Étude des facteurs de pertinence dans la recherche de microblogs.

Firas Damak

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 *15/07/2014* **par :**

Firas Damak

Étude des facteurs de pertinence dans la recherche de microblogs.

JURY

Claude CHRISMENT	Professeur, Université Toulouse 3	Président du Jury
Patrice BELLOT	Professeur, Aix-Marseille Université	Rapporteur
Patrick GALLINARI	Professeur, Université Pierre et Marie Curie	Rapporteur
Brigitte GRAU	Professeur, ENSIIE	Examinatrice
Christian SALLABERRY	MCF/HDR, UPPA	Examinateur
Mohand BOUGHANEM	Professeur, Université Toulouse 3	Directeur
Guillaume CABANAC	MCF, Université Toulouse 3	Co-encadrant
Karen PINEL-SAUVAGNAT	MCF, Université Toulouse 3	Co-encadrante

---

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, *Guillaume CABANAC* et *Karen PINEL-SAUVAGNAT*
Rapporteurs :
Patrice BELLOT et *Patrick GALLINARI*

Du plus profond de mon cœur, je dédie ce travail,
À Mes parents Ridha et Sabeh Pour lesquels j'exprime mon amour et ma gratitude pour leur sacrifice et leur soutien moral. Ils n'ont eu de cesse de m'encourager et de m'offrir des conditions favorables durant la période de mes études. Que DIEU leur préserve une bonne vie.
À Mes frères Farah et Mehdi Qu'ils trouvent dans ce travail l'expression de ma reconnaissance en leur souhaitant un avenir plein de succès et de bonheur.
À Ma meilleure amie Ines Pour son soutien moral et pour les moments inoubliables que nous avons passés ensemble tout au long de ces années.
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. . .

Remerciements

Il m'est agréable de manifester ici toute ma gratitude à tous ceux et à toutes celles qui m'ont aidé de près ou de loin, afin d'aboutir au couronnement de quatre années de travail et de labeur. Toutefois je ne peux me permettre d'omettre de citer les honorables personnes auxquelles j'adresse ma modeste reconnaissance. Il s'agit de Monsieur Mohand Boughanem, Mme Karen Pinel-Sauvagnat et Monsieur Guillaume Cabanac, qui m'ont été d'un grand apport durant les moments les plus difficiles et ont atténué le poids du dépaysement. Elles m'ont permis de surpasser des périodes difficiles et ont fait renaître en moi la volonté d 'aller de l'avant et de continuer mon chemin vers la réussite.

Je remercie chaleureusement Monsieur Claude Chrisment, Monsieur Christian Salaberry, Madame Brigitte Grau, Monsieur Patrick Gallinari et Monsieur Patrice Bellot d'avoir accepté de juger ce travail.

Mes vifs remerciements s'adressent également à tous mes amis de l'IRIT qui m'ont prêté main forte pour la réalisation du projet : Mădălina, Dana, Ali, Mohamed, Rafik, Bilel, Lamjed, Arlind, Laure Eya, Ismail. Je tiens à remercier mes amis quotidiens de Toulouse : Faeiz, Bou7a, Marwa, Amine, Yessine, Faty, Sameh, Khouloud, Sami et Cycy.

Enfin, je souhaite remercier toute ma famille et mes amis en Tunisie.

Résumé

Notre travail se situe dans le contexte de recherche d'information (RI) sociale et s'intéresse plus particulièrement à la recherche de microblogs. Les microblogs sont des messages de faible longueur à travers lesquels les utilisateurs publient des informations sur différents sujets : des opinions, des événements, des statuts... Les microblogs occupent aujourd'hui une part considérable de l'information générée sur le web. Dans Twitter, la plate-forme de microblogging la plus populaire, le nombre de microblogs par jour peut atteindre 500 millions. Les microblogs ont une forme différente des traditionnels documents. Leur taille est réduite par rapport aux blogs et aux articles publiés sur le web (140 caractères pour Twitter). De plus, les microblogs peuvent contenir une syntaxe spécifique telle que les #hashtags, les @citations ou bien encore des URLs. Les plateformes de microblogging représentent également un modèle de réseau social différent des autres réseaux sociaux. Les relations entre les utilisateurs ne sont pas forcément réciproques et les abonnements sont sans restrictions entre microbloggeurs.

