des documents.

4.3.1 *Collections de documents*

A notre connaissance, hormis les collections TREC Microblog[3] contenant des données Twitter uniquement, il n'existe pas à ce jour de collections de test standard pour l'évaluation de l'efficacité des signaux sociaux en recherche d'information. Pour répondre à ce besoin, nous avons proposé un enrichissement des collections existantes telles que IMDb (Internet Movies Database) et SBS (Social Book Search), par des données sociales issues de différents célèbres réseaux sociaux. Le choix de ces deux collections se justifie par des raisons techniques. Afin de collecter les signaux sociaux pour chaque document, il est primordial d'avoir les URLs des documents pour pouvoir les extraire à travers les APIs des réseaux sociaux qui prennent en argument ces URLs. Cet enrichissement des collections standard présente un avantage important par rapport aux expérimentations. Nous n'avons pas besoin de créer les jugements de pertinence des documents manuellement. Entre autres, grâce aux Qrels fournies par ces standards, on a la possibilité d'évaluer plusieurs configurations (selon les signaux collectés), qui seraient bien trop coûteux en temps et en personnes dans le cadre d'une expérimentation où c'est à nous de créer les Qrels. Dans ce qui suit, nous présentons les collections que nous avons exploitées.

4.3.1.1 *INEX Internet Movies Database 2011*

Chaque document de la collection INEX IMDb décrit un film, et est représenté par un ensemble de méta-données, listées dans le tableau 4.2. Pour chaque document, nous avons collecté les données sociales via leur API correspondante sur les cinq réseaux sociaux (Facebook, Twitter, Google+, Delicious et LinkedIn) listés dans le tableau 4.4. Nous les avons introduits dans la balise UGC (User Generated Content). Ce champ n'a pas été indexé. Seuls les champs ayant un statut indexé dans le tableau4.2 ont été effectivement indexés. Nous avons aussi utilisé 30 requêtes parmi les requêtes d'INEX IMDb[4] (voir le tableau 4.3). Dans ce cas, la tâche est devenue plus facile car nous n'avons pas besoin de faire une évaluation manuelle pour obtenir les jugements de pertinence, nous utilisons les Qrels fournies par INEX IMDb 2011.

Le tableau 4.1 montre 4 exemples de données sociales pour 4 documents d'IMDb. L'URL du document est donnée par la syntaxe suivante : *www.imdb.com/title/{id}/*. Par exemple, le document ayant le Id=*tt1730728* correspond au film *The Sea Is All I Know* dont son URL est *www.imdb.com/title/tt1730728/*, possède 31 actions de *j'aime*, 11 *partage*, 2 *commentaire*, 2 *tweet* et 0 action de *bookmark* et de *partage* sur LinkedIn.

3 http://trec.nist.gov/data/microblog.html
4 https://inex.mmci.uni-saarland.de/tracks/dc/2011/

	Facebook			Google+	Delicious	Twitter	LinkedIn
Id	J'aime	Partage	Commentaire	+1	Bookmark	Tweet	Partage
tt1730728	31	11	2	0	0	2	0
tt1922777	14763	13881	22914	341	12	2859	14
tt0372784	3990	3308	2363	134	0	1787	13
tt0145487	1319	1183	588	49	0	471	3

Tableau 4.1: Exemple de documents IMDb ayant des données sociales

Champ	Description	Statut
ID	Identifiant du film (le document)	-
Title	Le titre du film	Indexé
Year	L'année de sortie du film	Indexé
Rated	Classement des films selon le type du contenu	-
Released	Date de réalisation du film	Indexé
Runtime	Durée du film	Indexé
Genre	Genre de film (Action, Drame, etc.)	Indexé
Director	Le directeur du projet du film	Indexé
Writer	Les écrivains et les scénaristes du film	Indexé
Actors	Les acteurs principaux du film	Indexé
Plot	Résumé textuel du film	Indexé
Poster	Le lien URL de l'affiche du film	-
url	Le lien URL qui mène à la source originale du document	-
UGC	Les différentes données sociales	-

Tableau 4.2: Liste des différents champs indexés d'un document IMDb

Le tableau 4.4 présente des statistiques sur le nombre de signaux sociaux dans les 1000 documents retournés par chaque requête IMDb 2011 (il y a en tout 30 requêtes IMDb). Selon les moyennes des nombres de signaux dans les documents, nous remarquons que la densité des signaux de Facebook (en moyenne : 85.8 j'aime, 94.1 partage et 98.4 commentaire) est très élevée par rapport aux autres signaux (en moyenne : 2.5 +1, 0.9 bookmark, 17.2 tweet et 1.4 partage (LinkedIn)).

Réseaux Sociaux	Facebook			Google+	Delicious	Twitter	LinkedIn
Signaux Sociaux	J'aime	Partage	Commentaire	+1	Bookmark	Tweet	Partage
Minimum	0	0	0	0	0	0	0
Maximum	76842	43918	62281	1475	986	12223	299880
Total	2478498	2718918	2845169	73392	26143	499232	42787
Moyenne	85.8027	94.1258	98.4964	2.5407	0.9050	17.2830	1.4812

Tableau 4.4: Statistiques sur le nombre de signaux sociaux dans les documents retournés par les 30 requêtes

Requête	Description	Narration
action biker	search for all action movies with bikers in it.	As i like action movies, specially if bikers are in it, i like to get a list of all these movies.
ancient Rome era	find the movies about the era of ancient Rome.	I am interested in the movies about era of ancient Rome. I am looking for movies talking stories in the era of ancient Rome.
true story drugs +addiction -dealer	find movies about drugs (drug addiction but not drug dealers) that are based on a true story.	I am working with teens and I want to show them a movie about drugs that is based on a true story. A relevant movie is any true story based movie about drug use and addiction. Movies about drug dealers are not relevant. I would like to see as much information as possible about the movie in order to decide whether the movie is appropriate or not.

Tableau 4.3: Exemple de requêtes d'évaluation INEX IMDb

Le tableau 4.5 présente des statistiques sur le nombre de documents (1000 documents retournés par chaque requête IMDb, il y a en tout 30 requêtes) contenant ou pas des signaux sociaux. Par exemple, le nombre de documents qui contiennent le signal *j'aime* (la ligne "AVEC signaux" et colonne "J'aime" dans le tableau) est de 16903 documents sur 30000 documents retournés par les 30 requêtes.

Réseaux Sociaux	Facebook			Google+	Delicious	Twitter	LinkedIn
Signaux Sociaux	J'aime	Partage	Commentaire	+1	Bookmark	Tweet	Partage
AVEC signaux	16903	18656	13001	5259	3256	12390	3724
SANS signaux	13097	11344	16999	24741	26744	17610	26276

Tableau 4.5: Statistiques sur le nombre de documents (retournés par les 30 requêtes) contenant ou pas des signaux sociaux

4.3.1.2 *INEX Social Book Search*

Pour chaque document de la collection SBS (nous l'avons décrit dans le chapitre 3 section 3.5), nous avons collecté des données sociales via l'API de Facebook, soient les signaux *j'aime*, *partage* et *commentaire* rajoutés aux signaux existants extraits d'Amazon/LT (voir le tableau 4.6). Nous les avons mises dans la balise UGC (User Generated Content). Ce champ n'a pas été indexé. La liste complète des champs indexés est indiquée dans le tableau 4.9. Nous avons utilisé les 208 requêtes ainsi que leur jugements de pertinence fournies par INEX SBS track 2015[5].

5 http://social-book-search.humanities.uva.nl/#/data

	Facebook			Amazon		
Id	J'aime	Partage	Commentaire	Review	Rating	Tag
0553583859	128	48	10	279	4.5	13
0441014100	440	59	165	1561	4	17
0316545015	0	2	0	234	4.5	28
0765308843	4	15	13	59	4	07

Tableau 4.6: Exemple de documents SBS ayant des données sociales

Le tableau 4.6 montre un exemple de données sociales pour des documents SBS. L'URL du document est donnée par la syntaxe : *http://www.amazon.com/gp/product/{id}/*. Par exemple, le document ayant le Id=*0553583859* correspond au livre *Fields of Fire* dont l'URL est *http://www.amazon.com/gp/product/0553583859/*, possède 128 *j'aime*, 148 *partage* et 10 *commentaire* issus de Facebook, ainsi que 279 *review*, 4.5 en moyenne de *rating* et 13 de *tag* issus d'Amazon/LT.

Le tableau 4.7 présente des statistiques sur le nombre de signaux sociaux dans les documents retournés par les 208 requêtes de SBS 2015 (1000 documents par requête). Nous remarquons que la densité des signaux de Facebook est très élevée par rapport aux signaux d'Amazon/LibraryThing mais le nombre total de *rating* et de *review* est largement supérieur par rapport aux autres signaux.

Source	Facebook			Amazon		
Signaux Sociaux	J'aime	Partage	Commentaire	Review	Rating	Tag
Minimum	0	0	0	0	0	0
Maximum	7213	5892	5975	3378	3378	277
Total	4760698	5222491	5465011	23856118	23856118	1811968
Moyenne	41.3082	45.3152	47.4195	206.9981	206.9981	15.7223

Tableau 4.7: Statistiques sur le nombre de signaux sociaux dans les documents SBS retournés par les 208 requêtes

Le tableau 4.8 présente des statistiques sur le nombre de documents (1000 documents retournés par chaque requête SBS, il y a en tout 208 requêtes) contenant ou pas des signaux sociaux. Par exemple, le nombre de documents qui contiennent le signal *partage* (la ligne "AVEC signaux" et colonne "partage" dans le tableau) est de 32467 documents sur 115248 documents retournés par les 208 requêtes.

Source	Facebook			Amazon		
Signaux Sociaux	J'aime	Partage	Commentaire	Review	Rating	Tag
AVEC signaux	28668	32467	22702	82093	82093	44424
SANS signaux	86580	82781	92546	33155	33155	70824

Tableau 4.8: Statistiques sur le nombre de documents SBS (retournés par les 208 requêtes) contenant ou pas des signaux sociaux

Champ	Statut	Champ	Statut	Champ	Statut	Champ	Statut
book	Indexé	similarproducts	-	title	Indexé	imagecategory	-
dimensions	-	tags	Indexé	edition	Indexé	name	Indexé
reviews	Indexé	isbn	Indexé	dewey	-	role	Indexé
editorialreviews	Indexé	ean	Indexé	creator	Indexé	blurber	-
images	-	binding	Indexé	review	Indexé	dedication	Indexé
creators	Indexé	label	Indexé	rating	-	epigraph	Indexé
blurbers	-	listprice	-	authorid	-	firstwordsitem	Indexé
dedications	Indexé	manufacturer	Indexé	totalvotes	-	lastwordsitem	-
epigraphs	Indexé	numberofpages	-	helpfulvotes	-	quotation	Indexé
firstwords	Indexé	publisher	Indexé	date	Indexé	seriesitem	-
lastwords	Indexé	height	-	summary	Indexé	award	-
quotations	Indexé	width	-	editorialreview	Indexé	browseNode	-
series	Indexé	length	-	content	Indexé	character	-
awards	-	weight	-	source	Indexé	place	Indexé

Tableau 4.9: Liste des différents champs indexés d'un document SBS.

4.3.1.3 *Quantification des propriétés sociales*

Les propriétés sociales ont été identifiées en analysant différents signaux à travers plusieurs réseaux sociaux. Nous avons tout d'abord réparti les signaux en fonction de la propriété qu'ils sont censés représenter. Nous associons des signaux sociaux pour chaque propriété comme suit (voir le tableau 4.10) :

Propriétés	c_i	Signaux sociaux	Source	Collection
Popularité	c_1	Nombre de *Commentaires*	Facebook	SBS, IMDb
	c_2	Nombre de *Tweets*	Twitter	IMDb
	c_3	Nombre de *Partages(LIn)*	LinkedIn	IMDb
	c_4	Nombre de *Partages*	Facebook	SBS, IMDb
	c_5	Nombre de *Reviews*	Amazon/LibraryThing	SBS
	c_6	Nombre de *Tags*	Amazon/LibraryThing	SBS
Réputation	c_7	Nombre de *J'aimes*	Facebook	SBS, IMDb
	c_8	Nombre de *Mentions +1*	Google+	IMDb
	c_9	Nombre de *Bookmarks*	Delicious	IMDb
	c_{10}	Moyenne de *Ratings*	Amazon/LibraryThing	SBS

Tableau 4.10: Liste des signaux sociaux exploités dans la quantification

La répartition des signaux est effectuée selon leur nature et signification, qui correspondent à la définition que nous avons donnée en section 4.2.2.1. Dans le tableau 4.10, nous remarquons que les signaux sociaux estimant la *réputation* portent des opinions positives, par exemple, *marquer* un lien d'une ressource par un utilisateur sur Delicious signifie que ce lien a été rajouté à sa liste des favoris. Quand l'utilisateur clique sur le *j'aime*, *+1* ou *rating*, cela indique qu'il a apprécié le contenu de cette ressource. Donc la présence de cet ensemble de signaux sociaux dans une ressource augmente le degré de *réputation* de cette ressource. De même pour la *popularité*, les signaux sociaux exploités pour estimer cette dernière, nous permettent de savoir la position en termes de tendance et propagation de cette ressource sur le Web.

4.3.1.4 *Métriques d'évaluation*

Les métriques que nous avons utilisées sont : la précision (P@k, MAP) et le nDCG (Normalized Discounted Cumulative Gain) qui sont aussi suggérées par la campagne d'évaluation INEX SBS 2015. Pour plus de détails voir chapitre 2 section 2.4.2.

4.3.1.5 *Modèles de référence*

Chaque document des trois collections a été indexé en fonction des mots clés se trouvant dans les balises ayant le statut indexé dans les tableaux 4.2 et 4.9, en utilisant le moteur de Lucene Solr[6]. L'indexation est classique, nous avons utilisé le *EnglishAnalyzer*, qui filtre les mots grammaticaux en utilisant l'algorithme de racinisation Porter.

6 http://lucene.apache.org/solr/

Dans les expérimentations nous sélectionnons les 1000 premiers documents retournés par requête, pour les collections d'INEX IMDb et SBS.

Nous avons comparé nos approches à trois modèles de référence :

- *Lucene Solr*[7] : c'est un moteur de recherche populaire développé par *Apache Software Foundation* qui est basé sur le modèle vectoriel et la pondération du terme TF-IDF.
- *Modèle de langue (ML.Hiemstra)* [87] : il désigne un modèle d'appariement de recherche d'information classique qui utilise le principe de génération de la requête par le document. Le modèle de langue est utilisé dans notre cas pour calculer le score basé sur le contenu textuel.
- *Modèle BM25*[8] [171] : le modèle le plus utilisé est Okapi BM25 [171].

4.3.2 *Étude de corrélation des signaux sociaux*

Avant d'évaluer nos approches en termes de performances, nous avons tout d'abord analysé la distribution des signaux sociaux dans les documents, afin de vérifier s'il y a une corrélation entre la présence de signaux sur les documents et leur pertinence. Nos objectifs dans cette étude sont :

- réduire la plage d'incertitude associée à l'aboutissement de notre hypothèse sur l'utilité des signaux sociaux pour la détection des documents pertinents. Les études de corrélation nous aide à faire des prévisions relativement fiable.
- déterminer les signaux sociaux et les propriétés sociales qui sont en corrélation avec la pertinence, ainsi que faciliter l'interprétation des résultats.
- identifier les signaux redondants.
- la corrélation ne signifie pas la causalité [111] c-à-d qu'une étude de corrélation ne peut pas prouver définitivement une hypothèse de causalité, mais elle peut en exclure une ou en motiver une autre.

4.3.2.1 *Corrélation entre les signaux sociaux et la pertinence*

Selon une étude en Juin 2014 par Searchmetrics[9], les signaux sociaux représentent 5 des 6 facteurs les plus fortement corrélés avec les résultats de recherche Google. En outre, l'enquête BrightEdge[10] réalisée en 2013 a révélé que 84% des marketeurs de recherche disent que les signaux sociaux tels que le *j'aime*, *tweet*, et +1 sont plus important (53%) à leur référencement Web par rapport aux années précédentes.

En 2015, les signaux sociaux continuent de devenir de plus en plus un facteur très corrélé avec les résultats de Google. Bien que nous ne voyions pas beaucoup d'études scientifiques concernant ces signaux, quelques organismes de marketing tel que Searchmetrics continuent à les analyser. Selon la dernière étude menée par Searchmetrics [139]

7 https://lucene.apache.org/core/4_0_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity
8 https://lucene.apache.org/core/4_0_0/core/org/apache/lucene/search/similarities/BM25Similarity.html
9 http://www.searchmetrics.com/knowledge-base/ranking-factors-infographic-2014/
10 http://www.marketingcharts.com/uncategorized/social-signals-increasingly-important-to-seo-20695/

en 2015, les corrélations des signaux sociaux avec les classements sont pratiquement inchangées par rapport à 2014 et reste à un niveau élevé. Les premiers résultats retournés par Google ont plus de signaux sociaux, ce facteur augmente de façon exponentielle dans les premières places. La figure 4.2 montre les résultats 2015 de la corrélation entre les signaux sociaux et les résultats de Google.

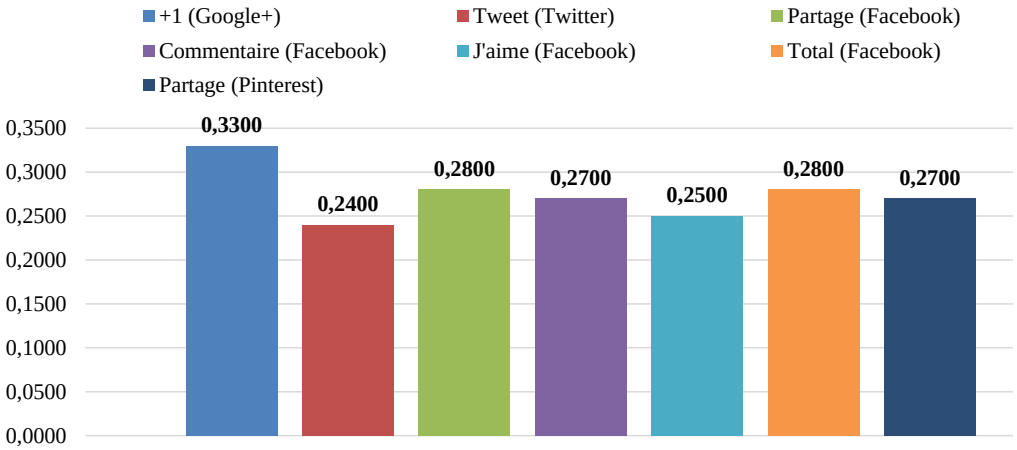

Figure 4.2: Corrélation des critères sociaux et les résultats de Google

Pour notre part, nous avons mesuré la corrélation entre les documents pertinents retournés par Lucene (pour les 30 requêtes d'IMDb et les 208 requêtes de SBS) et leur pertinence issue des Qrels. Nous avons utilisé le coefficient de corrélation de *Rho* Spearman [30], qui permet de vérifier l'existence d'une liaison entre deux variables. Dans notre cas, les deux variables sont le signal social et la pertinence. Si la valeur du *Rho* est positive, nous pouvons dire qu'il existe une certaine corrélation entre les deux variables [111]. La valeur de *Rho* est comprise entre $[-1, +1]$, plus le *Rho* de Spearman est proche de 1, plus la relation est forte et vice-versa. [30].

Les figures 4.3 et 4.4 présentent les valeurs de corrélations entre les signaux sociaux (individuellement et groupés en propriété) et la pertinence des documents IMDb et SBS, respectivement.

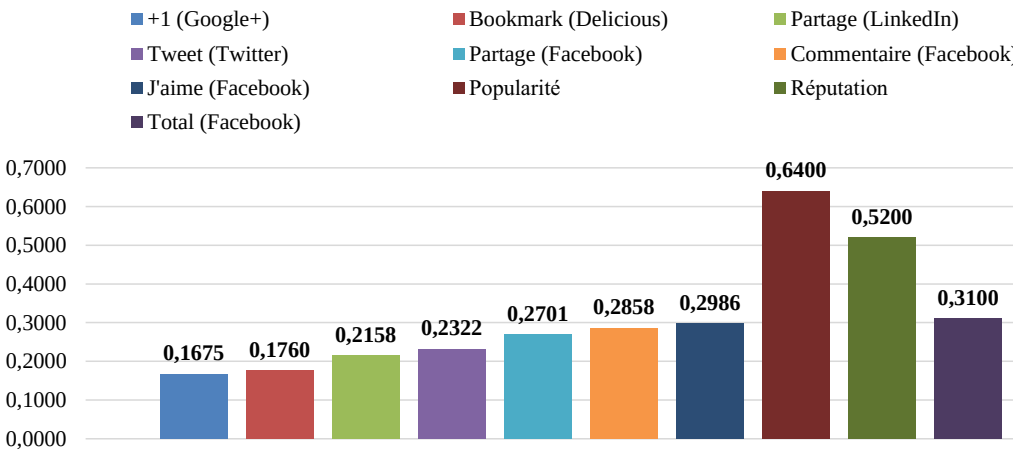

Figure 4.3: Corrélation des critères sociaux avec la pertinence sur la collection INEX IMDb

La figure 4.3 montre que le *j'aime* (0.29) a la plus forte corrélation parmi les signaux individuel, suivi par le nombre de *commentaire* (0.28). D'autres signaux enregistrent une corrélation positive incluant le nombre de *partage* (0.27) et le nombre de *tweets* (0.23). Cependant, les groupes de signaux sous forme de propriétés (*popularité* et *réputation*) ainsi que le total des signaux Facebook sont les plus corrélés à la pertinence.

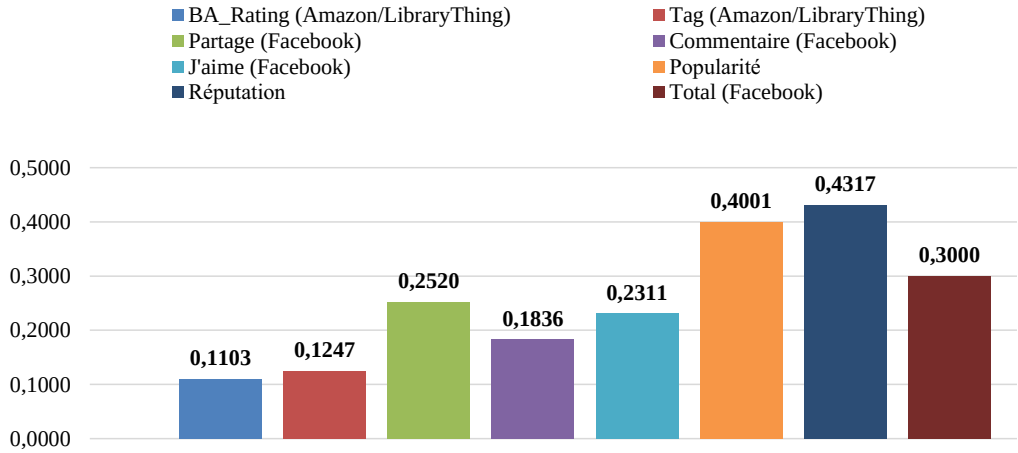

Figure 4.4: Corrélation des critères sociaux avec la pertinence sur la collection INEX SBS

Contrairement à la première étude, la figure 4.4 montre que le *partage* dépasse le *j'aime* et présente la plus forte corrélation (0.25) parmi les signaux individuels, suivi par le nombre de *j'aime* (0.23). Les autres signaux, *commentaire*, *tag* et *rating*, enregistrent, respectivement, les valeurs de corrélation 0.1836, 0.1247 et 0.1103. Concernant le groupement des signaux, les résultats restent largement meilleurs que ceux obtenus par les signaux considérés individuellement. Cependant, la *réputation* génère la corrélation la plus élevée par rapport à la *popularité* et le total des signaux de Facebook.

Enfin, l'analyse de corrélation montre que tous les signaux sociaux sont positivement corrélés avec la pertinence. Ce constat renforce nos hypothèses et soutient l'étude de causalité des signaux sur la détection des documents pertinents présentée dans la section 4.3.3.

4.3.2.2 *Corrélation entre les signaux sociaux deux à deux*

Pour examiner l'existence potentielle d'une relation (corrélation) entre les signaux sociaux pris deux à deux, nous calculons le chevauchement (corrélation) entre les listes de documents retournés (les top-1000) par requête en exploitant chacun des 2 signaux. On fait ensuite une moyenne des corrélations sur l'ensemble des requêtes.

Les deux tableaux 4.11 et 4.12, nous fournissent les scores de *Rho* Spearman qui sont normalisés selon la plage [0,1], où 0 signifie un classement complètement différent et 1 signifie un classement égal entre le couple des signaux. La diagonale inférieure des tableaux présente la corrélation des signaux sociaux basé sur le classement des documents de toutes les requêtes. Nous constatons que, les classements top-1000 fournies par les paires des signaux sociaux suivantes : (*tweet*, *partage*(LIn)), (*bookmark*, *tweet*) et (+1, *bookmark*) sont fortement corrélés, à savoir, les scores de similarité de ces paires sont supérieurs à 0,70 (voir tableau 4.11). Ces corrélations entre ces couples de signaux impliquent une certaine redondance, l'exploitation conjointe de ces signaux ne

serait probablement pas utile. Ces résultats seront confirmés dans les expérimentations menées dans la section 4.3.4.

4.3.2.3 *Corrélation et causalité*

Les études et analyses menées dans les sections précédentes donnent des explications sur une éventuelle corrélation entre les signaux sociaux (l'importance à priori d'une ressource) et la pertinence. Mais, la question de l'intérêt de ces signaux dans le processus de RI reste posée. Nous distinguons en effet les deux notions, corrélation et causalité.

- *La corrélation* est une sorte de constat, on examine des documents retournés selon une classe de pertinence, et on analyse les caractéristiques des documents pertinents. Par exemple, on peut s'apercevoir qu'il y a une corrélation entre les documents pertinents et le fait qu'ils aient beaucoup d'actions relevant d'activités comme le *j'aime* sur Facebook. On constate donc ce fait : il existe une corrélation entre le nombre de *j'aime* sur Facebook et les documents pertinents sur le système de recherche. Mais rien ne dit que ces documents sont pertinents parce qu'ils ont beaucoup de *j'aime*.

- *La causalité* est l'étude des causes directes d'un événement. Par exemple, on indexe une collection contenant un nombre important de documents, puis on utilise un modèle de recherche basique et un autre avec le même modèle basique mais qui prend en compte des critères sociaux. Le second modèle retourne des résultats meilleurs que le premier en termes de pertinence, donc on peut dire qu'il existe une causalité entre la sélection des documents pertinent et les critères sociaux prises en compte.

Donc pour résumer, avec la corrélation, on constate un fait, avec la causalité, on analyse la cause directe de ce fait, ce qui est très différent. Les études que nous allons présenter dans les sections suivantes prouvent de façon rigoureuse la causalité entre l'exploitation des signaux sociaux et la détection des documents pertinents.

Dans la section suivante, nous présentons une étude approfondie sur l'intérêt des facteurs sociaux pour améliorer les résultats d'une recherche d'information. Nous allons en effet nous baser sur les modèles de langue afin de modéliser ces critères sociaux sous forme de probabilité a priori $P(D)$.

Signaux sociaux	J'aime	partage	Commentaire	Partage (LIn)	Tweet	Bookmark	+1
J'aime	1						
Partage	0.61	1					
Commentaire	0.31	0.26	1				
Partage (LIn)	0.35	0.41	0.40	1			
Tweet	0.32	0.28	0.39	**0.77**	1		
Bookmark	0.34	0.48	0.51	0.31	**0.76**	1	
+1	0.34	0.61	0.40	0.32	0.30	**0.71**	1

Tableau 4.11: Corrélation deux à deux entre les signaux sociaux sur INEX IMDb

Signaux sociaux	J'aime	Partage	Commentaire	Rating	Tag
J'aime	1				
Partage	0.59	1			
Commentaire	0.27	0.22	1		
Rating	0.11	0.13	0.09	1	
Tag	0.08	0.09	0.12	0.57	1

Tableau 4.12: Corrélation deux à deux entre les signaux sociaux sur INEX SBS

4.3.3 *Évaluation de notre approche*

Dans cette section, nous avons mené des expériences avec des modèles basés uniquement sur le contenu textuel des documents (le modèle de Lucene Solr, le modèle de langue Hiemstra [87] et BM25 sans la probabilité a priori), ainsi que des approches combinant le contenu textuel et les caractéristiques sociales prises en compte individuellement et groupées sous forme de propriétés sociales. Nous notons que les meilleures valeurs de μ appartiennent à l'intervalle $\mu \in [90, 100]$ pour INEX IMDb et à l'intervalle $\mu \in [2400, 2500]$ pour INEX SBS. Les résultats obtenus sont présentés dans les tableaux 4.13 et 4.14. Concernant les paramètres de BM25, $b = 0.75$ et $k_1 = 1.2$.

4.3.3.1 *Résultats et discussions*

Nos résultats sont présentés comme suit :

1. *Prise en compte individuelle des signaux sociaux* : les tableaux 4.13 et 4.14 partie (a) montrent les résultats obtenus sur les deux collections en considérant les signaux sociaux séparément. Un astérisque indique que la différence avec les modèles textuels est statistiquement significative selon le test de Student [75].
2. *Combinaison de plusieurs signaux sociaux* : Les résultats listés dans les tableaux 4.13 et 4.14 partie (b) nous permettent de comprendre le groupement des signaux sous forme de propriétés sociales. Un double astérisque indique que la différence avec le modèle textuel (*ML.Hiemstra*) et la prise en compte individuelle des signaux est statistiquement significative selon le test de Student [75].

Modèles	P@10	P@20	nDCG	MAP
Base (A) : Sans Signaux Sociaux				
BM25	0.3500	0.3371	0.4113	0.2068
Lucene Solr	0.3411	0.3122	0.3919	0.1782
ML.Hiemstra	0.3700	0.3403	0.4325	0.2402
a) Prise en Compte des Signaux Individuellement				
J'aime	0.3938*	0.3620*	0.5130*	0.2832*
Partage	0.4061*	0.3649*	0.5262*	0.2905*
Commentaire	0.3857*	0.3551*	0.5121*	0.2813*
Tweet	0.3879*	0.3512*	0.4769*	0.2735*
+1	0.3826*	0.3468*	0.5017*	0.2704*
Bookmark	0.3730	0.3414	0.4621	0.2600
Partage (LIn)	0.3739	0.3432	0.4566	0.2515
b) Prise en Compte des Signaux sous Forme de Propriétés				
TotalFacebook	0.4209**	0.4102**	0.5681**	0.3125**
Popularité	0.4319**	0.4264**	0.5801**	0.3221**
Réputation	0.4405**	0.4272**	0.5900**	0.3260**
TousLesCritères	0.4408**	0.4262**	0.5974**	0.3300**
ToutesLesPropriétés	0.4629**	0.4509**	0.6203**	0.3557**

Tableau 4.13: Résultats de P@k, nDCG et MAP sur la collection INEX IMDb

Modèles [Run SBS'15]	P@10	P@20	nDCG	MAP
Base (A) : Sans Signaux Sociaux				
BM25 [Run1]	0.0601	0.0517	0.1581	0.0540
Lucene Solr	0.0528	0.0487	0.1300	0.0463
ML.Hiemstra	0.0607	0.0559	0.1620	0.0527
a) Prise en Compte des Signaux Individuellement				
J'aime	0.0857*	0.0689*	0.1864*	0.0741*
Partage	0.0902*	0.0711*	0.1900*	0.0872*
Commentaire	0.0799*	0.0678*	0.1807*	0.0701*
BA_Rating [Run3]	0.0730*	0.0559*	0.1748*	0.0620*
Log_Tag [Run4]	0.0770*	0.0531*	0.1742*	0.0610*
b) Prise en Compte des Signaux sous Forme de Propriétés				
TotalFacebook	0.0958**	0.0810**	0.1937**	0.0892**
Log_Tag + BA_Rating [Run5]	0.0753*	0.0548*	0.1800*	0.0620*
Popularité	0.0964**	0.0780**	0.1953**	0.0890**
Réputation	0.0972**	0.0801**	0.1974**	0.0897**
TousLesCritères	0.0973**	0.0787**	0.1981**	0.0900**
ToutesLesPropriétés	0.1021**	0.0888**	0.2004**	0.0923**
c) Meilleure configuration de la tâche SBS *suggestion* 2015				
MIIB Run6	0.1262**	0.1083**	0.2257**	0.1050**

Tableau 4.14: Résultats de P@k, nDCG et MAP sur la collection INEX SBS

Rang	Run	nDCG@10	MRR	MAP	R@1000	Apprentissage
1	Run6	0.186	0.394	0.105	0.374	Oui
3	Run2	0.130	0.290	0.074	0.374	Oui
8	Run5	0.095	0.235	0.062	0.374	Non
10	Run3	0.094	0.237	0.062	0.374	Non
11	Run4	0.094	0.232	0.061	0.375	Non
21	Run1	0.082	0.189	0.054	0.375	Non

Tableau 4.15: Résultats officiels à SBS'15. Les runs sont triés selon leur nDCG@10

Les deux tableaux 4.13 et 4.14 récapitulent les résultats de précision@k pour $k \in \{10, 20\}$ et de nDCG, ainsi que la MAP. La partie Base(A) liste les résultats des modèles de référence (sans prise en compte des signaux). Nous avons évalué notre approche à travers des configurations différentes, en prenant en compte les signaux sociaux séparément et groupés sous forme de propriétés sociales (*popularité* et *réputation*). Afin de vérifier si les résultats obtenus sont statistiquement significatifs par rapport au modèle de base, nous avons effectué le test de Student [75]. Nous attribuons un astérisque * (forte signification par rapport à *ML.Hiemstra*) et un double astérisque ** (très forte

signification par rapport à *ML.Hiemstra*) lorsque *p-value*$<$ 0.05 et *p-value*$<$ 0.01, respectivement. Le tableau 4.15 récapitule nos résultats officiels de SBS 2015 évalués en utilisant le nDCG@10, MRR (Mean Reciprocal Rank), MAP and R@1000 (Recall), tandis que nDCG@10 est la mesure d'évaluation officielle.

Nous remarquons dans tous les cas, avec la prise en compte des caractéristiques sociales, les résultats obtenus sont significativement meilleurs que ceux obtenus par les modèles de base. Nous discutons dans ce qui suit les résultats de chacune des configurations que nous avons étudié.

- **Intérêt des signaux pris en compte individuellement** : nous constatons que la prise en compte de chaque signal social individuellement améliore les résultats par rapport aux modèles de base. Selon le test de Student, la majorité des résultats montrent une amélioration statistiquement significative. Nous remarquons aussi que la nature du signal impacte les résultats. Selon les tableaux 4.13 et 4.14, les deux signaux *j'aime* et *partage* issus de Facebook apportent les meilleurs résultats par rapport aux autres signaux avec une amélioration de (19%, 22% sur IMDB) et (15%, 17% sur SBS), respectivement, en termes de nDCG par rapport au modèle *ML.Hiemstra*. Le *commentaire* arrive en troisième position avec une amélioration de 18% sur IMDb et 12% sur SBS en termes de nDCG par rapport au modèle *ML.Hiemstra*. Les signaux *tweet* et *+1* apportent relativement le même taux de précision par rapport au *commentaire*, cependant, le *+1* dépasse le *tweet* en termes de nDCG avec 5%. Le reste des signaux sur IMDb, *bookmark* de Delicious et *partage* de LinkedIn, ont un faible impact (statistiquement non significative). Concernant les signaux issus d'Amazon, le *tag* et le *rating*, améliorent significativement les résultats vis-à-vis du modèle textuel *ML.Hiemstra*. Le *tag* dépasse le *rating* en termes de *P*@10 de 5%, cependant, le *rating* est meilleur que le *tag* avec des taux d'amélioration de 1% en nDCG et 2% en MAP. Nous notons que les configurations *Log_Tag+BA_Rating* (équipe : MIIB, Run5), *BA_Rating* (Run4) et *Log_Tag* (Run3) ont été classées 8ème, 10ème et 11ème parmi 47 configurations, respectivement, dans la compagne d'évaluation SBS 2015.

- **Intérêt des propriétés** : les probabilités a priori basées sur les propriétés améliorent considérablement les résultats en termes de nDCG par rapport au modèle textuel *ML.Hiemstra* (*popularité* : 34%, 21% et *réputation* : 36%, 22% sur IMDb et SBS, respectivement), ainsi que par rapport à la prise en compte des signaux individuellement. La *popularité* dépasse le *partage* avec des améliorations de 10%, 3% sur IMDb et SBS, respectivement. La *réputation* surpasse le *j'aime* avec des améliorations de 15%, 6% sur IMDb et SBS, respectivement. De plus, la *réputation* apporte de meilleurs résultats par rapport à la *popularité*, on constate une amélioration légère en termes de nDCG de 2% sur IMDb et 1% sur SBS. Ces résultats sont justifiés par le fait que les signaux qui quantifient la *réputation* peuvent être considérés comme un engagement de l'utilisateur qui fournit son approbation explicitement. Par exemple, les ressources ayant des signaux plus positives (comme le *j'aime*, *rating*$>$3 et +1) sont plus susceptibles d'intéresser les utilisateurs que ceux qui ne possèdent pas ces signaux. Si plusieurs utilisateurs ont constaté que la ressource est utile, alors il est plus probable que d'autres utilisateurs puissent trouver ces ressources utiles aussi. Les signaux sociaux qui quantifient la *popularité* ne représentent pas des votes d'approbation, comme par exemple le *commentaire*

peut être positif ou négatif, mais ils représentent des facteurs de tendance et une mesure de la propagation de l'information. Par conséquent, une information populaire suscite toujours l'intérêt de l'utilisateur. L'autre point est que la combinaison des signaux sociaux à partir de divers réseaux sociaux offre un jugement collectif (*wisdom of crowds*) plus réaliste de la notoriété des ressources et renforce la confiance et la crédibilité. Par conséquent, le regroupement des signaux sociaux en fonction de leur sens et de leur nature, où certains signaux sont liés à la *popularité* et d'autres liés à la *réputation* est la solution la plus efficace par rapport à la prise en compte individuelle des signaux pour améliorer la RI.

- **Intérêt de la combinaison globale** : nous comparons également les différentes combinaisons des signaux, soient tous les critères *TousLesCritères*, en considérant tous les signaux associés à une ressource, ou bien en prenant uniquement ceux provenant d'un seul réseau social (*TotalFacebook*), et une dernière qui concerne la combinaison de signaux regroupés en propriétés (formule 4.4). Les combinaisons globales des signaux (nommé *TousLesCritères* dans les tableaux) et des deux propriétés (*ToutesLesPropriétés*) apportent les meilleurs résultats par rapport aux autres configurations (modèles de base, signaux et propriétés). Cependant, les meilleurs résultats de cette expérimentation sont obtenus par la configuration *ToutesLesPropriétés* avec des taux d'amélioration (+25% P@10 et +43% nDCG sur IMDb) et (+68% P@10 et +25% nDCG sur SBS) par rapport au modèle de langue *ML.Hiemstra*. Une conclusion intéressante vient de la façon dont tous les critères sont combinés. Nous remarquons que la combinaison des propriétés (nommé *ToutesLesPropriétés*) conduit à de meilleurs résultats (+4% nDCG sur IMDb et +5% sur SBS) que la combinaison de tous les critères (nommé *TousLesCritères* dans les tableaux). Ceci montre en effet qu'il est plus efficace d'appliquer un lissage sur les propriétés sociales que sur les signaux sociaux.

- **Meilleure configuration de la tâche SBS *suggestion* 2015**[11] : ce travail décrit dans [92] à été mené en collaboration avec *M. Imhof*. Nous avons utilisé l'algorithme d'apprentissage *random forests* [35] pour évaluer non seulement les différentes méta-données textuelles, mais aussi les critères non textuels. En particulier, nous avons exploité le *prix* et le *nombre de pages* d'un livre à l'égard de la préférence de l'utilisateur ainsi que les *ratings* de chaque livre. Nous avons supposé qu'un utilisateur qui ne dispose que de petits livres dans son catalogue signifie qu'il préfère les livres courts. En outre, nous avons exploité la note (*rating*) moyenne d'un livre comme un critère supplémentaire dans l'algorithme *random forests*. Pour permettre à l'algorithme d'intégrer l'importance de la note moyenne, nous avons considéré aussi le nombre de notes.

  Nous pouvons voir dans le tableau 4.14 partie (c) que la configuration nommée *MIIB Run6* en utilisant le *random forests* dépassent de loin l'efficacité des autres configurations n'utilisant aucun apprentissage des données, ainsi que les 46 différentes configurations proposées par les participants dans la tâche *SBS suggestion 2015*[12]. Au cours de nos expérimentations, nous avons constaté que la prise en compte de ces trois critères non textuels dans l'apprentissage contribue à aug-

11 http://social-book-search.humanities.uva.nl/#/results15
12 http://social-book-search.humanities.uva.nl/#/results15

menter la *nDCG*, ce qui signifie que ces critères améliorent la détection des documents pertinents, en particulier, le *rating*.

Dans la section suivante, nous présentons des expérimentations plus approfondies relatives à l'exploitation d'approches supervisées basées sur la sélection d'attributs et l'apprentissage automatique. Nous allons en effet nous baser sur les techniques de sélection d'attributs afin de détecter les meilleurs critères qui reflètent la pertinence et qui sont susceptibles d'être utiles par rapport à d'autres dans la recherche d'information en utilisant l'apprentissage automatique.

4.3.4 *Évaluation et approches basées sur l'apprentissage*

Le problème des méthodes précédentes réside dans la détermination de l'importance de chaque signal social par rapport à l'autre. Tant que nous n'avons pas évalué toutes les combinaisons pondérées de chaque signal, il est impossible de juger convenablement et de façon rigoureuse l'ordre d'importance de ces critères de pertinence, et d'identifier les meilleurs d'entre eux.

L'objectif principal de cette étude est d'identifier les meilleures critères sociaux de pertinence, ainsi que vérifier si la sélection d'attributs (*j'aime, partage, commentaire, bookmark, tweet, +1, partage(LIn), tag* et *rating*) améliore les résultats de recherche en termes de pertinence. Dans cette étude nous allons aborder deux questions :

- Quels sont les signaux les plus importants dans la tâche de RI vis-à-vis les algorithmes de sélection d'attributs ?
- Quel est l'impact de ces signaux, en termes de performance en RI, quand ils sont exploités par des techniques d'apprentissage.

4.3.4.1 *Étude d'importance des signaux sociaux*

Afin de mieux comprendre l'impact réel des différents signaux sociaux, nous avons évalué l'impact de chacun d'eux en utilisant des algorithmes de sélection d'attributs. Le but est de déterminer les meilleurs signaux à exploiter dans le modèle de recherche d'information. Les algorithmes de sélection d'attributs visent à identifier et supprimer le maximum d'information inutile, redondante et non pertinente en amont d'un processus à base d'apprentissage [81]. Ils permettent également de sélectionner de manière automatique les sous ensembles de critères de pertinence permettant d'avoir les meilleurs résultats. Nous avons utilisé Weka[13], un outil open-source écrit entièrement en Java et qui rassemble un bon ensemble de techniques d'apprentissage et des techniques de sélection d'attributs.

Nous avons procédé ainsi : les premiers 1000 documents pour chaque requête des deux collections (30 requêtes d'IMDb et 208 requêtes de SBS) ont été restitués avec le modèle par défaut de Lucene Solr. Nous avons ensuite calculé pour chaque document les probabilités a priori relatives à chaque signal social. Nous avons identifié les documents pertinents et les documents non pertinents selon les Qrels. L'ensemble obtenu contient 30000 documents pour IMDb et 115248 documents pour SBS, dont :

13 http://www.cs.waikato.ac.nz/ml

- 2765 documents IMDb pertinents et 2953 documents SBS pertinents.
- 27235 documents IMDb non pertinents et 112295 documents SBS non pertinents.

Les classes de ces ensembles sont déséquilibrées, or lorsque le nombre d'éléments d'une classe dans une collection d'apprentissage dépasse considérablement les autres échantillons des autres classes, un classifieur tend à prédire les échantillons de la classe majoritaire et peut ignorer complètement les classes minoritaires [212]. Pour cette raison, nous avons appliqué une approche de sous-échantillonnage (en réduisant le nombre d'échantillons qui ont la classe majoritaire) pour générer des collections équilibrées composées de :

- 2765 documents pertinents et 2765 documents non pertinents pour IMDb.
- 2953 documents pertinents et 2953 documents non pertinents pour SBS.

Les documents non pertinents pour cette étude ont été sélectionnés de manière aléatoire. Enfin, nous avons appliqué les algorithmes de sélection d'attributs sur les deux ensembles obtenus.

La sélection d'attributs a pour but de supprimer l'information redondante ou non pertinente qui n'apporterait aucun bénéfice à la classification des données et rendrait celle-ci plus difficile. Nous présentons dans ce qui suit un aperçu sur l'ensemble de ces méthodes :

- *CfsSubsetEval* : CFS, abréviation de "Correlation-based Feature Subset Selection", c'est un algorithme de filtrage simple qui classe les sous ensembles d'attributs selon une fonction d'évaluation heuristique basée sur la corrélation. Cette méthode utilise le coefficient de corrélation de Pearson entre deux variables v_1 et v_2 [213]. Le biais de la fonction d'évaluation tend vers des sous ensembles qui contiennent des attributs qui sont fortement corrélés avec la classe et non corrélés entre eux [80].
- *WrapperSubsetEval* : évalue les ensembles d'attributs en utilisant un modèle d'apprentissage. Une validation croisée est utilisée pour estimer la précision du système d'apprentissage pour un ensemble d'attributs [118].
- *ConsistencySubsetEval* : cette méthode évalue la qualité d'un sous ensemble d'attributs en utilisant leur niveau de constance par rapport à la constance de tous l'ensemble d'apprentissage [127]. Si celui-ci est plus faible, cela signifie que, pour une plus faible quantité de variables, les mêmes résultats sont obtenus. La constance de tout sous ensemble ne peut jamais être inférieur à celui de l'ensemble des attributs.
- *FilteredSubsetEval* : c'est une classe pour l'exécution d'un sous ensemble d'attributs arbitraire sur les données qui ont été passées à travers un filtre arbitraire, sachant que les filtres qui altèrent l'ordre ou le nombre d'attributs ne sont pas autorisés. La structure du filtre est basée exclusivement sur les données d'apprentissage [81].
- *ChiSquaredAttributeEval* : c'est une méthode qui détermine le rang d'un attribut par le calcul de la statistique du *Chi-carré* à l'égard de la classe [126]. Le test du

X^2, prononcé *Chi-deux* ou *Chi-carré*, est une loi à densité de probabilité. Cette loi est un test statistique permettant de tester l'indépendance entre deux variables aléatoires.

- *FilteredAttributeEval* : c'est une classe pour exécuter un attribut arbitraire sur les données qui ont été passées à travers un filtre arbitraire, sachant que les filtres qui modifient l'ordre ou le nombre d'attributs ne sont pas autorisés. La structure du filtre est basée exclusivement sur les données d'apprentissage [203].

- *GainRatioAttributeEval* : estime la valeur d'un attribut en mesurant le rapport de gain par rapport à la classe [79].

$$GainR(Class, Attribute) = \frac{(H(Class) - H(Class|Attribute))}{H(Attribute)} \tag{4.10}$$

Avec H représente l'entropie. L'entropie est une mesure de l'incertitude associée à une variable aléatoire. L'entropie pour une variable V est donnée par l'équation suivante :

$$H(V) = -\sum_i P(v_i) \cdot log_2(P(v_i)) \tag{4.11}$$

- *InfoGainAttributeEval* : estime la valeur d'un attribut en mesurant le gain d'information à l'égard de la classe [79].

$$GainR(Class, Attribute) = (H(Class) - H(Class|Attribute)) \tag{4.12}$$

- *OneRAttributeEval:* OneR, abréviation de "One Rule", est un classificateur simple qui génère un seul niveau d'arbre de décision. OneR estime la valeur d'un attribut en utilisant le classificateur de OneR. Le classificateur OneR utilise la validation croisée pour estimer la précision du système d'apprentissage pour un ensemble d'attributs [89]. Il combine l'arbre de décision C4.5 et la distribution gaussienne.

- *ReliefFAttributeEval* : estime la valeur d'un attribut par un échantillonnage répété d'une instance et en considérant la valeur de l'attribut donnée pour l'instance la plus proche de la même et différente classe. Il peut fonctionner à la fois sur des données de classe discret et continu [173].

- *SVMAttributeEval* : estime la valeur d'un attribut en utilisant un classifieur SVM. Les attributs sont classés par le carré du poids attribué par le classifieur SVM [78].

- *SymmetricalUncertAttributeEval* : estime la valeur d'un attribut en mesurant l'incertitude symétrique par rapport à la classe [80].

$$SymmU(Class, Attribute) = \frac{2 * (H(Class) - H(Class|Attribute))}{H(Class) + H(Attribute)} \tag{4.13}$$

Dans ce qui suit, nous présentons les résultats obtenus par chaque algorithme de sélection pour les deux collections d'INEX SBS et IMDb. Nous notons que nous avons utilisé pur chaque algorithme le paramétrage par défaut fournie par Weka.

			Popularité				Réputation		
Algorithmes	Métrique	Lucene Solr	Commentaire	Tweet	Partage (LIn)	Partage	J'aime	+1	Bookmark
CfsSubsetEval	[nombre de sélection]	5	5	5	0	5	5	2	0
WrapperSubsetEval	[nombre de sélection]	5	1	1	1	4	5	3	2
ConsistencySubsetEval	[nombre de sélection]	5	5	5	5	5	5	5	4
FilteredSubsetEval	[nombre de sélection]	5	5	5	0	5	5	2	0
	Moyenne	5	4	4	1.5	4.75	5	3	1.5
ChiSquaredAttributeEval	[rang]	1	4	5	7	2	3	6	8
FilteredAttributeEval	[rang]	1	4	5	7	2	3	6	8
GainRatioAttributeEval	[rang]	1	2	5	8	3	4	6	7
InfoGainAttributeEval	[rang]	1	4	5	7	2	3	6	8
OneRAttributeEval	[rang]	1	3	5	7	4	2	6	8
ReliefFAttributeEval	[rang]	1	4	8	6	2	3	5	7
SVMAttributeEval	[rang]	1	6	7	3	2	5	4	8
SymetricalUncertEval	[rang]	1	2	5	7	3	4	6	8
	Moyenne	1	3.62	5.62	6.5	2.5	3.37	5.62	7.75

Tableau 4.16: Sélection des signaux sociaux avec les algorithmes de sélection d'attributs (Application sur INEX IMDb)

			Popularité			Réputation	
Algorithmes	Métrique	Lucene Solr	Commentaire	Partage	Tag	J'aime	Rating
CfsSubsetEval	[nombre de sélection]	5	5	5	0	5	0
WrapperSubsetEval	[nombre de sélection]	5	1	5	2	4	4
ConsistencySubsetEval	[nombre de sélection]	5	4	5	3	5	4
FilteredSubsetEval	[nombre de sélection]	5	4	5	0	5	0
	Moyenne	5	3.5	5	1.25	4.75	2
ChiSquaredAttributeEval	[rang]	1	5	2	6	3	4
FilteredAttributeEval	[rang]	1	4	2	5	3	6
GainRatioAttributeEval	[rang]	1	4	2	6	3	5
InfoGainAttributeEval	[rang]	1	4	2	6	3	5
OneRAttributeEval	[rang]	1	4	3	5	2	6
ReliefFAttributeEval	[rang]	1	4	2	6	3	5
SVMAttributeEval	[rang]	1	4	2	5	3	6
SymetricalUncertEval	[rang]	1	4	2	6	3	5
	Moyenne	1	4.125	2.125	5.625	2.875	5.25

Tableau 4.17: Sélection des signaux sociaux avec les algorithmes de sélection d'attributs (Application sur INEX SBS)

4.3.4.2 *Résultats et discussions*

Notre objectif dans cette étude est de déterminer les signaux les plus importants pour la tâche de RI, ainsi que de vérifier si les résultats obtenus précédemment (corrélation et probabilité a priori du document) sont cohérents. Dans notre cas, les algorithmes de sélection consistent à donner un score à chaque signal en fonction de son intérêt vis-à-vis la classe de pertinence du document (pertinent ou non pertinent). Nous avons appliqué une validation croisée pour 5 itérations (5 cross-validation folds).

Les tableaux 4.16 et 4.17 présentent les résultats des algorithmes de sélection d'attributs. Certains algorithmes tels que *FilteredAttributeEval* et *SVMAttributeEval* utilisent des méthodes de recherche qui ordonnent les signaux sélectionnés. Le champ "rang" dans les tableaux 4.16 et 4.17 indique l'ordre d'importance des signaux retournés par l'algorithme (la valeur de "rang" est comprise entre 1 et n, avec n le nombre de tous les critères évalués). Par rapport à la collection IMDb (voir tableau 4.16), nous évaluons 8 critères d'où $n = 8$, ainsi que sur SBS (voir tableau 4.17) nous évaluons 6 critères d'où $n = 6$. D'autres algorithmes tels que *FilteredSubsetEval* et *CfsSubsetEval* utilisent des méthodes de recherche qui retournent le nombre de fois qu'un signal a été sélectionné durant les 5 itérations de la validation croisée, soit le champ "nombre de sélection" dans les tableaux (cette valeur est comprise entre 0 et 5). L'ensemble des signaux choisi sera alors évalué sur la base de deux métriques : le "rang" et le "nombre de sélection". Nous notons que le critère préféré par l'ensemble des algorithmes est le *Lucene Solr*, avec un (rang $= 1$) et (nombre de sélection $= 5$), qui représente le modèle de base VSM. Un signal social fortement préféré par l'algorithme de sélection est un signal bien classé (rang $= 1$) et fortement sélectionné (nombre de sélection $= 5$).

Selon les deux tableaux 4.16 et 4.17, nous remarquons que les signaux *partage*(LIn), *bookmark*, *tag* et *rating* sont faiblement préférés par les algorithmes de sélection, avec des moyennes des rangs de 6.5, 7.75, 5.62, 5.25, respectivement, et des moyennes de sélection de 1.5, 1.5, 1.25, 2, respectivement. Le signal +1 est modérément préféré (avec des moyennes de rangs $= 5.6$ et de nombre de sélection $= 3$) mais il est sélectionné par chaque algorithme, ce qui indique son importance même s'il n'est pas le meilleur. Ainsi, les signaux de Facebook : *j'aime*, *partage*, *commentaire* et le *tweet* sont les mieux classés et souvent validés au cours des 5 itérations de la validation croisée.

En comparant ces résultats avec les résultats figurant dans les tableaux 4.13 et 4.14, nous avons remarqué que les mêmes facteurs de pertinence mis en avant par l'étude de corrélation et la probabilité a priori sont mis en avant par l'étude avec les techniques de sélection d'attributs. Les signaux sociaux (*partage* (LIn), *bookmark*, *tag* et *rating*) qui fournissent relativement les résultats les plus bas sont les moins préférés par les algorithmes de sélection. Cependant, les signaux (*j'aime* et *partage*) qui fournissent les meilleurs résultats en termes de précision et nDCG sont sélectionnés à chaque itération et sont bien classés par les différents algorithmes de sélection.

4.3.4.3 *Approches basées sur l'apprentissage*

Nous avons également conduit une série d'expérimentation en exploitant ces signaux dans des approches supervisées basées sur des techniques d'apprentissage. Nous avons utilisé les résultats retournés par Lucene Solr en utilisant toutes les requêtes des deux collections d'INEX, chacune à part, comme collection d'apprentissage. Nous avons ensuite utilisé trois algorithmes d'apprentissage, ce choix s'explique par le fait qu'ils ont

souvent montré leur efficacité dans la RI en exploitant des critères de pertinence : SVM [197], J48 [214] (une implantation de C4.5 [164]) et Naive Bayes [214]. L'entrée de chaque algorithme est un vecteur de signaux sociaux, soit tous les signaux ou juste les signaux sélectionnés par un algorithme de sélection précis. Chaque signal est représenté par sa quantité dans les documents. Les algorithmes d'apprentissage prédisent la classe de pertinence pour chaque document (pertinent ou non pertinent). Nous avons appliqué une validation croisée pour 5 itérations (5 cross-validation folds). La figure 4.5 illustre le processus d'apprentissage que nous avons mis en place pour l'évaluation des signaux sociaux.

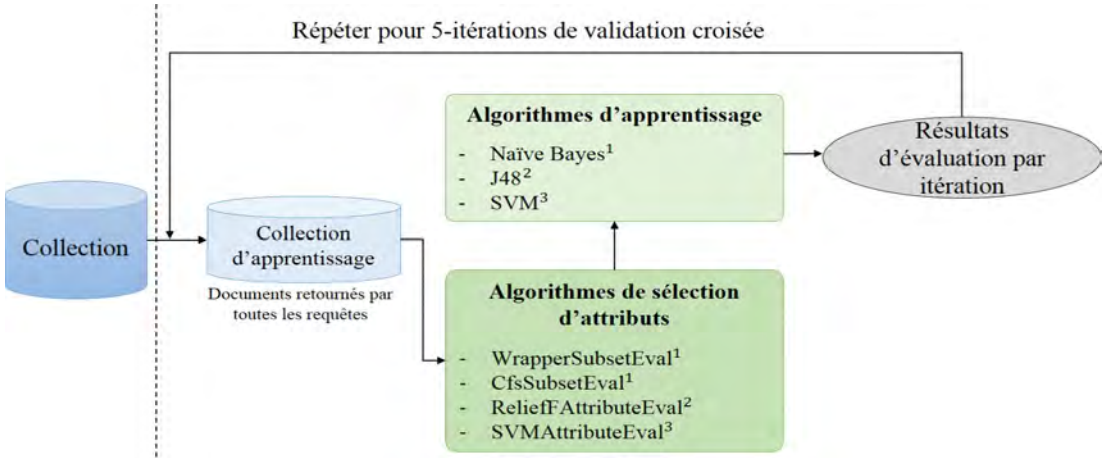

Figure 4.5: Processus d'apprentissage automatique

Nous rappelons que la phase des algorithmes de sélection d'attributs a fait ressortir deux ensembles de signaux :

1. dans le cas des algorithmes *CfsSubsetEval* et *FilteredSubsetEval*, les signaux sélectionnés sont : *commentaire, partage* et *j'aime* pour la collection de SBS et *commentaire, tweet, partage, j'aime* et *+1* pour la collection d'IMDb.

2. dans le cas des autres algorithmes de sélection, tous les signaux sociaux étudiés sur les deux collections sont sélectionnés : *commentaire, partage, tag, j'aime* et *rating* pour la collection de SBS et *commentaire, tweet, partage(LIn), partage, j'aime, +1* et *bookmark* pour la collection d'IMDb.

La question à ce stade est liée à la spécification du vecteur de signaux d'entrée pour les algorithmes d'apprentissage, soit on prend tous les signaux, soit on garde uniquement ceux sélectionnés par les techniques de sélection d'attributs. Dans ce cas, avec quels algorithmes d'apprentissage ces derniers seront combinés.

Afin de prendre en compte les signaux choisis par les algorithmes de sélection dans des modèles d'apprentissage, nous nous sommes basés sur les travaux de *Hall* et *Holmes* [81].

Hall et *Holmes* [81] ont étudié l'efficacité de certaines techniques de sélection d'attributs en les confrontant avec les techniques d'apprentissage. Étant donné que la performance des facteurs diffère d'une technique d'apprentissage à une autre, ils ont identifié les meilleures techniques de sélection d'attributs permettant de retrouver les facteurs les plus performants en fonction des techniques d'apprentissage à utiliser.

En se basant sur leur étude, nous avons utilisé les mêmes couples des techniques d'apprentissage et des techniques de sélection d'attributs :

- l'ensemble des signaux sélectionnés par l'algorithme *WrapperSubsetEval* (WRP) sont appris par le modèle Naïve Bayes.
- l'ensemble des signaux sélectionnés par l'algorithme *CfsSubsetEval* (CFS) sont appris par le modèle Naïve Bayes.
- l'ensemble des signaux sélectionnés par l'algorithme *ReliefFAttributeEval* (RLF) sont appris par le modèle J48.
- l'ensemble des signaux sélectionnés par l'algorithme *SVMAttributeEval* (SVM) sont appris par le modèle SVM.

Classifieurs	Techniques de sélection	Tous les critères
NaïveBayes	**0.4927** (CFS) 0.4802 (WRP)	0.4802
SVM	0.4874 (SVM)	0.4874
J48	**0.5562** (RLF)	**0.5562**

Tableau 4.18: Résultats de l'apprentissage automatique (P@20) sur INEX IMDb

Classifieurs	Techniques de sélection	Tous les critères
NaïveBayes	0.1223 (CFS) 0.1100 (WRP)	0.1100
SVM	0.1114 (SVM)	0.1114
J48	**0.1301** (RLF)	**0.1301**

Tableau 4.19: Résultats de l'apprentissage automatique (P@20) sur INEX SBS

Les tableaux 4.18 et 4.19 présentent les résultats des trois algorithmes d'apprentissage des signaux ressortis de l'étude avec les techniques de sélection d'attributs. Nous constatons que seul l'algorithme de CFS confirme l'hypothèse lancé par *Hall* et *Holmes*. C'est en effet le seul pour lequel les résultats obtenus avec la sélection d'attributs, soient 0.4927 (IMDb) et 0.1223 (SBS), dépassent l'utilisation de tous les signaux, 0.4802 (IMDb) et 0.1100 (SBS). Nous avons montré que les approches d'apprentissage automatique ont une meilleure efficacité (précision) avec les approches de sélection d'attributs.

Nous remarquons ensuite que tous les modèles d'apprentissage surpassent les modèles textuels (modèle Lucene Solr, BM25 et Hiemstra) ainsi que nos premières propositions basées sur la probabilité a priori. Nous constatons enfin que l'arbre de décision J48 est le modèle le plus approprié, il prend en considération tous les signaux sociaux, les taux d'amélioration par rapport à Naïve Bayes et SVM sont de 13% et 15% sur IMDb ainsi que de 6% et 17% sur SBS, respectivement.

En outre, le J48 donne les meilleures améliorations par rapport à toutes les approches précédentes, le taux d'amélioration par rapport au modèle par défaut de Lucene Solr

(configuration nommée *Lucene Solr* dans les tableaux 4.13 et 4.14) est de 78% pour la collection IMDb et de 167% pour la collection SBS, alors que par rapport aux meilleurs résultats obtenus par le modèle basé sur les probabilités a priori (configuration nommée *ToutesLesPropriétés* dans les tableaux 4.13 et 4.14) est de 23% pour la collection IMDb et de 46% pour la collection SBS.

4.3.5 *Bilan*

Nous avons proposé dans ce chapitre des modèles de recherche basés sur les signaux sociaux et les propriétés sociales. Ces signaux sont considérés comme une connaissance a priori du document permettant de mesurer son intérêt et sa pertinence indépendamment de la requête. Les propriétés sociales ont été définies par ces signaux issus de plusieurs réseaux sociaux. Afin de montrer la corrélation éventuelle entre la présence des signaux sociaux sur des documents et la pertinence de ces documents, nous avons mené une étude mesurant la distribution des signaux dans les documents pertinents.

En analysant les corrélations des signaux avec la pertinence des documents, nous notons que tous les signaux sociaux présentent une corrélation positive avec la pertinence. Nous avons montré que les signaux sociaux pris individuellement ou de manière groupée améliorent les performances des systèmes de RI. Certains signaux sociaux tels que le *j'aime* et le *partage* permettent d'améliorer la qualité des résultats de manière très importante, tandis que d'autres tels que le *bookmark* et le *partage*(LIn) fournissent de modestes améliorations. En outre, la probabilité a priori basée sur les propriétés, *popularité* et *réputation*, améliorent considérablement les résultats par rapport aux modèles de base et la prise en compte des signaux individuellement.

Notre travail présente cependant quelques limites. D'abord, nous avons considéré que les signaux sont tous de même importance. Ils ne se différencient que par leur nombre vis-à-vis de la ressource à laquelle ils sont associés. Selon les études que nous avons menées sur la corrélation les techniques de sélection d'attributs, il semblerait que certains soient plus important que d'autres pour la recherche d'information. Ensuite, nous n'avons pas pu évaluer l'impact d'autres signaux, tels que le *j'aime* de LinkedIn et les auteurs des différents signaux, sur le processus de RI. La récupération de ces informations n'est pas accessible via les APIs des réseaux sociaux actuels.

Une autre limite de cette première étude réside dans la non prise en compte des facteurs temporels (temporalité des signaux, date de publication de la ressource, date du signal). Nous pensons qu'un comptage simple de la quantité des signaux associés à une ressource privilégieront les ressources anciennes. Nous traitons plus finement ces aspects dans les chapitres qui suivent.

Partie IV

EXPLOITATION DE LA TEMPORALITÉ ET LA DIVERSITÉ DES SIGNAUX SOCIAUX

B2C et B2B n'existent plus. Ce qui importe désormais, c'est le P2P. Le People-to-People.

— Jowyang

5

TEMPORALITÉ DES SIGNAUX SOCIAUX

Introduction

Comme nous l'avons vu dans le chapitre 3, les signaux sociaux sont de plus en plus exploités par les moteurs de recherche [189]. En effet, plus ces signaux sont fréquents sur une ressource plus son importance a priori croît [10]. Cependant, dans les travaux existants les signaux sont pris en compte indépendamment du moment où l'action (le signal) s'est produite et de la date de publication de la ressource (document). Ils sont souvent pris en compte uniquement par rapport à leurs fréquences dans la ressource. C'est également la manière dont nous les avons considéré dans notre approche.

Dans ce chapitre, nous nous intéressons particulièrement à la temporalité associée à ces signaux. Nous supposons que l'importance a priori d'un document dépend non seulement de la qualité (on suppose que l'apprentissage fait de la qualité) et la quantité de ces signaux mais aussi de la date de leur création ainsi que la date de publication de la ressource. De ce fait, plutôt que d'estimer cette importance (probabilité) a priori par un simple comptage des signaux liés au document, nous intégrons également la date de publication de la ressource, pour ne pas pénaliser les nouvelles ressources, et les dates des signaux pour privilégier les signaux récents. Nous évaluons la performance de notre approche sur deux types de collection, IMDb (Internet Movies Database) et SBS (Social Book Search), enrichies de plusieurs données sociales collectées à partir de plusieurs réseaux sociaux.

5.1 Hypothèses et questions de recherche

Nous pensons que l'impact d'un signal social dépend également du temps, c'est-à-dire la date à laquelle l'action de l'utilisateur est réalisée. Nous considérons que les signaux récents devraient avoir un impact supérieur vis-à-vis des signaux anciens dans le calcul de l'importance d'une ressource. La récence des signaux peut indiquer certains intérêts récents à la ressource. Dans notre seconde hypothèse, nous considérons que le nombre de signaux d'une ressource doit être pris en compte au regard de l'âge (date de publication) de cette ressource. En général, une ressource ancienne en termes de durée d'existence a de fortes chances d'avoir beaucoup plus de signaux qu'une ressource récente. Ceci conduit donc à pénaliser les ressources récentes vis-à-vis de celles qui sont anciennes.

Nous proposons dans ce chapitre un modèle d'estimation de l'importance a priori d'une ressource qui tient compte les caractéristiques temporelles des actions des utilisateurs comme connaissance a priori dans un modèle de recherche. Les questions de recherche abordées dans ce chapitre sont les suivantes :

1. Comment prendre en compte la date de création des signaux et la date de publication de la ressource pour estimer la probabilité a priori de cette ressource ?
2. Quel est l'impact du temps associé aux signaux sociaux sur la performance de la recherche d'information ?

5.2 Approche basée sur la temporalité des signaux

Nous proposons d'estimer l'importance sociale d'une ressource en exploitant le moment où l'interaction (signal) s'est produite ainsi que la date de publication de la ressource. Afin de prendre en compte cette importance dans l'évaluation de pertinence, nous nous appuyons sur les modèles de langue [159] pour combiner la pertinence textuelle d'une ressource vis-à-vis d'une requête et son importance socio-temporelle modélisée comme une probabilité a priori. Nous reprenons le même modèle général basé sur les signaux sociaux, qui a été présenté dans le chapitre précédent section 4.2.2, mais en prenant en compte l'aspect temporel. Finalement, notre principale contribution de ce chapitre est sur l'estimation de $P(D)$ en exploitant les signaux sociaux et leurs caractéristiques temporelles.

5.2.1 *Préliminaires et notations*

L'information sociale que nous exploitons dans le cadre de notre modèle est représentée par le quintuplet <U, R, A, T, RS> où U, R, A, T, RS sont des ensembles finis d'instances : *Utilisateurs*, *Ressources*, *Actions*, *Temps* et *Réseaux sociaux*.

Les notations des instances *Ressources*, *Actions* et *Réseaux sociaux* ont été présentées dans le chapitre précédent section 4.2.1. De plus, dans ce chapitre, nous introduisons l'aspect temporel qui est défini comme suit :

5.2.1.1 *Temps*

Le temps T intervient à deux niveaux dans notre approche :

- l'historique de chaque action, soit $T_{a_i} = \{t_{1,a_i}, t_{2,a_i}, ...t_{k,a_i}\}$ l'ensemble de k moments (date) à laquelle une action a_i s'est produite. Un instant de temps t_{k,a_i} représente la date et l'heure (datetime, on pourrait considérer différentes échelles : jour, année,...) de l'action effectuée par un utilisateur u sur une ressource D.
- la date de publication de la ressource, soit $T_D = \{t_{D_1}, t_{D_2}, ...t_{D_n}\}$ l'ensemble de n date à laquelle chaque ressource D de la collection R a été créée. t_D est la date de publication de la ressource D, cette date est fournie en format datetime.

5.2.2 *Prise en compte de la date du signal social*

Nous supposons que les ressources associées aux signaux frais (récents) devraient être favorisées par rapport à celles qui sont associées à des signaux anciens. Chaque fois qu'un signal apparaît, il est associé à sa date d'occurrence. Nous proposons de compter les occurrences d'un signal en le pondérant (en le *boostant*) avec sa date d'apparition, soit $Count_{t_a}$. La formule correspondante est la suivante :

$$Count_{t_a}(t_{j,a_i}, D) = \sum_{j=1}^{k} f(t_{j,a_i}, D) \tag{5.1}$$

La pondération de l'occurrence peut se faire de différentes manières. Une façon simple est de prendre par exemple une fonction linéaire inversement proportionnelle à la date d'apparition :

$$f(t_{j,a_i}, D) = \frac{1}{t_{actuel} - t_{j,a_i}} \tag{5.2}$$

ou une pondération exponentielle qui boosterait d'avantage les signaux "récents" vis-à-vis les signaux "anciens" :

$$f(t_{j,a_i}, D) = \exp\left(-\frac{\parallel t_{actuel} - t_{j,a_i} \parallel^2}{2\sigma^2}\right) \tag{5.3}$$

Afin d'éviter la division par zéro qui peut être provoquée par la formule 5.2, nous utilisons la formule 5.3 où $f(t_{j,a_i}, D)$ représente la fonction temporelle du signal, estimée en utilisant le noyau Gaussien [194]. Cette fonction calcule la distance temporelle entre la date actuelle t_{actuel} et la date de l'action t_{j,a_i}. $\sigma \in R_+$ est le paramètre du noyau Gaussien.

Nous notons que plus la distance euclidienne relative au temps $\parallel t_{actuel} - t_{j,a_i} \parallel^2$ augmente, plus la valeur du noyau Gaussien diminue. Par conséquent, les signaux sociaux les plus récents sont les plus favorisés. Le signal qui se produit à l'heure actuelle ($t_{actuel} - t_{j,a_i} = 0$) aura une fréquence égale à 1. Plus la date du signal s'éloigne de la date actuelle plus la fréquence diminue.

La probabilité a priori $P(D)$ est estimée en utilisant la formule 4.4 du chapitre 4 mais en remplaçant le $Count()$ par $Count_{t_a}()$. Nous notons que si la date du signal est ignorée dans le calcul $f(t_{j,a_i}, D) = 1\ \forall t_{j,a_i}$ (chaque occurrence d'un signal est captée une fois).

Concernant la prise en compte du temps dans le *rating*, nous intégrons le temps pour chaque note *rating*, ensuite nous calculons la moyenne des notes biaisées par leurs temporalités. La formule générale est la suivante :

$$BA_t(D) = \frac{moy(r_t) \cdot |r_t| + \sum_{D' \in R} moy(r'_t) \cdot |r'_t|}{|r_t| + \sum_{D' \in R} |r'_t|}, \tag{5.4}$$

Avec :

- $r_t = \{\{r_{ti}\}\}$ et $i = 1, nr$. nr est le nombre de *ratings* associés au document D, $nr = |r_t|$. r_{ti} est le ième *rating* biaisé par sa date de création t_{r_i} donné par l'utilisateur i au document D, en utilisant la formule gaussienne 5.3 comme suit :

$$r_{ti} = r_i \cdot \exp\left(-\frac{\parallel t_{actuel} - t_{r_i} \parallel^2}{2\sigma^2}\right) \tag{5.5}$$

- *moy* est la fonction de moyenne des *ratings* biaisés par leur dates de création du document D.

- r'_t est l'ensemble des *ratings* dans toute la collection R.

Nous notons que nous estimons le $P(a_i = rating)$ en utilisant la formule 4.8, en remplaçant $BA(D)$ par $BA_t(D)$.

5.2.3 *Prise en compte de la date de publication de document*

La date de publication d'une ressource joue un rôle important sur la vie sociale de cette ressource dans les réseaux sociaux, c'est-à-dire qu'une vieille ressource a une plus grande chance d'avoir un grand nombre d'interactions par rapport à une ressource publiée récemment. Donc, pour résoudre ce problème, nous proposons de normaliser la distribution des signaux sociaux associés à une ressource par la date de publication de la ressource. La formule correspondante est la suivante :

$$Count_{t_D}(a_i, D) = Count(a_i, D) \cdot A(D) \tag{5.6}$$

Où :

$$A(D) = \exp\left(-\frac{\| t_{actuel} - t_D \|^2}{2\sigma^2}\right) \tag{5.7}$$

Avec :

- $A(D)$ représente la fonction temporelle du document, estimée en utilisant le noyau Gaussien pour les mêmes raisons citées précédemment. Cette fonction calcule la distance temporelle entre la date actuelle t_{actuel} et la date de la ressource t_D.
- Paramètre du noyau Gaussien $\sigma \in R_+$.

La probabilité a priori $P(D)$ est estimée en utilisant la formule 4.4 du chapitre 4 mais en remplaçant le $Count()$ par $Count_{t_D}()$ pour le document et $Count_{t_C}()$ pour la collection.

Par rapport au signal *rating*, la formule correspondante est la suivante :

$$BA_{T_D}(D) = BA(D) \cdot A(D) \tag{5.8}$$

Avec $BA_{T_D}(D)$ représente la moyenne bayésienne des *ratings* biaisée par l'âge du document. Nous notons que nous estimons le $P(a_i = rating)$ en utilisant la formule 4.8, en remplaçant $BA(D)$ par $BA_{T_D}(D)$.

5.3 Expérimentations et résultats

Afin de valider notre approche, nous avons effectué une série d'expérimentations sur les collections IMDb (Internet Movie Database) et SBS (Social Book Search). Notre objectif principal dans ces expériences est d'évaluer l'impact du temps associé aux signaux sur le système de recherche d'information vis-à-vis à la fois d'approches qui ne prennent pas en compte ces facteurs et aussi celles qui ne considèrent pas du tout cette notion de probabilité a priori.

5.3.1 *Cadre expérimental*

5.3.1.1 *Données expérimentales*

Nous avons utilisé les mêmes collections d'INEX présentées dans le chapitre précédent en utilisant les mêmes requêtes et les mêmes jugements de pertinence. Nous avons

également suivi le même processus d'indexation et de collecte de données sociales (voir chapitre 4 section 4.3.1). Cependant, pour chaque document IMDb et SBS, nous avons collecté d'autres types de données sociales liées à la temporalité des signaux, via l'API de Facebook (voir les tableaux 5.1 et 5.2). Dans notre étude, nous nous sommes concentrés sur l'évaluation des 1000 premiers documents retournés par les différentes approches.

Les tableaux 5.1 et 5.2 montrent un exemple de données sociales temporelles pour les documents d'IMDb et SBS. Par exemple, dans le tableau 5.1, le document IMDb ayant le Id=*tt1730728* a été partagé la dernière fois sur Facebook le *2013-09-11T20:55:47*, commenté la dernière fois le *2012-03-01T11:07:32* et publié pour la première fois le *2010-09-29T05:08:09*. Tandis que dans le tableau 5.2 contient des documents SBS, nous avons une information supplémentaire par rapport aux documents IMDb, c'est la date de chaque *rating* associé au document. Par exemple, le document ayant le Id=*0553583859* ses deux premiers *rating* ont eu lieu le : *1999-12-15* et *2000-01-04*.

	Facebook		
Id	Dernier Partage	Dernier Commentaire	Date de publication du Document
tt1730728	2013-09-11T20:55:47	2012-03-01T11:07:32	2010-09-29T05:08:09
tt1922777	2014-09-29T02:49:01	2014-09-28T00:41:01	2011-05-07T19:00:57

Tableau 5.1: Exemple de deux documents IMDb ayant des données sociales

	Facebook			Amazon
Id	Dernier Partage	Dernier Commentaire	Date de publication	Date Rating
0553583859	2014-03-10T32:01:32	2014-03-18T00:01:43	2008-12-14T02:13:22	1999-12-15
	-	-	-	2000-01-04
0441014100	2015-01-01T21:17:41	2015-03-21T00:31:11	2010-02-11T08:01:21	2008-02-09

Tableau 5.2: Exemple de documents SBS ayant des données sociales

Malheureusement, les dates des différentes actions que nous souhaitons exploiter ne sont pas disponibles, sauf la date de chaque *rating* issue d'Amazon/LibraryThing et les dates des dernières actions issues de Facebook (*commentaire* et *partage*). Par conséquent, nous représentons les résultats en utilisant la formule 5.1 biaisée (prise en compte du temps) uniquement par la date du dernier *commentaire* et *partage* ainsi que les dates des *ratings*.

5.3.1.2 *Métriques d'évaluation*

Pour évaluer la performance de notre approche, nous calculons les métriques de précision (P@k, MAP) et le nDCG (Normalized Discounted Cumulative Gain) qui sont aussi suggérées par la compagne d'évaluation INEX SBS 2015. Pour plus d'informations voir chapitre 2 section 2.4.2.

5.3.1.3 *Modèles de référence*

Nous avons utilisé le moteur de Lucene Solr pour l'indexation et la recherche. En plus des modèles de base textuels présentés dans le chapitre précédent (le modèle par défaut

de Lucene Solr, le modèle de langue Hiemstra et BM25), nous avons considéré des configurations de notre approche, qui ne prennent pas en compte la date de l'action et la date de publication de la ressource, comme modèles de référence.

5.3.2 *Résultats et discussions*

Nous avons mené des expériences avec des modèles basés uniquement sur le contenu textuel des documents (le modèle de Lucene Solr, BM25 et le modèle de langue Hiemstra sans la probabilité a priori), ainsi que des approches combinant le contenu textuel et les caractéristiques sociales avec prise en compte de leur aspect temporel. Nous notons que la meilleure valeur de μ (paramètre utilisé dans le lissage de Dirichlet) appartient à l'intervalle suivant : $\mu \in [90, 100]$ pour IMDb et à l'intervalle $\mu \in [2400, 2500]$ pour SBS. Les résultats obtenus sont présentés dans les tableaux 5.3 et 5.4.

Modèles	P@10	P@20	nDCG	MAP
Base (A) : Sans Probabilité a Priori				
BM25	0.0601	0.0517	0.1581	0.0517
Lucene Solr	0.0528	0.0487	0.1300	0.0463
ML.Hiemstra	0.0607	0.0559	0.1620	0.0527
Base (B) : Sans Prise en Compte du Temps				
J'aime	0.0857	0.0689	0.1864	0.0741
Partage	0.0901	0.0711	0.1900	0.0872
Commentaire	0.0799	0.0678	0.1807	0.0701
TotalFacebook	0.0958	0.0810	0.1937	0.0892
BA_Rating	0.0730	0.0559	0.1748	0.0620
Log_Tag	0.0770	0.0531	0.1742	0.0610
TousLesCritères	0.0973	0.0787	0.1981	0.0900
(a) Avec Prise en Compte de la Date de l'Action T_a				
Partage^{T_a}	0.0910*	0.0787*	0.1918*	0.0884*
Commentaire^{T_a}	0.0846*	0.0708*	0.1866*	0.0753*
BA_Rating^{T_a}	0.0941**	0.0732**	0.1904**	0.0885**
(b) Avec Prise en Compte de la Date de Publication de la Ressource T_D				
J'aime^{T_D}	0.0891*	0.0708*	0.1900*	0.0873*
Partage^{T_D}	0.0917*	0.0796*	0.1947*	0.0903*
Commentaire^{T_D}	0.0881*	0.0711*	0.1882*	0.0777*
TotalFacebook^{T_D}	0.0957*	0.0873*	0.1959*	0.0928*
BA_Rating^{T_D}	0.0790*	0.0695*	0.1808*	0.0685*
Log_Tag^{T_D}	0.0782*	0.0599*	0.1771*	0.0666*
TousLesCritères^{T_D}	0.1078**	0.0973**	0.2080**	0.0986**

Tableau 5.3: Résultats de P@k, nDCG et MAP sur la collection INEX SBS

Les tableaux 5.3 et 5.4 récapitulent les résultats de précision@k pour $k \in \{10, 20\}$ et de nDCG, ainsi que la MAP. Nous avons évalué notre approche à travers des configu-

Modèles	P@10	P@20	nDCG	MAP
Base (A) : Sans Probabilité a Priori				
BM25	0.3500	0.3371	0.4113	0.2068
Lucene Solr	0.3411	0.3122	0.3919	0.1782
ML.Hiemstra	0.3700	0.3403	0.4325	0.2402
Base (B) : Sans Prise en Compte du Temps				
J'aime	0.3938	0.3620	0.5130	0.2832
Partage	0.4061	0.3649	0.5262	0.2905
Commentaire	0.3857	0.3551	0.5121	0.2813
TotalFacebook	0.4209	0.4102	0.5681	0.3125
Tweet	0.3879	0.3512	0.4769	0.2735
+1	0.3826	0.3468	0.5017	0.2704
Bookmark	0.3730	0.3414	0.4621	0.2600
Partage (LIn)	0.3739	0.3432	0.4566	0.2515
TousLesCritères	0.4408	0.4262	0.5974	0.3300
a) Avec Prise en Compte de la Date de l'Action T_a				
Partage^{T_a}	0.4148*	0.3681*	0.5472*	0.2970*
Commentaire^{T_a}	0.3861*	0.3601*	0.5207*	0.2844*
b) Avec Prise en Compte de la Date de Publication de la Ressource T_D				
J'aime^{T_D}	0.4091*	0.3620*	0.5308*	0.2907*
Partage^{T_D}	0.4177*	0.3721*	0.5544*	0.2989*
Commentaire^{T_D}	0.3912*	0.3683*	0.5285*	0.2874*
TotalFacebook^{T_D}	0.4302	0.4258	0.5827	0.3200
Tweet^{T_D}	0.3918*	0.3579*	0.4903*	0.2779*
+1^{T_D}	0.3900	0.3511	0.5246	0.2748
Bookmark^{T_D}	0.3732	0.3427	0.4671	0.2618
Partage^{T_D} (LIn)	0.3762	0.3449	0.4606	0.2542
TousLesCritères^{T_D}	0.4484*	0.4305*	0.6200*	0.3366*

Tableau 5.4: Résultats de P@k, nDCG et MAP sur la collection INEX IMDb

rations différentes, en prenant en compte les signaux sociaux séparément et avec prise en compte de : a) leur date de création et b) la date de publication de la ressource. Afin de vérifier si les résultats obtenus sont statistiquement significatifs par rapport aux modèles de base (Base(B)), nous avons effectué le test de Student [75]. Nous attribuons * (forte signification par rapport à (Base(B)) et ** (très forte signification par rapport à (Base(B)) lorsque *p-value*$<$ 0.05 et *p-value*$<$ 0.01, respectivement. Nous discutons dans ce qui suit les résultats de chacune des configurations que nous avons étudiées.

5.3.2.1 *Prise en compte de la date de signal*

La partie (a) dans les tableaux 5.3 et 5.4 présente les résultats obtenus en intégrant la date de l'action. Dans notre cas, les dates du dernier *commentaire* et *partage* sur Facebook ainsi que les dates de chaque signal *rating*. Les résultats, obtenus dans les deux collections (IMDb et SBS), montrent que le nDCG, la précision@k et la MAP sont en général meilleurs par rapport à celles obtenues lorsque le temps de l'action est ignoré (Base (B)). Concernant la collection IMDb, les taux d'amélioration en termes de nDCG sont de 4% pour le *partage* et de 2% pour le *commentaire*.

Tandis que par rapport à la collection SBS, on enregistre des taux d'améliorations de *partage* 1%, *commentaire* 7.42% et *rating* 42.75%. Ces résultats aussi bien pour IMDb et pour SBS sont statistiquement significatifs. Nous remarquons dans le tableau 5.3 que le $rating^{T_a}$ apporte les meilleurs résultats par rapport aux deux autres signaux $partage^{T_a}$ et $commentaire^{T_a}$, ceci revient à l'exploitation de la date de chaque *rating* et pas uniquement la date de la dernière action. En conséquence, ces résultats confirment notre hypothèse, que les ressources associées aux signaux frais devraient être favorisées par rapport à ceux associées aux anciens signaux. Cependant, nous n'avons pas vraiment évalué l'impact réel de notre proposition au regard de tous les signaux sociaux, et de toutes les dates auxquelles les actions considérées se sont produites (hormis pour le *rating*). L'exploitation de la date de la dernière action n'est pas suffisante pour tirer des conclusions efficaces sur l'ensemble des signaux. Mais les résultats obtenus par le $rating^{T_a}$, pour lequel toutes les conditions étaient satisfaites, nous encouragent à investir davantage cette piste.

5.3.2.2 *Prise en compte de la date de publication de document*

Nous avons étudié également la performance de la recherche d'information par l'intégration de la date de publication de la ressource. Les résultats présentés dans la partie (b) des tableaux 5.3 et 5.4 montrent que le nDCG et les précisions sont meilleurs par rapport à ceux obtenus lorsque la date de publication de la ressource est ignorée (modèles de base (A) et (B)). Le signal $partage^{T_D}$ semble le plus approprié parmi les autres signaux selon ses résultats, tandis que la combinaison de tous les signaux Facebook $TotalFacebook^{T_D}$ améliore davantage la précision mais pas le nDCG. Les meilleurs résultats sont obtenus par la combinaison de tous les critères (nommé $TousLesCritères^{T_D}$ dans les tableaux 5.3 et 5.4) en prenant en compte la date de publication du document, avec un taux d'amélioration en termes de nDCG de 4% sur IMDb et 5% sur SBS, par rapport à *TousLesCritères* où le temps est ignoré. En effet, une ressource qui génère une activité sociale de 100 actions de *j'aime* pendant une heure de temps, ne suscite pas la même importance et le même intérêt temporel pour les utilisateurs par rapport à une ressource qui a réuni 100 actions de *j'aime* durant une semaine.

5.3.2.3 *Évaluation de l'impact de la temporalité des signaux*

Nous avons évalué l'impact réel de la temporalité des différents signaux sociaux en utilisant les algorithmes de sélection d'attributs appliqués aux deux collections (IMDb et SBS). L'objectif est de déterminer les meilleurs signaux dépendants du temps que nous pouvons exploiter effectivement en RI, ainsi que de vérifier si les résultats obtenus précédemment (probabilité a priori du document) sont cohérents. Nous avons utilisé

Weka[1] pour ces expérimentations. Nous notons que nous avons procédé de la même manière décrite dans le chapitre 4 section 4.3.4.1, mais en prenant en compte des signaux biaisés par leur date de création ainsi que l'âge du document.

Les tableaux 5.5 et 5.6 présentent le nombre de fois qu'un signal social a été sélectionné par un algorithme (cette valeur est comprise entre 0 et 5) et le rang du signal (sélection par ordre de préférence) par rapport aux autres signaux considérés. Par rapport à SBS la valeur de "rang" est comprise entre 1 et 9, avec 9 le nombre total des critères évalués. Concernant IMDb, la valeur de "rang" est comprise entre 1 et 10. Par exemple, dans le tableau 5.5 le signal *tweet* a été sélectionné 5 fois par l'algorithme *CfsSubsetEval*, 3 fois par l'algorithme *WrapperSubsetEval* et classé 5eme parmi les 10 critères sur le tableau par l'algorithme *FiltredAttributeEval*. Nous discutons les résultats dans ce qui suit.

- **Date de publication du document** : selon les deux tableaux 5.5 et 5.6, nous remarquons que les signaux biaisés par la date de publication du document : *partage*(LIn), *bookmark*, *tag* et *rating* sont faiblement favorisés par les algorithmes de sélection, avec des moyennes des rangs de 9, 10, 8.75, 8.25, respectivement, et des moyennes de sélection de 1.75, 1.5, 8.75, 3, respectivement. Tandis que l'action +1 est modérément favorisée (avec des moyennes de rangs $= 7$ et un nombre de sélection $= 3.25$). Cependant, ils sont tous sélectionnés par chaque algorithme à l'exception de *partage*(LIn) et *bookmark* qui n'ont pas été sélectionnés par les deux algorithmes *CfsSubsetEval* et *FilteredSubsetEval*, ce qui indique leur faible impact. Les signaux de Facebook : *j'aime* et *partage* sont les mieux classés par rapport aux autres signaux, ils sont fortement validés au cours des 5 itérations de la validation croisée (avec des moyennes de rangs 2.6 et 2.3, respectivement). Le *commentaire* et le *tweet* viennent en seconde position, ils sont souvent sélectionnés au cours des 5 itérations de la validation croisée.

- **Date de l'action** : concernant les signaux biaisés par leur date d'action, nous remarquons à travers les deux tableaux 5.5 et 5.6 que le *rating* vient en première position par rapport à tous les autres signaux, il est sélectionné dans 5 itérations de la validation croisée par tous les algorithmes ainsi qu'il vient en moyenne dans le 2ème rangs après le modèle de base *Lucene Solr*. Une des raisons de ces résultats revient à l'efficacité de l'exploitation des dates de chaque action *rating*. Par conséquent, notre hypothèse est pleinement vérifiée à travers ces résultats. Il est à noté aussi que le *partage* vient en seconde position suivit par le *commentaire* (avec une moyenne de sélection de 3.5 (SBS), 4.25 (IMDb) et avec une moyenne de rang 6.6 (SBS), 8 (IMDb)). Il est a rappelé que par rapport à ces deux signaux, nous avons seulement exploité la date de la dernière action.

En comparant ces résultats avec les résultats figurant dans les tableaux 5.3 et 5.4, nous remarquons que les mêmes facteurs de pertinence mis en avant par le modèle basé sur les probabilités a priori sont mis en avant par l'étude avec les techniques de sélection d'attributs. Les signaux sociaux pris en compte avec leur temporalité, (*rating* et *partage*), qui fournissent les meilleurs résultats (statistiquement significatifs) sont les plus favorisés et les mieux classés par les différents algorithmes de sélection.

1 http://www.cs.waikato.ac.nz/ml

			Date de publication du document T_D							Date de l'action T_a	
Algorithmes	Métrique	Lucene	Commentaire	Tweet	Partage (LIn)	Partage	J'aime	+1	Bookmark	Partage	Commentaire
CfsSubsetEval	[nombre de sélection]	5	5	5	0	5	5	3	0	5	5
WrapperSubsetEval	[nombre de sélection]	5	3	3	2	5	5	3	2	5	2
ConsistencySubsetEval	[nombre de sélection]	5	5	5	5	5	5	5	4	5	5
FilteredSubsetEval	[nombre de sélection]	5	5	5	0	5	5	3	0	5	5
	Moyenne	5	4.5	4.5	1.75	5	5	3.25	1.5	5	4.25
ChiSquaredAttributeEval	[rang]	1	5	6	9	2	3	7	10	4	8
FilteredAttributeEval	[rang]	1	6	5	9	2	3	7	10	4	8
GainRatioAttributeEval	[rang]	1	6	5	9	3	2	7	10	4	8
InfoGainAttributeEval	[rang]	1	5	6	9	3	2	7	10	4	8
OneRAttributeEval	[rang]	1	5	6	9	3	2	7	10	4	8
ReliefFAttributeEval	[rang]	1	6	5	9	2	3	7	10	4	8
SVMAttributeEval	[rang]	1	6	5	9	2	3	7	10	4	8
SymetricalUncertEval	[rang]	1	6	5	9	2	3	7	10	4	8
	Moyenne	1	5.625	5.375	9	2.375	2.625	7	10	4	8

Tableau 5.5: Sélection des signaux sociaux temporellement dépendants par les algorithmes de sélection d'attributs (Application sur INEX IMDb)

			Date de publication du document T_D					Date de l'action T_a		
Algorithmes	Métrique	Lucene Solr	Commentaire	Partage	Tag	J'aime	Rating	Commentaire	Partage	Rating
CfsSubsetEval	[nombre de sélection]	5	5	5	2	5	2	3	5	5
WrapperSubsetEval	[nombre de sélection]	5	1	5	2	4	4	4	5	5
ConsistencySubsetEval	[nombre de sélection]	5	4	5	3	5	4	4	5	5
FilteredSubsetEval	[nombre de sélection]	5	4	5	2	5	2	3	5	5
	Moyenne	5	3.5	5	2.25	4.75	3	3.5	5	5
ChiSquaredAttributeEval	[rang]	1	6	3	9	5	8	7	4	2
FilteredAttributeEval	[rang]	1	7	4	9	5	8	6	2	3
GainRatioAttributeEval	[rang]	1	7	2	9	5	8	6	4	3
InfoGainAttributeEval	[rang]	1	6	2	8	5	9	7	4	3
OneRAttributeEval	[rang]	1	7	4	8	5	9	6	3	2
ReliefFAttributeEval	[rang]	1	6	3	9	5	8	7	4	2
SVMAttributeEval	[rang]	1	6	3	9	5	8	7	4	2
SymetricalUncertEval	[rang]	1	6	3	9	5	8	7	4	2
	Moyenne	1	6.375	3	8.75	5	8.25	6.625	3.625	2.375

Tableau 5.6: Sélection des signaux sociaux temporellement dépendants par les algorithmes de sélection d'attributs (Application sur INEX SBS)

5.3.2.4 *Corrélation de temporalité des signaux avec la pertinence*

Afin d'analyser la temporalité des signaux sociaux et de déterminer s'il y a un lien entre la date de création du signal et la pertinence des documents, ainsi qu'entre le signal vis-à-vis l'âge du document et la pertinence. Pour cette étude, nous avons procédé de la même manière que celle décrite dans le chapitre 4 section 4.3.2.1.

La figure 5.1 présente les valeurs de corrélation entre la temporalité des signaux sociaux et la pertinence des documents IMDb. Par exemple, le signal *+1* biaisé par l'âge de la ressource donne un $Rho = 0.2675$, et le signal *commentaire* biaisé par sa date de création donne un $Rho = 0.2460$.

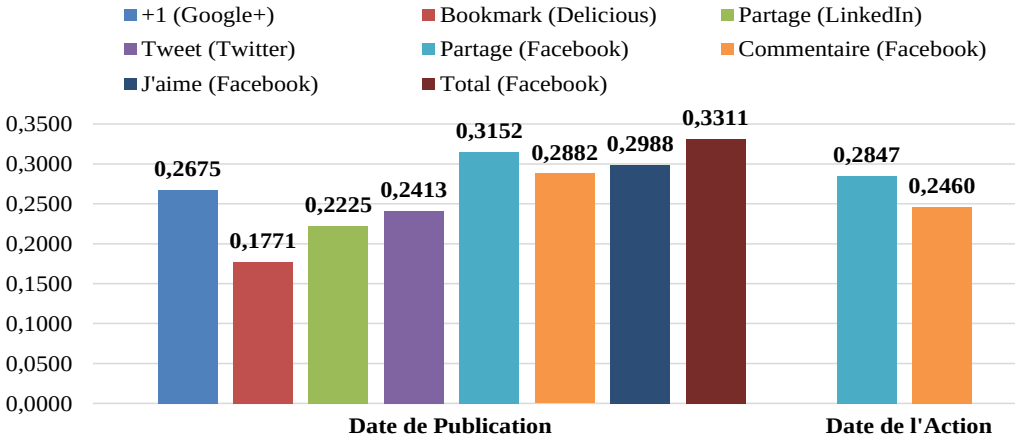

Figure 5.1: Corrélation entre la temporalité des signaux et la pertinence sur la collection IMDb

La figure 5.2 présente les valeurs de corrélations entre la temporalité des signaux et la pertinence des documents SBS. Par exemple, le signal *rating* biaisé par la date de création de chacune de ses actions enregistre une corrélation de $Rho = 0.2720$ avec la pertinence des documents, alors qu'avec la prise en compte de la date de publication de la ressource, il enregistre une valeur de corrélation de $Rho = 0.19$.

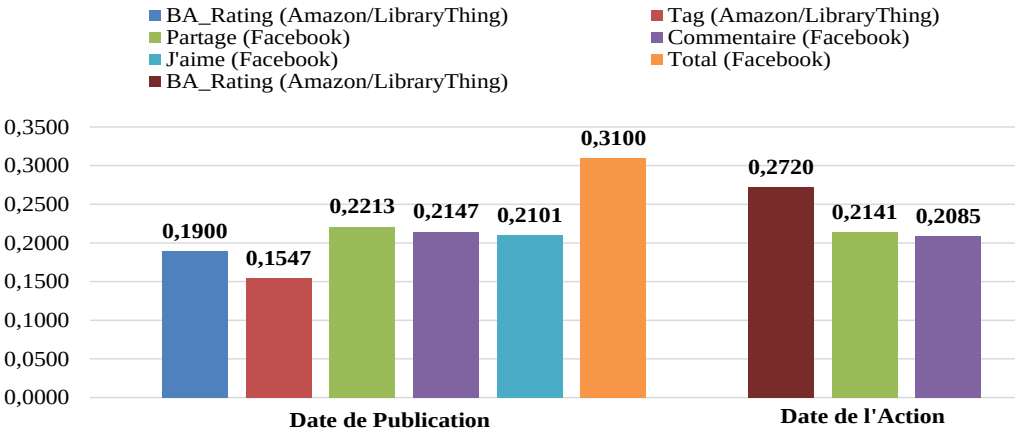

Figure 5.2: Corrélation entre la temporalité des signaux et la pertinence sur la collection SBS

- **Date de publication du document** : l'étude montre que le *partage* (0.3152 sur IMDb et 0.2213 sur SBS) a la plus forte corrélation parmi les signaux individuel,

suivi par le *j'aime* (0.2988 sur IMDb et 0.2101 sur SBS) et le *commentaire* (0.2882 sur IMDb et 0.2147 sur SBS). Cependant, le groupe de signaux de Facebook *total* (0.3311 sur IMDb et 0.31 sur SBS) possède la plus forte corrélation avec la pertinence. Enfin, la différence des résultats de corrélation entre SBS et IMDb réside sur les deux signaux *commentaire* et *j'aime*, sur SBS le *commentaire* vient en deuxième position après le *le partage* alors que sur IMDb il vient en troisième position après le *partage* et le *j'aime*.

- **Date de l'action** : concernant la prise en compte de la date de l'action, l'ordre de corrélation des signaux reste le même, c'est-à-dire que le signal *partage* apporte une corrélation (0.2847 sur IMDb et 0.2141 sur SBS) meilleure que celle apportée par le *j'aime* (0.2460 sur IMDb et 0.2085 sur SBS). Cependant, le signal *rating* enregistre une corrélation très forte par rapport à celles obtenues par le *j'aime* et le *partage*.

Enfin, l'analyse de corrélation montre que tous les signaux considérés avec leur aspect temporel (date de l'action et l'âge du document) sont positivement corrélés avec la pertinence. Cette observation renforce nos hypothèses lancées auparavant ainsi qu'elle confirme l'étude de causalité de la temporalité des signaux sur la détection des documents pertinents. Notre étude confirme également l'intérêt de la dimension temporelle des signaux sociaux exploités pour améliorer la recherche d'information : les ressources pertinentes ont un nombre élevé d'actions sociales mais aussi des interactions récentes, et ce nombre est proportionnel à la durée de vie de la ressource.

5.3.3 *Bilan*

Dans ce chapitre, nous avons étudié l'impact de la temporalité des signaux associés à une ressource, ainsi que la date de publication de cette ressource sur la performance d'un système de recherche d'information. Nous avons proposé d'estimer les probabilités a priori du document en tenant compte de ces deux facteurs. Les expérimentations menées sur les collections de données d'INEX IMDb et SBS montrent que la prise en compte des caractéristiques sociales et leurs aspects temporels dans un modèle textuel améliore la qualité des résultats de recherche renvoyés. La contribution principale de ce travail était de montrer que ces facteurs socio-temporels sont fructueux pour les systèmes de recherche d'information. Les résultats obtenus en termes de performances confirment l'intérêt de considérer le temps (date de publication de l'action et la ressource) dans la mesure de l'importance de la ressource. Une question importante que nous n'avons pas abordée peut être suffisamment est l'expérimentation du temps associé à chaque action des réseaux sociaux. Malheureusement, actuellement les APIs des réseaux sociaux ne permettent pas l'extraction de ces informations, par conséquent, nous avons étudié uniquement le temps du signal *rating* issu de la collection SBS de façon complète en exploitant la date de chaque occurrence de *rating* dans le document.

Dans le chapitre suivant, nous allons introduire un nouveau facteur lié à la diversité des signaux sociaux au sein du document. De plus, une analyse sur la qualité des signaux sera présentée.

6

QUALITÉ ET DIVERSITÉ DES SIGNAUX SOCIAUX

Introduction

Nous avons montré dans les chapitres précédents que les signaux sociaux et leur temporalité associés aux ressources Web peuvent être considérés comme une information additionnelle, qui peut jouer un rôle pour mesurer une importance a priori de la ressource indépendamment de la requête. Dans ce chapitre, nous pensons que, en plus de l'estimation de l'importance a priori basée sur la fréquence et la temporalité des signaux liés à la ressource [10, 13], la diversité des signaux au sein d'une ressource peut être aussi un indice qui peut dénoter un intérêt qui dépasse un réseau social ou une communauté.

Dans ce chapitre, nous évaluons l'impact de la diversité des facteurs de pertinence sociale, dont ce terme "diversité" est souvent utilisé dans les approches de l'état de l'art comme diversité liée au contenu textuel restitué vis-à-vis d'un besoin en information [8, 163]. Tandis que la diversité des signaux sociaux est un nouveau concept que selon notre connaissance, nous sommes les premiers à l'avoir introduit dans l'espace de la recherche d'information. Nos expérimentations sont menées sur deux types de collection, IMDb (Internet Movies Database) et SBS (Social Book Search), contenant respectivement 167438 et 2.8 millions de documents ainsi que leurs données sociales collectées à partir de plusieurs réseaux sociaux.

6.1 Hypothèse et questions de recherche

La diversité que nous voulons étudier, désigne la variété des signaux sociaux sous toutes leurs formes au sein du même document. Nous supposons que la multiplicité de traces et de signaux sociaux générés par les utilisateurs fait l'objet d'une sorte de témoignage de plusieurs communautés sur la qualité d'un document ou une ressource Web. Cette diversité d'interactions avec une ressource serait un atout pour mesurer l'importance sociale de cette ressource. Donc, nous croyons que la diversité peut être considérée comme un facteur de pertinence sociale, qui contribuerait à l'amélioration de la recherche d'information.

Nous proposons dans ce chapitre un modèle d'estimation de l'importance a priori d'une ressource qui prend en compte la diversité des signaux comme connaissance a priori dans un modèle de recherche. Les questions de recherche abordées dans ce chapitre sont les suivantes :

1. Comment estimer la diversité des signaux d'une ressource ?
2. Quel est l'impact de la diversité des signaux sur le système de RI ?
3. Est-ce que la qualité du signal est influencée par son réseau social ?

6.2 Approche basée sur la qualité et la diversité des signaux

Dans ce chapitre, nous proposons d'estimer l'importance sociale d'une ressource en exploitant la diversité des signaux sociaux associés. Afin de prendre en compte cette importance dans l'évaluation de pertinence, nous nous appuyons sur les modèles de langue pour combiner la pertinence textuelle d'une ressource vis-à-vis d'une requête et son importance socio-diverse modélisée comme une probabilité a priori. Nous reprenons le même modèle général basé sur les signaux sociaux, qui a été présenté dans le chapitre 4 section 4.2.2, mais en prenant en compte le facteur de diversité des signaux.

La diversité des signaux associés à une ressource reflètent la variété des communautés sociales qui interagissent et portent un intérêt à une ressource. Sachant que les utilisateurs ne sont pas du même réseau social, ils n'ont pas les mêmes fonctionnalités donc ils n'interagissent pas de la même façon. Cette multitude et diversification des sources qui jugent la ressource à travers ses signaux, peut être un moyen qui casse le risque de la manipulation et le suivisme aveugle sur le jugement positif ou négatif d'un contenu, un utilisateur du réseau LinkedIn ne voit pas clairement ce que pensent les utilisateurs de Facebook sur la ressource en question, de même pour les autres réseaux sociaux. Par conséquent, nous proposons d'intégrer le facteur de diversité des signaux au sein d'un document dans le modèle de recherche. Notre principale contribution décrite dans ce chapitre est sur l'estimation de $P(D)$ en exploitant la diversité des signaux au sein du document. Nous détaillons cet aspect dans la section suivante.

6.2.1 *Diversité des signaux au sein d'un document*

Nous croyons que la diversité des signaux au sein d'une ressource est un indice qui peut mesurer l'intérêt du document au-delà d'un réseau social ou d'une communauté donnée. La diversité et la distribution (répartition) quantitative des signaux sociaux au sein d'une ressource peuvent être des facteurs de pertinence, c.-à-d. une ressource dominée par un seul type de signal doit être défavorisée par rapport à une ressource ayant une équi-répartition des signaux. On propose d'évaluer cette diversité en utilisant l'indice de diversité de Shannon-Wiener [157], l'entropie. Cet indice est introduit en écologie pour mesurer la biodiversité. Il est donné par la formule suivante :

$$Diversite_s(D) = -\sum_{i=1}^{m} P_x(a_i^x) \cdot log(P_x(a_i^x)) \tag{6.1}$$

Avec $P_x(a_i^x)$ est défini dans la section précédente, et m le nombre de signaux sociaux étudiés. $x \in X$ se réfère à la propriété sociale (*popularité* ou *réputation*) estimée à partir d'un ensemble d'actions spécifiques.

L'indice de Shannon est souvent accompagné par l'indice d'équitabilité de Piélou [157]. La formule correspondante est la suivante :

$$Diversite_s^{Equit}(D) = \frac{Diversite_s(D)}{MAX(Diversite_s(D))} = \frac{Diversite_s(D)}{log(m)} \tag{6.2}$$

La probabilité a priori $P_x(D)$ est estimée en utilisant la formule 4.4 multipliée par le facteur de diversité. La formule correspondante est la suivante :

$$P_x(D) = \left(\prod_{a_i^x \in A} P_x(a_i^x) \right) \cdot Diversite_s^{Equit}(D) \tag{6.3}$$

6.2.2 *Influence des réseaux sociaux sur la qualité de leurs signaux*

L'intuition derrière cette étude est que les signaux sociaux reflétant la pertinence distinguent les documents pertinents des non-pertinents. Ces facteurs sociaux n'ont pas le même comportement avec les documents pertinents et les documents non-pertinents. Pour évaluer l'impact de l'origine d'un signal vis-à-vis son réseau de provenance, nous avons observé la distribution des signaux dans les documents pertinents et non pertinents. Si la distribution d'un signal est la même pour les documents pertinents et non-pertinents, ce signal ne permettra pas ainsi de différencier les deux classes de documents, et ne sera pas considéré comme facteur utile à cette tâche. Dans le cas contraire, lorsque la distribution des scores d'un signal est différente entre les documents pertinents et non-pertinents, ce facteur permettra dans ce cas de différencier les deux classes de documents, et il sera par conséquent considéré comme facteur utile.

Pour arriver à nos fins, nous proposons d'étudier l'ensemble des signaux ainsi que le rapport entre la qualité du signal et son réseaux social où il a été généré. Cette étude consiste à analyser la fréquence et la moyenne des signaux sociaux dans les documents pertinents et les non-pertinents. En effet, nous essayons d'expliquer le lien entre le signal et la détection des documents pertinents, et de comprendre si le signal est lié à sa fréquence ou uniquement à sa présence dans les documents pertinents.

6.3 Expérimentations et résultats

Les expérimentations ont été menées sur les deux collections IMDb et SBS. Afin de comprendre l'impact de la diversité des signaux sociaux dans un document, en plus des modèles de base textuels présentés dans les chapitres précédents, nous avons considéré comme modèles de référence, des configurations de notre approche qui ne prennent pas en compte la diversité des signaux.

6.3.1 *Résultats et discussion*

6.3.1.1 *Diversité des signaux au sein d'un document*

Les tableaux 6.1 et 6.2 récapitulent les résultats de précision@k pour $k \in \{10, 20\}$ et de nDCG, ainsi que la MAP. Nous avons évalué notre approche à travers des configurations différentes, en prenant en compte la diversité des signaux sociaux au sein d'un document. Nous avons déjà montré que la prise en compte de ces signaux sociaux indépendamment de la diversité améliore significativement la recherche d'information (voir chapitre 4) par rapport aux modèles basés uniquement sur la pertinence thématique [11, 10]. Afin de vérifier si les résultats obtenus sont statistiquement significatifs par rapport aux modèles de base, nous avons effectué le test de Student [75]. Les résultats (*) dans les tableaux 6.1 et 6.2 indiquent que les améliorations sont statistiquement

significatives avec un valeur-p (*p-value*) < 0.05. Nous discutons dans ce qui suit les résultats que nous avons obtenus.

Modèles	P@10	P@20	nDCG	MAP
Base : Sans Prise en Compte de Diversité				
TotalFacebook	0.4209	0.4102	0.5681	0.3125
Popularité	0.4316	0.4264	0.5801	0.3221
Réputation	0.4405	0.4272	0.5900	0.3260
TousLesCritres	0.4408	0.4262	0.5974	0.3300
ToutesLesProprits	0.4629	0.4509	0.6203	0.3557
Avec Prise en Compte de Diversité				
TotalFacebook^{Div}	0.4227	0.4187	0.5713	0.3167
Popularité^{Div}	0.4403	0.4288	0.5983	0.3320
Réputation^{Div}	0.4480	0.4306	0.6110	0.3319
TousLesCritères^{Div}	0.4463	0.4318	0.6174	0.3325
ToutesLesPropriétés^{Div}	0.4689	0.4563	0.6245	0.3571

Tableau 6.1: Résultats de P@k, nDCG et MAP sur la collection INEX IMDb

Modèles	P@10	P@20	nDCG	MAP
Base : Sans Prise en Compte de Diversité				
TotalFacebook	0.0958	0.0810	0.1937	0.0892
Popularité	0.0964	0.0780	0.1953	0.0890
Réputation	0.0972	0.0801	0.1974	0.0897
TousLesCritères	0.0973	0.0787	0.1981	0.0900
ToutesLesPropriétés	0.1021	0.0888	0.2004	0.0923
Avec Prise en Compte de Diversité				
TotalFacebook^{Div}	0.0960*	0.0840*	0.1945*	0.0907*
Popularité^{Div}	0.0967*	0.0862*	0.1970*	0.0916*
Réputation^{Div}	0.0988*	0.0891*	0.1994*	0.0928*
TousLesCritères^{Div}	0.1011*	0.0915*	0.2031*	0.0952*
ToutesLesPropriétés^{Div}	0.1089*	0.0976*	0.2148*	0.0984*

Tableau 6.2: Résultats de P@k, nDCG et MAP sur la collection INEX SBS

Selon les tableaux 6.1 et 6.2 le nDCG et les précisions sont en général meilleurs que le nDCG et les scores de précision lorsque la diversité est ignorée (modèles de base). La prise en compte de la diversité des signaux qui définissent la *réputation* (voir chapitre 4 section 4.10) apporte les meilleurs résultats et statistiquement significatifs comparativement à ceux obtenus par *TotalFacebook* et la *popularité* (nommés dans les tableaux TotalFacebook^{Div} et Popularité^{Div}). La diversité des avis positifs renforce la *réputation* de

la ressource par rapport à la diversité des signaux liés à la propagation de la ressource dans les réseaux sociaux (*popularité*).

Les meilleurs résultats sont obtenus par la combinaison de toutes les propriétés sociales $ToutesLesPropriétés^{Div}$ qui enregistrent un taux d'amélioration de 6.60% en MAP et 7.18% nDCG par rapport à *ToutesLesPropriétés*. Par conséquent, l'ensemble des résultats montre l'intérêt de la diversité des signaux pour donner plus de crédibilité sociale à travers la multitude des sources d'approbation et de jugement des ressources, d'où son utilité pour améliorer les systèmes de recherche d'information. En effet, si plusieurs utilisateurs de différentes communautés sociales ont trouvé qu'une ressource est utile, alors il est plus probable que d'autres utilisateurs la trouveront utile aussi. De plus, une ressource avec une diversification de sources de signaux peut faire face aux risques des signaux spams. Au niveau d'une ressource, un signal peut être généré par un robot, cependant, il est difficile de faire un robot pour tous les signaux issus de plusieurs réseaux sociaux.

6.3.1.2 *Distribution du facteur diversité et la pertinence*

Afin d'avoir un visuel sur la distribution du facteur de diversité des signaux sociaux aux sein des documents, nous avons pris l'ensemble des documents retournés par les requêtes IMDb et SBS et nous avons calculé le score de diversité pour chaque document. Ensuite, nous avons associé à chaque document les jugements de pertinence fournis par les Qrels d'INEX. Nous avons représenté le chevauchement entre la pertinence et la diversité des signaux à travers les graphiques 6.1 et 6.2.

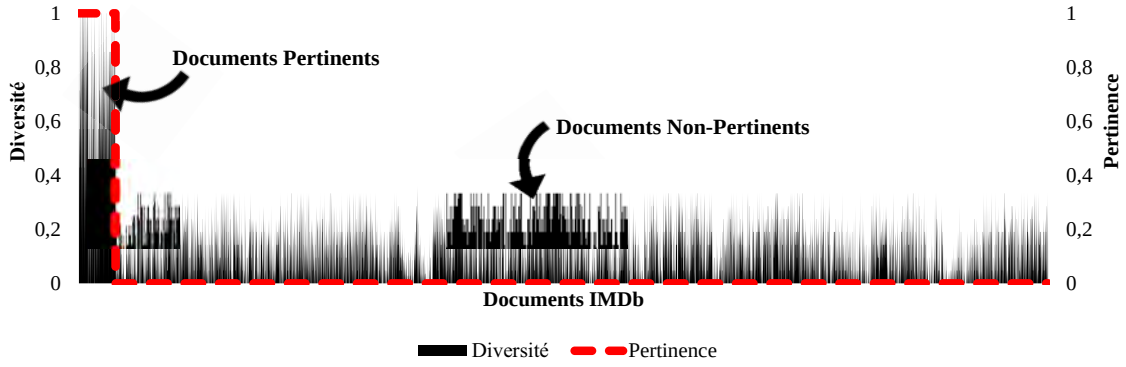

Figure 6.1: Diversité des signaux sociaux par rapport à la pertinence sur IMDb

Les figures 6.1 et 6.2 présentent les scores de diversité des signaux pour chaque document. Afin de mieux différencier entre les documents pertinents et les non-pertinents, nous avons trié les documents selon leur pertinence (en utilisant une classe binaire, 0 : non-pertinent ou 1 : pertinent). Nous remarquons que la majorité des documents pertinents (restitués) délimités par le rectangle rouge, contiennent des scores de diversité des signaux largement supérieurs à ceux enregistrés dans les documents non-pertinents, excepté un seul segment de documents non-pertinents sur SBS qui montre des scores élevés. La diversité des signaux obtient ses meilleures scores avec les documents pertinents et reflètent probablement ainsi la pertinence.

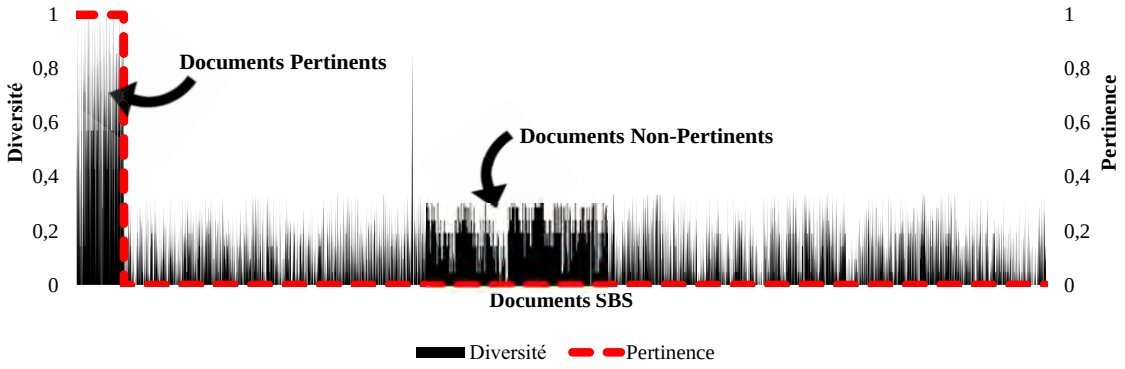

Figure 6.2: Diversité des signaux sociaux par rapport à la pertinence sur SBS

6.3.1.3 *Influence des réseaux sociaux sur la qualité des signaux*

Afin de mieux comprendre l'effet de ces signaux sociaux sur le processus de sélection des documents pertinents, nous analysons leurs distributions dans les différents documents renvoyés par les différentes requêtes issues des deux collections (30 requêtes d'IMDb et 208 requêtes de SBS).

Les tableaux 6.3 et 6.4 illustrent la distribution des différents signaux dans l'ensemble des documents (pertinents et non-pertinents) renvoyés pas les 30 requêtes d'IMDb ainsi que les 208 requêtes de SBS. Dans chacun des tableaux, nous avons 3 colonnes principales présentant des chiffres liés au nombre de signaux dans les documents (pertinents et non pertinents) listés pour chaque signal. Une première colonne liste le nombre de signaux en total et en moyenne dans les documents pertinents, la deuxième colonne liste le nombre de documents pertinents ou les signaux ne sont pas présents, et la dernière colonne présente le nombre de signaux en total et en moyenne dans les documents non pertinents.

En analysant ces tableaux, nous remarquons clairement que la fréquence moyenne des signaux dans les documents pertinents est plus élevée par rapport aux documents non-pertinents (ex. les moyennes des *j'aime* sont de 362 et 1118 actions dans les documents pertinents d'IMDb et SBS, respectivement, alors que dans les documents non-pertinents elle est de 61 actions sur IMDb et 20 actions sur SBS). Nous remarquons également que les signaux de Facebook (*partage*, *j'aime* et *commentaire*) et le *rating* capturent la majorité des documents pertinents, *partage* (2357 documents), *j'aime* (2210), *commentaire (1988)* sur IMDb et *partage* (2251), *j'aime* (2183), *rating* (2118), *commentaire* (2043) sur SBS (voir les figures 6.5 et 6.6), sachant qu'ils sont aussi nombreux dans les documents non-pertinents mais avec une moyenne beaucoup plus petite. Ceci est dû au taux d'engagement des utilisateurs sur Facebook et à sa croissance dynamique [4]. Donc, la distinction entre les documents pertinents et les documents non-pertinents est sensible beaucoup plus à la fréquence du signal, c-à-d que les documents pertinents sont caractérisés par un nombre très élevé de signaux Facebook ainsi que le *rating* d'Amazon par rapport aux documents non-pertinents (voir les figures 6.3 et 6.4).

Les signaux *tweet*, *+1* et *tag* viennent en seconde position avec une fréquence moyenne respectivement de 97, 29 et 520 actions dans les documents pertinents (voir les figures 6.5 et 6.6). Le signal issu de Delicious (*bookmark*) est le critère le plus faible parmi ces signaux, il n'est présent que dans 429 documents pertinents avec une fréquence moyenne

de 13 actions par document seulement. Pour le signal issu de LinkedIn, nous remarquons que 95% de ses actions de *partage* sont concentrées dans 601 documents pertinents avec une fréquence moyenne de 67 actions. Le nombre de documents pertinents capturés par ce signal est très faible par rapport aux signaux de Facebook. Ceci est dû au taux d'engagement sur LinkedIn qui est très faible par rapport à Facebook [4], mais le signal *partage(LIn)* représente la source la plus fiable en termes de confiance et de crédibilité par rapport aux autres signaux sociaux. En effet, LinkedIn est plutôt conçu pour les affaires et pour un public professionnel, il offre un profil et des options dans un sens qui conduit les utilisateurs à se comporter sérieusement. Alors que, Facebook est conçu beaucoup plus pour connecter avec la famille et les amis dans un sens où les utilisateurs peuvent se comporter spontanément et sans protocole. Par conséquent, sur LinkedIn, les interactions sont soigneusement créées, c'est pour cette raison que la présence du signal *partage* de LinkedIn dans un document représente un indice de pertinence.

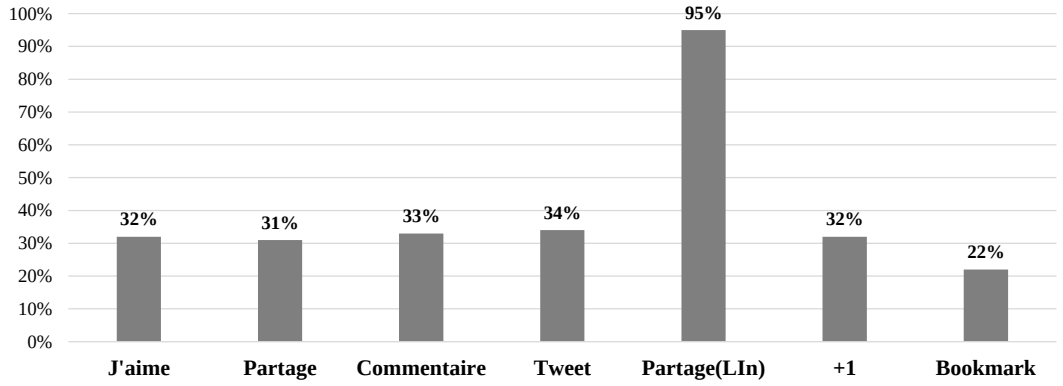

Figure 6.3: Pourcentage des signaux dans les documents pertinents IMDb

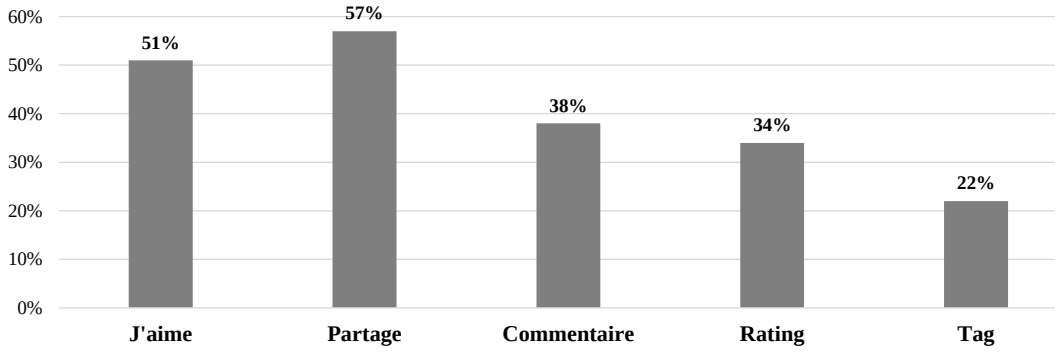

Figure 6.4: Pourcentage des signaux dans les documents pertinents SBS

	Documents pertinents contenants des signaux			Documents pertinents sans signaux	Documents non-pertinents	
	Nombre de documents	Nombre d'actions	Moyenne	Nombre de documents	Nombre d'actions	Moyenne
J'aime	2210	800458	362.1981	555	1678040	61.6133
Partage	2357	856009	363.1774	408	1862909	68.4012
Commentaire	1988	944023	474.8607	777	1901146	69.8052
Tweet	1735	168448	97.0884	1030	330784	12.1455
+1	790	23665	29.9556	1975	49727	1.8258
Bookmark	429	5654	13.1794	2336	20489	0.7523
Partage(LIn)	601	**40446**	67.2985	2164	**2341**	0.0859
Total	Documents pertinent : 2765				Documents non-pertinents : 27235	

Tableau 6.3: Statistiques sur la distribution des signaux dans les documents (pertinents et non-pertinents) issus d'IMDb retournés par les 30 requêtes

	Documents pertinents contenants des signaux			Documents pertinents sans signaux	Documents non-pertinents	
	Nombre de documents	Nombre d'actions	Moyenne	Nombre de documents	Nombre d'actions	Moyenne
J'aime	2183	2442052	1118.6248	770	2318646	20.6478
Partage	2251	3002068	1333.3700	702	2220423	19.7731
Commentaire	2043	2086534	1021.0896	910	3378477	30.0857
Rating	2118	8049416	3799.8273	835	15806702	140.7605
Tag	753	391878	520.0984	2200	1420090	12.6460
Total	Documents pertinent : 2953				Documents non-pertinents : 112295	

Tableau 6.4: Statistiques sur la distribution des signaux dans les documents (pertinents et non-pertinents) issus de SBS retournés par les 208 requêtes

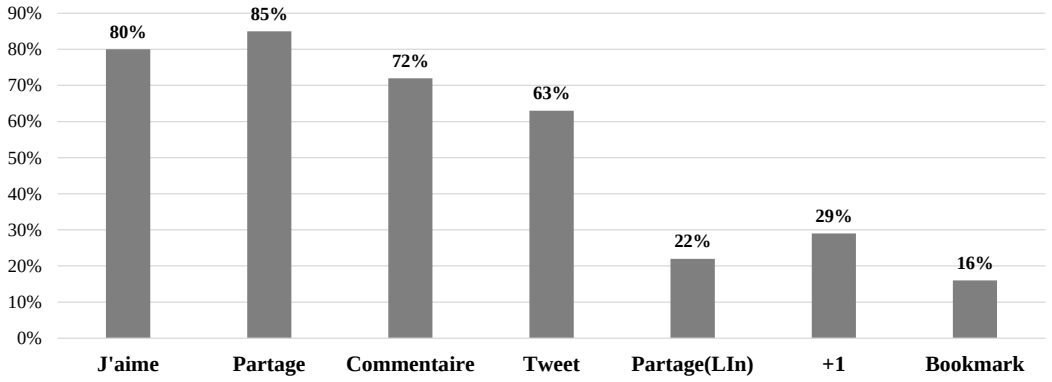

Figure 6.5: Pourcentage des documents pertinents IMDb contenants des signaux

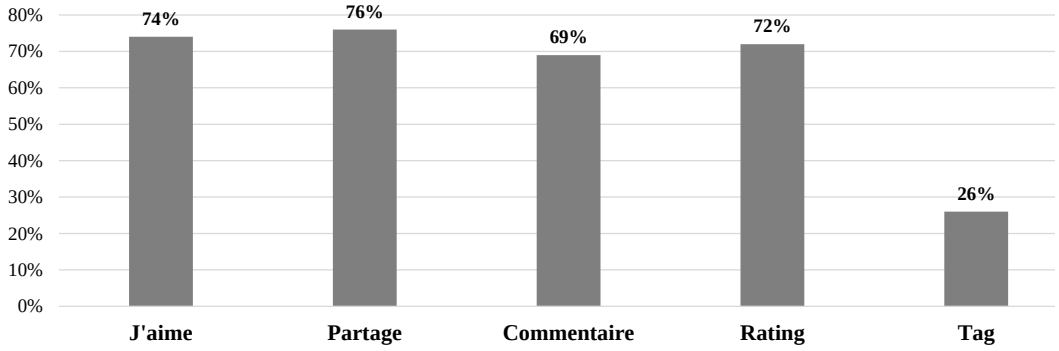

Figure 6.6: Pourcentage des documents pertinents SBS contenants des signaux

Enfin, selon cette étude statistique, nous avons constaté que chaque réseau social a sa propre influence sur la qualité de ses signaux. La qualité des signaux, provenant de Facebook, Twitter, Google+, Delicious ainsi que de Amazon/LibraryThing, dépend de leurs fréquences, plus les signaux sont fréquents sur le document, plus son importance a priori augmente. Cependant, le signal de LinkedIn ne dépend pas uniquement de sa fréquence parce qu'il a en lui-même une puissance de confiance mature par rapport aux autres signaux. Ceci revient à la maturité des utilisateurs de LinkedIn qui sont mieux réputés par rapport à d'autres utilisateurs des réseaux sociaux.

6.3.2 *Bilan*

Dans ce chapitre, nous avons étudié l'impact de la diversité des signaux au sein d'une ressource sur la performance d'un système de recherche d'information, ainsi qu'analyser la qualité de chaque signal par rapport à son réseau social. Nous avons proposé d'estimer les probabilités a priori du document en tenant compte de la diversité comme un facteur supplémentaire dans le modèle présenté dans le chapitre 4. Les expérimentations menées sur les collections de données d'INEX IMDb et SBS montrent que la prise en compte des caractéristiques sociales et leur diversité dans un modèle textuel améliore la qualité des résultats de recherche renvoyés.

Dans ce chapitre nous n'avons pas pu évaluer la diversité des signaux *bookmark*, *partage(LIn)*, *tweet* et *+1*. Ces signaux ne sont pas présents sur les documents d'INEX SBS.

Des études plus approfondies sur ce point doivent être réalisées en intégrant d'autres types de signaux sociaux. Enfin, Ces résultats nous encouragent à prendre en compte la diversité des signaux, en particulier ceux de Facebook. Il reste maintenant à trouver une méthode d'extraction rapide de ce genre d'informations à partir des réseaux sociaux.

6.4 Agrégation des résultats

Les différentes méthodes, présentées précédemment dans les chapitres 4, 5 et 6, ont été évaluées chacune à part. Dans cette section, nous voulons savoir si la prise en compte de tous les facteurs vus précédemment combinés en même temps améliore les résultats.

Les tableaux 6.5, 6.6 et 6.7 récapitulent les résultats de précision@k pour $k \in \{10, 20\}$ et de nDCG, ainsi que la MAP, obtenus par la prise en compte de la diversité et la temporalité des signaux au sein de la même fonction de score. Nous listons uniquement les configurations prenant en compte plusieurs signaux, les seuls pour lesquelles la diversité a un sens. Nous attribuons un astérisque * (forte signification par rapport aux configurations correspondantes qui prennent pas en compte l'aspect temporel combiné avec la diversité *Div*).

Modèles	P@10	P@20	nDCG	MAP
Avec Prise en Compte de Diversité et la Date de Publication de la Ressource T_D				
TotalFacebook$^{Div}_{T_D}$	0.4417	0.4289	0.5966	0.3273
TousLesCritères$^{Div}_{T_D}$	0.4568	0.4334	0.6311	0.3427
ToutesLesPropriétés$^{Div}_{T_D}$	0.4588	0.4410	0.6508	0.3491

Tableau 6.5: Résultats de P@k, nDCG et MAP sur la collection INEX IMDb

Modèles	P@10	P@20	nDCG	MAP
Avec Prise en Compte de Diversité et la Date de Publication de la Ressource T_D				
TotalFacebook$^{Div}_{T_D}$	0.0978	0.0868	0.1970	0.0932
TousLesCritères$^{Div}_{T_D}$	0.1092	0.0988	0.2095	0.0997
ToutesLesPropriétés$^{Div}_{T_D}$	0.1114	0.1031	0.2164	0.1012

Tableau 6.6: Résultats de P@k, nDCG et MAP sur la collection INEX SBS

Modèles	P@10	P@20	nDCG	MAP
Avec Prise en Compte de Diversité et la Date de l'Action T_a: *Rating*				
TousLesCritères$^{Div}_{T_a}$	0.1187*	0.1024*	0.2234*	0.1074*
ToutesLesPropriétés$^{Div}_{T_a}$	0.1214*	0.1158*	0.2287*	0.1101*

Tableau 6.7: Résultats de P@k, nDCG et MAP sur la collection INEX SBS

En comparant ces résultats présentés dans les deux tableaux 6.5, 6.6 et 6.7 avec ceux obtenus par nos expérimentations précédentes [92, 10, 11, 9, 12, 13, 14], nous constatons

que la prise en compte conjointe de la diversité et le temps (âge de la ressource, date de l'action) améliore les résultats sur les deux collections, respectivement, par rapport à toutes les autres configurations (sans prise de ces critères ou bien en les considérant de manière individuelle). Nous discutons dans ce qui suit les résultats de chacune des configurations que nous avons étudié.

6.4.1 *Prise en compte de la diversité et la date de publication du document*

Les tableaux 6.5 et 6.6 présentent les résultats obtenus en combinant la diversité et l'âge de la ressource pour estimer la pertinence a priori de cette ressource. Les résultats obtenus par la configuration $ToutesLesPropriétés^{Div}_{T_D}$ enregistrent un taux d'amélioration de +2% en termes de nDCG par rapport à $TousLesCritères^{Div}_{T_D}$ sur SBS, et +3% sur IMDb, qui prend en considération la fréquence des signaux, leur diversité et l'âge de la ressource. Par conséquent, le facteur de l'âge de la ressource devient plus efficace quand il est combiné avec la diversité des signaux. En effet, une ressource qui a eu une grande diversité de signaux sociaux durant deux jours n'a pas la même importance qu'une ressource qui a eu une diversité de signaux durant deux semaines. De plus, une ressource contenant une fréquence importante d'une variété de signaux signifie que le jugement de cette ressource est plus significatif qu'une ressource dominée par un seul signal. Une ressource avec une diversité de signaux élevée implique une propagation élevée de cette ressource chez plusieurs communautés de réseaux sociaux.

6.4.2 *Prise en compte de la diversité et la date de l'action*

Le tableau 6.7 présente les résultats obtenus en combinant la diversité et la fraîcheur du signal. Nous notons que la combinaison de la diversité avec la date des dernières actions *partage* et *commentaire* issues de Facebook n'apporte pas des améliorations, par conséquent nous présentons dan le tableau 6.7 uniquement les résultats obtenus par la prise en compte des dates du *ratings* avec la diversité des signaux sur la collection SBS.

Nous pouvons constater que la configuration $ToutesLesPropriétés^{Div}_{T_a}$ enregistre une amélioration de +2.5% en termes de nDCG par rapport à $TousLesCritères^{Div}_{T_a}$, 6.5% par rapport à $ToutesLesPropriétés^{Div}$ (voir tableau 6.2) et 20% par rapport à la prise en compte du *rating* $BA\_Rating^{T_a}$ sans la diversité (voir tableau 5.3). Par conséquent, le facteur de la date de l'action devient également plus efficace quand il est combiné avec la diversité des signaux. En effet, une ressource contenant une variété de signaux frais issus de plusieurs réseaux sociaux augmente son intérêt et sa signification temporelle en terme de pertinence vis-à-vis de l'utilisateur. Ceci signifie aussi la propagation rapide de cette ressource dans les espaces sociaux.

Partie V

CONCLUSION

Les vrais informaticiens confondent toujours Halloween et Noël car pour eux :

31 Oct = 25 Dec

— Andrew Rutherford

7 CONCLUSION

7.1 Synthèse des contributions

Le travail présenté dans cette thèse rentre dans le contexte de la recherche d'information, et se situe au carrefour de la RI classique et les réseaux sociaux. En particulier, nous abordons le problème d'intégration des signaux sociaux dans le processus de recherche d'information. Les questions de recherche auxquelles nous nous sommes intéressés concernent tout d'abord la manière d'exploiter et de traduire ces signaux de nature différente (un *commentaire*, un *tweet* sont différents d'un *j'aime*, d'un *+1* ou d'un *rating*) en des propriétés (facteurs) exploitables par les approches de RI. Ensuite comment les intégrer et les prendre en compte efficacement dans un modèle de RI.

L'examen des travaux de l'état de l'art nous a conduit à poser trois principales questions de recherche qui caractérisent nos travaux : i) la première porte sur la définition de solution pour quantifier l'importance d'une ressource à travers les signaux qui lui sont associés, et prendre en compte cette importance dans le processus de RI, ii) l'étude de l'impact de la temporalité et la diversité des signaux sociaux associés à un document sur la performance des systèmes de RI et iii) la définition d'un cadre formel permettant la combinaison de la pertinence thématique d'un document vis-à-vis d'une requête et son importance sociale.

1. Nous avons proposé dans le chapitre 4 une approche qui exploite les différents signaux sociaux pour évaluer l'importance d'une ressource. Cette importance est calculée comme probabilité de tous les signaux associées à une ressource. Les probabilités sont estimées en fonction du nombre d'occurrence d'un signal par rapport à l'ensemble des signaux associés à un document. Cette probabilité est aussi intégrée de manière naturelle dans un modèle de langue combinant la pertinence thématique et l'importance sociale d'une ressource; vue comme une probabilité a priori. Nous avons également montré que ces signaux, traduisant des actions différents des utilisateurs; jouent des rôles différents dans l'évaluation de l'importance de la ressource. Un *j'aime*, un *+1*, *bookmark*, a un sens différent que de publier un *commentaire*, ou envoyer un *tweet* mentionnant la ressource en question. Pour ce faire, nous avons proposé de regrouper ces signaux en fonction des propriétés qu'ils traduisent, soit la *popularité* ou la *réputation*, et nous avons étendu le modèle de calcul de probabilité pour prendre en compte explicitement ces propriétés.

   Nous avons réalisé plusieurs expérimentations sur deux collections de test issues d'INEX, soit IMDb et SBS (Social Book Search). Comme ces collections ne contiennent pas tous les signaux sociaux potentiellement intéressants pour nos travaux, nous avons collecté des signaux (*j'aime, partage, commentaire, tweet, bookmark, +1*, etc) via plusieurs réseaux sociaux (Facebook, Twitter, Delicious, Google+, etc).

   En préambule de l'évaluation de performance proprement dite, nous avons analysé les corrélations entre ces facteurs sociaux et la pertinence des documents.

Nous avons montré que la présence des signaux dans des documents est corrélée avec la pertinence des documents. Ces corrélations diffèrent entre les signaux, les signaux de type *j'aime* et *partage* ont des corrélations les plus fortes parmi tous les autres signaux.

En termes de performance, nous avons montré que la prise en compte des signaux individuellement améliore significativement les résultats de recherche en termes de précision et nDCG par rapport aux modèles de base considérés. Il semble que la prise en compte des signaux sous forme de propriétés est la configuration qui apporte les meilleurs résultats. Nous avons, par ailleurs, conduit d'autres expérimentations basées sur des approches supervisées. Nous avons là aussi montré que ces approches combinées avec des algorithmes de sélection d'attributs surpassent toutes les configurations que nous avons expérimentées.

Notre travail présente cependant quelques limites. D'abord, nous considérons que tous les types des signaux ont la même importance (1 *j'aime* est équivalent à 1 commentaire). Selon les études analytiques (corrélation et techniques de sélection d'attributs), il semblerait que certains soient plus importants que d'autres pour la recherche d'information.

2. Concernant notre deuxième contribution, nous avons étendu notre première approche en intégrant la temporalité des signaux associés à une ressource, ainsi que la date de publication de cette ressource. Nous avons proposé d'estimer les probabilités a priori du document en tenant compte de ces deux facteurs. Les expérimentations menées sur les collections de données d'INEX IMDb et SBS montrent que la prise en compte des caractéristiques sociales et leurs aspects temporels dans un modèle textuel améliore la qualité des résultats de recherche.

   La limite de notre approche réside dans l'impact du temps sur les signaux sociaux considérés. Actuellement les APIs des réseaux sociaux ne permettent pas l'extraction du moment où une action s'est produite. Nous avons évalué uniquement trois signaux (*j'aime, partage* et *rating*) pour lesquels la date existe. Pour les signaux, *j'aime* et *partage*, ils ont uniquement la date de la dernière action. Le seul signal pour lequel la date de chaque action est disponible, est le *rating*, et c'est pour ce signal que l'on a obtenu les meilleurs résultats en termes de performance.

3. Pour la troisième contribution, nous avons étudié l'impact de la diversité des signaux au sein d'une ressource sur la performance d'un système de RI, et nous avons aussi analysé la qualité de chaque signal par rapport à son réseau social. Nous avons proposé d'estimer la probabilité d'importance a priori du document en tenant compte de la diversité comme un facteur supplémentaire dans le modèle présenté dans le chapitre 4. La diversité est estimée comme une entropie. Les expérimentations menées sur les collections de données d'INEX IMDb et SBS montrent que la prise en compte des caractéristiques sociales et leur diversité dans un modèle textuel améliore la qualité des résultats de recherche renvoyés.

   Cependant, une question importante que nous n'avons pas abordée, concerne l'expérimentation de la diversité des signaux *bookmark*, *partage(LIn)*, *tweet* et *+1* au sein des documents de la collection INEX SBS. Ces signaux ne sont pas présents sur les documents SBS.

7.2 Perspectives

En ce qui concerne les défis de la RI sociale abordés dans le chapitre 1, nous avons abordé dans cette thèse les questions liées à la définition et l'évaluation de la pertinence sociale à travers les signaux sociaux. Dans les travaux futurs et en perspective, nous avons l'intention d'étudier à court terme :

- d'autres types de signaux sociaux (ex. *stumble* de StumbleUpon, *Pins* de Pinterest, etc) en améliorant la manière de prendre en compte le temps dans le calcul de l'intérêt de la ressource,
- l'importance des réseaux sociaux et des acteurs sociaux de ces signaux et leur impact sur la pertinence. Nous envisageons d'approfondir nos recherches pour évaluer la qualité des signaux (en termes d'importance et de pertinence) à partir de l'évaluation de leur acteurs (créateurs), ainsi que la qualité et la spécificité de leur réseaux sociaux (environnement de leur création). Un *j'aime* issu de Facebook n'est pas le même que celui issu de LinkedIn. De même, Facebook n'est pas LinkedIn.

  Il existe un déséquilibre important entre certains réseaux sociaux en termes de nombre d'utilisateurs actifs, ainsi que la quantité des données sociales générées. En plus de ces dernières, la diversité et l'hétérogénéité des réseaux et des acteurs influencent sur la qualité des signaux, en termes de fréquence, de signification et de taux d'engagement sur un laps de temps donné. Par conséquent, parmi les questions qu'on peut poser nous pouvons lister : comment normaliser ce déséquilibre et contrôler cette hétérogénéité ? comment arriver à mieux évaluer l'impact du signal, en prenant en compte son créateur et son réseau social ? quelle est la méthode la plus appropriée pour pondérer ces signaux sociaux ?
- la polarité du contenu textuelle des signaux (ex. *commentaire*, *tweet*, *statut*). La partie textuelle a été prise dans le cas de nos travaux par un simple comptage de présence ou d'absence du signal sans regarder son contenu sémantique (par exemple la polarité du texte, *commentaire* positif ou négatif, ou neutre). Nous souhaitons analyser ce contenu et le transformer en un *vote* (ou en un *rating*) de la ressource,
- la personnalisation des résultats de recherche en fonction des interactions sociales de l'utilisateur et leur temporalité (temporalité des intérêts), c-à-d la construction de profils en se basant sur les ressources auxquelles les signaux de l'utilisateur ont été associés à travers le temps. En effet, le profil de l'utilisateur changera dynamiquement selon le croisement de ses interactions dans le temps.

et à long terme, nous envisageons d'exploiter les signaux sociaux à d'autres cadres, par exemple :

- l'analyse du comportement des utilisateurs des réseaux sociaux lors d'une catastrophe dans une région précise. Les contenus générés (*commentaires*, *statuts*, *tweet*, etc) par les utilisateurs en temps réel sur les différents réseaux sociaux peuvent faire l'objet d'une source d'information immédiate à exploiter dans des systèmes spécifiques. Les utilisateurs des réseaux sociaux peuvent approuver ou désapprouver ces informations à travers les signaux tels que le *j'aime* et le *+1*. Ce type de

système permettra aux intervenants en urgence (ex. protection civile) de détecter les informations pertinentes en temps réel, d'avoir une vision globale de tout ce qui ce passe et d'agir rapidement dans les régions réclamant une intervention immédiate selon des priorités (ex. gravité des dégâts, zones sensibles, etc)

BIBLIOGRAPHY

[1] *SWSM '09: Proceedings of the 2Nd ACM Workshop on Social Web Search and Mining*, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-806-3. 605099.

[2] Lada Adamic and Eytan Adar. How to search a social network. *Social networks*, 27(3):187–203, 2005.

[3] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 183–194. ACM, 2008.

[4] O. Alonso and V. Kandylas. A study on placement of social buttons in web pages. *arXiv*, 2014.

[5] Omar Alonso, Michael Gamon, Kevin Haas, and Patrick Pantel. Diversity and relevance in social search, 2012.

[6] Gianni Amati, Giuseppe Amodeo, Marco Bianchi, Giuseppe Marcone, Fondazione Ugo Bordoni, Carlo Gaibisso, Giorgio Gambosi, Alessandro Celi, Cesidio Di Nicola, and Michele Flammini. Fub, iasi-cnr, univaq at trec 2011 microblog track. In *TREC*, 2011.

[7] Nawal Ould Amer, Philippe Mulhem, and 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.*, pages 55–70, 2015. URL http://coria2015.lip6.fr/wp-content/uploads/2015/03/6.pdf.

[8] Albert Angel and Nick Koudas. Efficient diversity-aware search. In *SIGMOD*, pages 781–792. ACM, 2011.

[9] Ismail Badache and Mohand Boughanem. Exploitation des signaux sociaux pour estimer la pertinence a priori d'une ressource. In *CORIA 2014 - Conférence en Recherche d'Infomations et Applications - 11th French Information Retrieval Conference, Nancy, France, March 18-21, 2015.*, pages 171–186, 2014. URL http://asso-aria.org/coria/2014/coria/CORIA-14.pdf.

[10] Ismail Badache and Mohand Boughanem. Social Priors to Estimate Relevance of a Resource . In *ACM Information Interaction in context (IIiX), Regensburg, Germany, 26/08/2014-29/08/2014*, pages 106–114, http://www.acm.org/, août 2014. ACM. URL http://www.irit.fr/publis/SIG/2014_IIiX_BADACHE.pdf.

[11] Ismail Badache and Mohand Boughanem. Harnessing social signals to enhance a search. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Warsaw, Poland, August 11-14, 2014 - Volume II*, pages 303–309, 2014. doi: 10.1109/WI-IAT.2014.48. URL http://dx.doi.org/10.1109/WI-IAT.2014.48.

[12] Ismail Badache and Mohand Boughanem. Pertinence a priori basée sur la diversité et la temporalité des signaux sociaux. In *CORIA 2015 - Conférence en Recherche d'Infomations et Applications - 12th French Information Retrieval Conference, Paris, France, March 18-20, 2015.*, pages 23–38, 2015. URL http://coria2015.lip6.fr/wp-content/uploads/2015/03/58.pdf.

[13] Ismail Badache and Mohand Boughanem. Document priors based on time-sensitive social signals. In *Advances in Information Retrieval - 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 - April 2, 2015. Proceedings*, pages 617–622, 2015. doi: 10.1007/978-3-319-16354-3_68. URL http://dx.doi.org/10.1007/978-3-319-16354-3_68.

[14] Ismail Badache and Mohand Boughanem. A priori relevance based on quality and diversity of social signals. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, pages 731–734, 2015. doi: 10.1145/2766462.2767807. URL http://doi.acm.org/10.1145/2766462.2767807.

[15] Ricardo Baeza-Yates. User generated content: how good is it? In *Proceedings of the 3rd workshop on Information credibility on the web*, pages 1–2. ACM, 2009.

[16] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[17] Krisztian Balog, Maarten De Rijke, and Wouter Weerkamp. Bloggers as experts: feed distillation using expert retrieval models. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 753–754. ACM, 2008.

[18] Shenghua Bao, Guirong Xue, Xiaoyuan Wu, Yong Yu, Ben Fei, and Zhong Su. Optimizing web search using social annotations. In *Proceedings of the 16th international conference on World Wide Web*, pages 501–510. ACM, 2007.

[19] Nicholas J Belkin. Cognitive models and information transfer. *Social Science Information Studies*, 4(2):111–129, 1984.

[20] Nicholas J Belkin and W Bruce Croft. Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM*, 35(12):29–38, 1992.

[21] Lamjed Ben Jabeur. *Leveraging social relevance: Using social networks to enhance literature access and microblog search*. PhD thesis, Université de Toulouse, Université Toulouse III-Paul Sabatier, 2013.

[22] Matthias Bender, Tom Crecelius, Mouna Kacimi, Sebastian Michel, Thomas Neumann, Josiane Xavier Parreira, Ralf Schenkel, and Gerhard Weikum. Exploiting social relations for query expansion and result ranking. In *Data engineering workshop, 2008. ICDEW 2008. IEEE 24th International Conference on*, pages 501–506. IEEE, 2008.

[23] Marin Bertier, Rachid Guerraoui, Vincent Leroy, and Anne-Marie Kermarrec. Toward personalized query expansion. In *Proceedings of the Second ACM EuroSys Workshop on Social Network Systems*, pages 7–12. ACM, 2009.

[24] Sumit Bhatia and Prasenjit Mitra. Adopting inference networks for online thread retrieval. In *AAAI*, volume 10, pages 1300–1305, 2010.

[25] Claudio Biancalana, Alessandro Micarelli, and Claudio Squarcella. Nereau: a social approach to query expansion. In *Proceedings of the 10th ACM workshop on Web information and data management*, pages 95–102. ACM, 2008.

[26] Team Bing. Comment and like stuff on facebook directly from bing. *Microsoft*, 2013. URL https://blogs.bing.com/search/2013/05/10/comment-and-like-stuff-on-facebook-directly-from-bing/.

[27] Kerstin Bischoff, Claudiu S Firan, Wolfgang Nejdl, and Raluca Paiu. Can all tags be used for search? In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 193–202. ACM, 2008.

[28] Freimut Bodendorf and Carolin Kaiser. Detecting opinion leaders and trends in online social networks. In *Proceedings of the 2nd ACM workshop on Social web search and mining*, pages 65–68. ACM, 2009.

[29] Toine Bogers, Marijn Koolen, Kamps Jaap, Gabriella Kazai, and Michael Preminger. Overview of the inex 2014 social book search track. In *Conference and Labs of the Evaluation Forum*, pages 462–479.

[30] Sorana-Daniela Bolboaca and Lorentz Jäntschi. Pearson versus spearman, kendall's tau correlation analysis on structure-activity relationships of biologic active compounds. *Leonardo Journal of Sciences*, 5(9):179–200, 2006.

[31] Mohamed Reda Bouadjenek. *Infrastructure and Algorithms for Information Retrieval Based On Social Network Analysis/Mining*. PhD thesis, Versailles-Saint-Quentin-en-Yvelines, 2013.

[32] Mohamed Reda Bouadjenek, Hakim Hacid, and Mokrane Bouzeghoub. Sopra: A new social personalized ranking function for improving web search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 861–864. ACM, 2013.

[33] Mohand Boughanem and Jacques Savoy. *Recherche d'information: état des lieux et perspectives*. Hermès science publ., 2008.

[34] Mohand Boughanem, Wessel Kraaij, and Jian-Yun Nie. Modeles de langue pour la recherche d'information. *Les systemes de recherche d'informations*, pages 163–182, 2004.

[35] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[36] Andrei Broder. A taxonomy of web search. In *ACM Sigir forum*, volume 36, pages 3–10. ACM, 2002.

[37] Marco Buijs and Marco R. Spruit. The social score - determining the relative importance of webpages based on online social signals. In *KDIR 2014 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, Rome, Italy, 21 - 24 October, 2014*, pages 71–77. SciTePress, 2014. doi: 10.5220/0005076400710077. URL http://dx.doi.org/10.5220/0005076400710077.

[38] Vannevar Bush. As we may think. *ACM SIGPC Notes*, 1(4):36–44, 1979.

[39] Vannevar Bush and As We May Think. The atlantic monthly. *As we may think*, 176 (1):101–108, 1945.

[40] Peng Cao, Jinhua Gao, Yubao Yu, Shenghua Liu, Yue Liu, and Xueqi Cheng. Ictnet at microblog track trec 2011. In *TREC*. Citeseer, 2011.

[41] David Carmel, Haggai Roitman, and Elad Yom-Tov. Who tags the tags?: a framework for bookmark weighting. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1577–1580. ACM, 2009.

[42] David Carmel, Naama Zwerdling, Ido Guy, Shila Ofek-Koifman, Nadav Har'El, Inbal Ronen, Erel Uziel, Sivan Yogev, and Sergey Chernov. Personalized social search based on the user's social network. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1227–1236. ACM, 2009.

[43] David Carmel, Haggai Roitman, and Elad Yom-Tov. Social bookmark weighting for search and recommendation. *The VLDB Journal—The International Journal on Very Large Data Bases*, 19(6):761–775, 2010.

[44] S. V. Chelaru, C. Orellana-Rodriguez, and I. S. Altingovde. Can social features help learning to rank youtube videos? In *WISE*, pages 552–566, Berlin, 2012.

[45] Sergiu Chelaru, Claudia Orellana-Rodriguez, and Ismail Sengor Altingovde. How useful is social feedback for learning to rank youtube videos? *World Wide Web*, pages 1–29, 2013.

[46] Shihn-Yuarn Chen and Yi Zhang. Improve web search ranking with social tagging. In *1st International Workshop on Mining Social Media*, 2009.

[47] Stanley F Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393, 1999.

[48] Fuxing Cheng, Xin Zhang, Ben He, Tiejian Luo, and Wenjie Wang. A survey of learning to rank for real-time twitter search. In *Pervasive computing and the networked world*, pages 150–164. Springer, 2013.

[49] Xu Cheng, Cameron Dale, and Jiangchuan Liu. Statistics and social network of youtube videos. In *Quality of Service, 2008. IWQoS 2008. 16th International Workshop on*, pages 229–238. IEEE, 2008.

[50] Heting Chu. *Information representation and retrieval in the digital age*. Information Today, Inc., 2003.

[51] Cyril Cleverdon, Jack Mills, and Michael Keen. Factors determining the performance of indexing systems volume 1. design. *Cranfield: College of Aeronautics*, 1966.

[52] W Bruce Croft, Donald Metzler, and Trevor Strohman. *Search engines: Information retrieval in practice*. Addison-Wesley Reading, 2015.

[53] Sally Jo Cunningham and David M Nichols. How people find videos. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 201–210. ACM, 2008.

[54] Firas Damak. *Etude des facteurs de pertinence dans la recherche de microblogs.* PhD thesis, Université Paul Sabatier, 2014.

[55] Firas Damak, Lamjed Ben Jabeur, Guillaume Cabanac, Karen Pinel-Sauvagnat, Lynda Tamine, and Mohand Boughanem. Irit at trec microblog 2011. In *TREC*. Citeseer, 2011.

[56] Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg, and Lillian Lee. How opinions are received by online communities: a case study on amazon. com helpfulness votes. In *Proceedings of the 18th international conference on World wide web*, pages 141–150. ACM, 2009.

[57] Brian D Davison. Topical locality in the web. In *SIGIR*, pages 272–279. ACM, 2000.

[58] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[59] David Dearman and Khai N Truong. Why users of yahoo!: answers do not answer questions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 329–332. ACM, 2010.

[60] Hongbo Deng, Irwin King, and Michael R Lyu. Formal models for expert finding on dblp bibliography data. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 163–172. IEEE, 2008.

[61] Nicholas A Diakopoulos and David A Shamma. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1195–1198. ACM, 2010.

[62] Pavel A Dmitriev, Nadav Eiron, Marcus Fontoura, and Eugene Shekita. Using annotations in enterprise search. In *Proceedings of the 15th international conference on World Wide Web*, pages 811–817. ACM, 2006.

[63] Anlei Dong, Ruiqiang Zhang, Pranam Kolari, Jing Bai, Fernando Diaz, Yi Chang, Zhaohui Zheng, and Hongyuan Zha. Time is of the essence: improving recency ranking using twitter data. In *Proceedings of the 19th international conference on World wide web*, pages 331–340. ACM, 2010.

[64] Yerach Doytsher, Ben Galon, and Yaron Kanza. Querying geo-social data by bridging spatial networks and social networks. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, pages 39–46. ACM, 2010.

[65] Gideon Dror, Yehuda Koren, Yoelle Maarek, and Idan Szpektor. I want to answer; who has a question?: Yahoo! answers recommender system. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1109–1117. ACM, 2011.

[66] Yajuan Duan, Long Jiang, Tao Qin, Ming Zhou, and Heung-Yeung Shum. An empirical study on learning to rank of tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 295–303. Association for Computational Linguistics, 2010.

[67] Miles Efron. Information search and retrieval in microblogs. *Journal of the American Society for Information Science and Technology*, 62(6):996–1008, 2011.

[68] Brynn M Evans and Ed H Chi. Towards a model of understanding social search. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 485–494. ACM, 2008.

[69] Luc Fayard. dicocitations - cookies. 2014. URL http://www.dicocitations.com/citation/informatique/1/0.php.

[70] Luc Fayard. dicocitations - cookies. 2014. URL http://www.blogdumoderateur.com/55-citations-inspirantes-sur-les-medias-sociaux/.

[71] Yupeng Fu, Rongjing Xiang, Yiqun Liu, Min Zhang, and Shaoping Ma. Finding experts using social network analysis. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 77–80. IEEE Computer Society, 2007.

[72] George W. Furnas, Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, 1987.

[73] Salton Gerard and J McGILL Michael. Introduction to modern information retrieval, 1983.

[74] Cyril W Gleverdon and Cyril W Cleverdon. Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. 1962.

[75] William Sealy Gosset. The probable error of a mean. *Biometrika*, 6(1):1–25, March 1908.

[76] Venkat N Gudivada, Vijay V Raghavan, William I Grosky, and Rajesh Kasanagottu. Information retrieval on the world wide web. *IEEE Internet Computing*, (5):58–68, 1997.

[77] Manish Gupta, Rui Li, Zhijun Yin, and Jiawei Han. Survey on social tagging techniques. *ACM SIGKDD Explorations Newsletter*, 12(1):58–72, 2010.

[78] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[79] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[80] Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[81] Mark A. Hall and Geoffrey Holmes. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Trans. on Knowl. and Data Eng.*, 15(6):1437–1447, November 2003. ISSN 1041-4347. doi: 10.1109/TKDE.2003.1245283. URL http://dx.doi.org/10.1109/TKDE.2003.1245283.

[82] Robert A Hanneman and Mark Riddle. Introduction to social network methods, 2005.

[83] Donna Harman. Towards interactive query expansion. In *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 321–331. ACM, 1988.

[84] Brent Hecht, Jaime Teevan, Meredith Ringel Morris, and Daniel J Liebling. Searchbuddies: Bringing search engines into the conversation. *ICWSM*, 12:138–145, 2012.

[85] Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina. Can social bookmarking improve web search? In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 195–206. ACM, 2008.

[86] Paul Heymann, Daniel Ramage, and Hector Garcia-Molina. Social tag prediction. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 531–538. ACM, 2008.

[87] D. Hiemstra. A linguistically motivated probabilistic model of information retrieval. In *ECDL Conference*, volume 1513 of *Lecture Notes in Computer Science*, pages 569–584, 1998. ISBN 978-3-540-65101-7.

[88] Djoerd Hiemstra and Wessel Kraaij. Twenty-one at trec-7: Ad-hoc and cross-language track. 1999.

[89] Robert C Holte. Very simple classification rules perform well on most commonly used datasets. *Machine learning*, 11(1):63–90, 1993.

[90] L. Hong, O. Dan, and B. D.f Davison. Predicting popular messages in twitter. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 57–58, NY, USA, 2011. ACM.

[91] Peter Hui and Michelle Gregory. Quantifying sentiment and influence in blogspaces. In *Proceedings of the First Workshop on Social Media Analytics*, pages 53–61. ACM, 2010.

[92] Melanie Imhof, Ismail Badache, and Mohand Boughanem. Multimodal social book search. In *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization-Fourth International Conference of the Cross-Language Evaluation Forum*, pages 9–pages, 2015.

[93] Yoshiyuki Inagaki, Narayanan Sadagopan, Georges Dupret, Anlei Dong, Ciya Liao, Yi Chang, and Zhaohui Zheng. Session based click features for recency ranking. In *AAAI*, volume 10, pages 1334–1339, 2010.

[94] Bogdan Ionescu, Adrian Popescu, Mihai Lupu, Alexandru L Gınsca, and Henning Müller. Retrieving diverse social images at mediaeval 2014: Challenge, dataset and evaluation. In *MediaEval 2014 Workshop, Barcelona, Spain*, 2014.

[95] Bogdan Ionescu, Adrian Popescu, Anca-Livia Radu, and Henning Müller. Result diversification in social image retrieval: a benchmarking framework. *Multimedia Tools and Applications*, pages 1–31, 2014.

[96] Bogdan Ionescu, Anca-Livia Radu, María Menéndez, Henning Müller, Adrian Popescu, and Babak Loni. Div400: a social image retrieval result diversification dataset. In *Proceedings of the 5th ACM Multimedia Systems Conference*, pages 29–34. ACM, 2014.

[97] Bogdan Ionescu, Adrian Popescu, Mihai Lupu, Alexandru Lucian Gînsca, Bogdan Boteanu, and Henning Müller. Div150cred: A social image retrieval result diversification with user tagging credibility dataset. *ACM Multimedia Systems-MMSys, Portland, Oregon, USA*, 2015.

[98] Lamjed Ben Jabeur, Lynda Tamine, and Mohand Boughanem. Featured tweet search: Modeling time and social influence for microblog retrieval. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on*, volume 1, pages 166–173. IEEE, 2012.

[99] Lamjed Ben Jabeur, Firas Damak, Lynda Tamine, Guillaume Cabanac, Karen Pinel-Sauvagnat, and Mohand Boughanem. Irit at trec microblog track 2013. In *Text REtrieval Conference-TREC 2013*, page 0, 2013.

[100] Bernard J Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, 60(11):2169–2188, 2009.

[101] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002.

[102] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[103] Frederick Jelinek. Interpolated estimation of markov source parameters from sparse data. *Pattern recognition in practice*, 1980.

[104] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.

[105] Thorsten Joachims. A support vector method for multivariate performance measures. In *Proceedings of the 22nd international conference on Machine learning*, pages 377–384. ACM, 2005.

[106] K.S. Jones, C.J. Van Rijsbergen, British Library. Research, and Development Department. *Report on the Need for and Provision of an Ideal Information Retrieval Test Collection*. British Library Research and Development reports. University Computer Laboratory, 1975. URL https://books.google.fr/books?id=cuGnSgAACAAJ.

[107] T Joyce and RM Needham. The thesaurus approach to information retrieval. *American Documentation*, 9(3):192–197, 1958.

[108] Bastian Karweg, Christian Hütter, and Klemens Böhm. Evolving social search based on bookmarks and status messages from social networks. In *CIKM*, pages 1825–1834. ACM, 2011.

[109] G. Kazai and N. Milic-Frayling. Effects of social approval votes on search performance. In *Information Technology: New Generations, ITNG '09.*, pages 1554–1559, 2009. doi: 10.1109/ITNG.2009.281.

[110] Gabriella Kazai and Natasa Milic-Frayling. Trust, authority and popularity in social information retrieval. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 1503–1504. ACM, 2008.

[111] David A Kenny. Correlation and causation. *New York: Wiley. Koeske, GF, and Koeske, RD (1989)'Construct validity of the Maslach Burnout Inventory: A critical review and reconceptualization.'Journal of Applied Behavioral Science*, 25(2):131–144, 1979.

[112] Ali Khodaei and Omar Alonso. Temporally-aware signals for social search. In *SIGIR Workshop on Time-aware Information Access*, 2012. URL http://ceur-ws.org/Vol-842/crowdsearch-khodaei.pdf.

[113] Ali Khodaei and Cyrus Shahabi. Social-textual search and ranking. In *WWW 2012 CrowdSearch workshop*, 2012.

[114] Lars Kirchhoff, Katarina Stanoevska-Slabeva, Thomas Nicolai, Matthes Fleck, and K Stanoevska. Using social network analysis to enhance information retrieval systems. *Applications of social network analysis (ASNA), Zurich*, 7:1–21, 2008.

[115] Sebastian Marius Kirsch, Melanie Gnasa, and Armin B Cremers. Beyond the web: Retrieval in social information spaces. In *Advances in Information Retrieval*, pages 84–95. Springer, 2006.

[116] Donald E. Knuth. Computer Programming as an Art. *Communications of the ACM*, 17(12):667–673, December 1974.

[117] Donald Ervin Knuth. *The art of computer programming: sorting and searching*, volume 3. Pearson Education, 1998.

[118] Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324, 1997.

[119] Marijn Koolen, Toine Bogers, and Jaap Kamps. Overview of the sbs 2015 suggestion track.

[120] Marijn Koolen, Gabriella Kazai, and Nick Craswell. Wikipedia pages as entry points for book search. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 44–53. ACM, 2009.

[121] Marijn Koolen, Toine Bogers, Maria Gäde, Mark Hall, Hugo Huurdeman, Jaap Kamps, Mette Skov, Elaine Toms, and David Walsh. Overview of the clef 2015 social book search lab. pages 545–564, 2015.

[122] Gerald Kowalski. Information retrieval systems: theory and implementation. *Computers and Mathematics with Applications*, 5(35):133, 1998.

[123] Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt. A few chirps about twitter. In *Proceedings of the first workshop on Online social networks*, pages 19–24. ACM, 2008.

[124] Yinghao Li, Wing Pong Robert Luk, Kei Shiu Edward Ho, and Fu Lai Korris Chung. Improving weak ad-hoc queries using wikipedia asexternal corpus. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 797–798. ACM, 2007.

[125] Jimmy Lin and Miles Efron. Overview of the trec-2013 microblog track. In *Proceedings of TREC*, volume 2013, 2013.

[126] Huan Liu and Rudy Setiono. Chi2: Feature selection and discretization of numeric attributes. In *2012 IEEE 24th International Conference on Tools with Artificial Intelligence*, pages 388–388. IEEE Computer Society, 1995.

[127] Huan Liu and Rudy Setiono. A probabilistic approach to feature selection-a filter solution. In *ICML*, volume 96, pages 319–327. Citeseer, 1996.

[128] Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

[129] Hans Peter Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, 1(4):309–317, 1957.

[130] Zhunchen Luo, Miles Osborne, Sasa Petrovic, and Ting Wang. Improving twitter retrieval by exploiting structural information. In *AAAI*, 2012.

[131] Nan Ma, Jiancheng Guan, and Yi Zhao. Bringing pagerank to the citation analysis. *Information Processing & Management*, 44(2):800–810, 2008.

[132] Craig Macdonald and Iadh Ounis. The trec blogs06 collection: Creating and analysing a blog test collection. *Department of Computer Science, University of Glasgow Tech Report TR-2006-224*, 1:3–1, 2006.

[133] Craig Macdonald and Iadh Ounis. Voting for candidates: adapting data fusion techniques for an expert search task. In *CIKM*, pages 387–396. ACM, 2006.

[134] Craig Macdonald and Iadh Ounis. Voting techniques for expert search. *Knowledge and information systems*, 16(3):259–280, 2008.

[135] David JC MacKay and Linda C Bauman Peto. A hierarchical dirichlet language model. *Natural language engineering*, 1(03):289–308, 1995.

[136] Matteo Magnani and Danilo Montesi. Toward conversation retrieval. In *Digital Libraries*, pages 173–182. Springer, 2010.

[137] Matteo Magnani, Danilo Montesi, and Luca Rossi. Conversation retrieval for microblogging sites. *Information retrieval*, 15(3-4):354–372, 2012.

[138] Thomas Mandl. Recent developments in the evaluation of information retrieval systems: Moving towards diversity and practical relevance. *Informatica*, 32(1), 2008.

[139] Tober Marcus, Furch Daniel, Londenberg Kai, Massaron Lucas, and Jan Grundmann. Search ranking factors and rank correlations. *SearchMetrics*, 2015. URL http://www.searchmetrics.com/knowledge-base/ranking-factors/.

[140] Melvin Earl Maron and John L Kuhns. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM (JACM)*, 7(3):216–244, 1960.

[141] James Mayfield and Paul McNamee. Single n-gram stemming. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 415–416. ACM, 2003.

[142] Donald Metzler and Congxing Cai. Usc/isi at trec 2011: Microblog track. In *TREC*. Citeseer, 2011.

[143] David RH Miller, Tim Leek, and Richard M Schwartz. A hidden markov model information retrieval system. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 214–221. ACM, 1999.

[144] Calvin N Mooers. The next twenty years in information retrieval; some goals and predictions. *American Documentation*, 11(3):229–236, 1960.

[145] M. R. Morris and J. Teevan. Exploring the complementary roles of social networks and search engines. In *Human-Computer Interaction Consortium Workshop*, pages 1–10, Asilomar, CA, 2012.

[146] Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich. What do people ask their social networks, and why?: a survey study of status message q&a behavior. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1739–1748. ACM, 2010.

[147] Rinkesh Nagmoti, Ankur Teredesai, and Martine De Cock. Ranking approaches for microblog search. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 153–157. IEEE, 2010.

[148] Michael G Noll and Christoph Meinel. *Web search personalization via social bookmarking and tagging*. Springer, 2007.

[149] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11(122-129):1–2, 2010.

[150] Brendan O'Connor, Michel Krieger, and David Ahn. Tweetmotif: Exploratory search and topic summarization for twitter. In *ICWSM*, 2010.

[151] Iadh Ounis, Craig Macdonald, and Ian Soboroff. Overview of the trec-2008 blog track. Technical report, DTIC Document, 2008.

[152] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In *WWW*, pages 161–172, Brisbane, Australia, 1998.

[153] Aditya Pal and Scott Counts. Identifying topical authorities in microblogs. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 45–54. ACM, 2011.

[154] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[155] P. Pantel, M. Gamon, O. Alonso, and K. Haas. Social annotations: Utility and prediction modeling. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 285–294, New York, NY, USA, 2012. ACM.

[156] Jie Peng, Craig Macdonald, Ben He, and Iadh Ounis. Combination of document priors in web information retrieval. In *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, RIAO '07, pages 596–611, Paris, France, France, 2007. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE. URL http://dl.acm.org/citation.cfm?id=1931390.1931446.

[157] EC Pielou. Shannon's formula as a measure of specific diversity: its use and misuse. *American Naturalist*, pages 463–465, 1966.

[158] Ari Pirkola. Morphological typology of languages for ir. *Journal of Documentation*, 57(3):330–348, 2001.

[159] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *SIGIR'98*, pages 275–281, USA, 1998. ACM. ISBN 1-58113-015-5. doi: 10.1145/290941.291008. URL http://doi.acm.org/10.1145/290941.291008.

[160] Jay M Ponte and W Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281. ACM, 1998.

[161] Martin F Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[162] MF Porter. Readings in information retrieval. 1997.

[163] Lu Qin, Jeffrey Xu Yu, and Lijun Chang. Diversifying top-k results. *VLDB Endowment*, 5(11):1124–1135, 2012.

[164] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993. ISBN 1-55860-238-0.

[165] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[166] M. Raza. A new level of social search: Discovering the user's opinion before he make one. In *Microsoft Research*, pages 1–6, 2011.

[167] B Ribeiro-Neto and R Baeza-Yates. Modern information retrieval: The concepts and technology behind search. 2011.

[168] CJ Rijsbergen. Van: Information retrieval. *London: Butterwoths*, 1979.

[169] Stephen E Robertson. The methodology of information retrieval experiment. *Information retrieval experiment*, 1:9–31, 1981.

[170] Stephen E Robertson. On term selection for query expansion. *Journal of documentation*, 46(4):359–364, 1990.

[171] Stephen E Robertson and Steve 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*, pages 232–241. Springer-Verlag New York, Inc., 1994.

[172] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at trec-3. *NIST SPECIAL PUBLICATION SP*, pages 109–109, 1995.

[173] Marko Robnik-Šikonja and Igor Kononenko. An adaptation of relief for attribute estimation in regression. In *Machine Learning: Proceedings of the Fourteenth International Conference (ICML'97)*, pages 296–304, 1997.

[174] Tom Rowlands, David Hawking, and Ramesh Sankaranarayana. New-web search with microblog annotations. In *Proceedings of the 19th international conference on World wide web*, pages 1293–1296. ACM, 2010.

[175] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[176] Gerard Salton. Automatic information organization and retrieval. 1968.

[177] Gerard Salton. A comparison between manual and automatic indexing methods. Technical report, Cornell University, 1968.

[178] Gerard Salton. The smart retrieval system—experiments in automatic document processing. 1971.

[179] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[180] Mark Sanderson. *Test collection based evaluation of information retrieval systems*. Now Publishers Inc, 2010.

[181] Mark Sanderson and Justin Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 162–169. ACM, 2005.

[182] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, volume 12, pages 44–49. Citeseer, 1994.

[183] Miguel-Angel Sicilia, Nikolaos Th Korfiatis, Marios Poulos, and George Bokos. Evaluating authoritative sources using social networks: an insight from wikipedia. *Online Information Review*, 30(3):252–262, 2006.

[184] Amit Singhal and Fernando Pereira. Document expansion for speech retrieval. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 34–41. ACM, 1999.

[185] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 21–29. ACM, 1996.

[186] Mark D Smucker, James Allan, and Ben Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 623–632. ACM, 2007.

[187] Xiaodan Song, Yun Chi, Koji Hino, and Belle Tseng. Identifying opinion leaders in the blogosphere. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 971–974. ACM, 2007.

[188] Laure Soulier. *Définition et évaluation de modèles de recherche d'information collaborative basés sur les compétences de domaine et les rôles des utilisateurs*. PhD thesis, Université de Toulouse, 2014.

[189] Danny Sullivan. What social signals do google & bing really count? *Search Engine Land*, 1, 2010.

[190] James Surowiecki. *The wisdom of crowds*. Anchor, 2005.

[191] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62 (2):406–418, 2011.

[192] Unknown. The best place to hide a dead body is page 2 of google search results. 2014. URL http://www.insivia.com/quoter/the-best-place-to-hidea-dead-body-is-page-2-of-google-search-results/.

[193] David Vallet, Iván Cantador, and Joemon M Jose. Personalizing web search with folksonomy-based user and document profiles. In *Advances in Information Retrieval*, pages 420–431. Springer, 2010.

[194] Jean-Philippe Vert, Koji Tsuda, and Bernhard Schölkopf. A primer on kernel methods. *Kernel Methods in Computational Biology*, pages 35–70, 2004.

[195] Sarah Vieweg, Amanda L Hughes, Kate Starbird, and Leysia Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1079–1088. ACM, 2010.

[196] Yana Volkovich and Andreas Kaltenbrunner. Evaluation of valuable user generated content on social news web sites. In *Proceedings of the 20th international conference companion on World wide web*, pages 139–140. ACM, 2011.

[197] Jan Vosecky, Kenneth Wai-Ting Leung, and Wilfred Ng. Searching for quality microblog posts: Filtering and ranking based on content analysis and implicit links. In *Database Systems for Advanced Applications*, pages 397–413. Springer, 2012.

[198] Qihua Wang and Hongxia Jin. Exploring online social activities for adaptive search personalization. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 999–1008. ACM, 2010.

[199] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[200] Bingjie Wei, Shuai Zhang, Rui Li, and Bin Wang. A time-aware language model for microblog retrieval. Technical report, DTIC Document, 2012.

[201] Thijs Westerveld, Wessel Kraaij, and Djoerd Hiemstra. Retrieving web pages using content, links, urls and anchors. 2002.

[202] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, pages 80–83, 1945.

[203] Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[204] Fang Wu and Bernardo A Huberman. Social structure and opinion formation. *arXiv preprint cond-mat/0407252*, 2004.

[205] Shengliang Xu, Shenghua Bao, Ben Fei, Zhong Su, and Yong Yu. Exploring folksonomy for personalized search. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 155–162. ACM, 2008.

[206] Gui-Rong Xue, Hua-Jun Zeng, Zheng Chen, Yong Yu, Wei-Ying Ma, WenSi Xi, and WeiGuo Fan. Optimizing web search using web click-through data. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 118–126. ACM, 2004.

[207] Sihem Amer Yahia, Michael Benedikt, and Philip Bohannon. Challenges in searching online communities. In *IEEE Data Eng. Bull*. Citeseer, 2007.

[208] Sihem Amer Yahia, Michael Benedikt, Laks VS Lakshmanan, and Julia Stoyanovich. Efficient network aware search in collaborative tagging sites. *Proceedings of the VLDB Endowment*, 1(1):710–721, 2008.

[209] Yuto Yamaguchi, Tsubasa Takahashi, Toshiyuki Amagasa, and Hiroyuki Kitagawa. Turank: Twitter user ranking based on user-tweet graph analysis. In *Web Information Systems Engineering–WISE 2010*, pages 240–253. Springer, 2010.

[210] Yusuke Yanbe, Adam Jatowt, Satoshi Nakamura, and Katsumi Tanaka. Towards improving web search by utilizing social bookmarks. In *Web Engineering*, pages 343–357. Springer, 2007.

[211] Xiao Yang and Zhaoxin Zhang. Combining prestige and relevance ranking for personalized recommendation. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, pages 1877–1880, 2013. doi: 10.1145/2505515.2507885. URL http://doi.acm.org/10.1145/2505515.2507885.

[212] Show-Jane Yen and Yue-Shi Lee. Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. In De-Shuang Huang, Kang Li, and GeorgeWilliam Irwin, editors, *Intelligent Control and Automation*, volume 344 of *Lecture Notes in Control and Information Sciences*, pages 731–740. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-37255-4. doi: 10.1007/978-3-540-37256-1_89. URL http://dx.doi.org/10.1007/978-3-540-37256-1_89.

[213] Lei Yu and Huan Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *ICML*, volume 3, pages 856–863, 2003.

[214] Quan Yuan, Gao Cong, and Nadia Magnenat Thalmann. Enhancing naive bayes with various smoothing methods for short text classification. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 645–646. ACM, 2012.

[215] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, April 2004. ISSN 1046-8188. doi: 10.1145/984321.984322. URL http://doi.acm.org/10.1145/984321.984322.

[216] Xiaoxun Zhang, Lichun Yang, Xian Wu, Honglei Guo, Zhili Guo, Shenghua Bao, Yong Yu, and Zhong Su. sdoc: exploring social wisdom for document enhancement in web mining. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 395–404. ACM, 2009.

[217] Dejin Zhao and Mary Beth Rosson. How and why people twitter: the role that micro-blogging plays in informal communication at work. In *Proceedings of the ACM 2009 international conference on Supporting group work*, pages 243–252. ACM, 2009.

[218] Lulin Zhao, Yi Zeng, and Ning Zhong. A weighted multi-factor algorithm for microblog search. In *Active Media Technology*, pages 153–161. Springer, 2011.

[219] Ding Zhou, Jiang Bian, Shuyi Zheng, Hongyuan Zha, and C Lee Giles. Exploring social annotations for information retrieval. In *Proceedings of the 17th international conference on World Wide Web*, pages 715–724. ACM, 2008.

[220] Paul Zikopoulos, Chris Eaton, et al. *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media, 2011.

[221] Justin Zobel, Alistair Moffat, and Kotagiri Ramamohanarao. Inverted files versus signature files for text indexing. *ACM Transactions on Database Systems (TODS)*, 23 (4):453–490, 1998.