Les utilisateurs de plateformes de microblogging, outre la publication de microblogs, effectuent également des recherches. Les motivations de ces recherches sont diverses. Certaines sont similaires à la recherche sur le web (comme par exemple la recherche d'actualités), et d'autres sont spécifiques à la recherche de microblogs (comme par exemple la recherche temps réel ou d'informations sociales). Dans Twitter, 1,6 milliards de requêtes sont ainsi émises chaque jour.

Les modèles de RI doivent s'adapter aux spécificités des microblogs : fraîcheur, aspect social et spécificités syntaxiques doivent ainsi être pris en compte. C'est dans ce contexte de recherche d'information dans les microblogs que se situent plus particulièrement nos travaux. Nous nous plaçons plus précisément dans le cadre de la recherche adhoc. L'objectif est de retrouver les microblogs répondant à un besoin d'information spécifié par un utilisateur.

Nos travaux visent à améliorer la qualité des résultats de recherche d'information adhoc dans les microblogs. Nos contributions se situent à plusieurs niveaux :

-Afin de déterminer exactement les facteurs limitant les performances des modèles de recherche classiques dans un corpus de microblogs, nous avons mené à une analyse de défaillance d'un modèle de recherche usuel. Nous avons sélectionné les

microblogs pertinents mais non retrouvés par le modèle de recherche. Ensuite, nous avons identifié les facteurs empêchant leur restitution. Nous avons trouvé que le problème principal vient de la concision des microblogs. Cette concision engendre une correspondance limitée entre les termes des microblogs et les termes des requêtes, même s'ils sont sémantiquement similaires.

-Afin de compenser l'impact de la concision des microblogs, nous avons proposé et testé plusieurs solutions. Nous avons proposé d'étendre les requêtes (i) en exploitant des ressources de type actualités, (ii) en utilisant la base lexicale Wordnet, (iii) en applicant des techniques de réinjection de pertinence de l'état de l'art qui ont souvent prouvé leur efficacité : Rocchio pour identifier les termes susceptibles de ramener la pertinence ainsi que pour la pondération des termes de la nouvelle requête, et le mécanisme naturel d'extension de requêtes du modèle BM25. Dans Rocchio, nous avons testé différentes méthodes de calcul de poids de termes d'expansion. Nous avons enfin étendu les microblogs grâce aux liens (URLs) qu'ils contiennent. Nos expérimentations ont montré que l'emploi des URLs et l'expansion de requêtes sont primordiales pour la RI dans les microblogs. La plupart de ces expérimentations (expansion de requêtes et de microblogs) ont été réalisées en se basant sur le modèle vectoriel et sur le modèle probabiliste comme modèle de restitution. Ceci nous a permis de comparer les comportements des deux modèles sur les microblogs et avec les deux types d'expansion. De manière générale, nous avons trouvé que le modèle vectoriel est plus performant que modèle probabiliste au niveau de la sélection des microblogs pertinents (meilleur rappel). Cependant, le modèle probabiliste met plus en valeur les microblogs pertinents restitués par rapport à tous les microblogs restitués (meilleure précision).

-Un deuxième volet de notre travail concerne l'étude des critères utilisés pour identifier les microblogs pertinents. Nous avons repris les critères souvent utilisés dans l'état de l'art (critères de contenu, critères sur l'importance des auteurs, critères sur les URLs) et nous les avons évalués. Nous avons réalisé cette analyse selon 3 axes. Dans le premier axe, nous avons analysé l'impact de la combinaison des scores des critères avec le score de pertinence du contenu, calculé avec un modèle de RI usuel. Dans le deuxième axe, nous avons étudié le comportement des critères dans les documents pertinents et les avons comparés avec leurs comportements dans les documents non pertinents. Dans le troisième axe, nous avons utilisé des techniques d'apprentissage ainsi que des algorithmes de sélection de critères qui peuvent être utiles en entrée de ces techniques d'apprentissages. De manière générale, nous avons montré que les critères en relation avec les URLs publiées dans les tweets sont les plus discriminants. Les critères liés aux auteurs ne reflètent pas la pertinence.

-Afin de prendre en compte l'aspect temporel dans la restitution des microblogs pertinents vis-à-vis d'un besoin d'information, nous avons proposé trois méthodes qui intègrent le temps dans le calcul de la pertinence. Cette intégration du temps

n'a cependant pas montré son intérêt dans nos méthodes.

Pour réaliser nos expérimentations, nous nous sommes basés sur le corpus fourni par la campagne d'évaluation internationale TREC (Text Retrieval Conference) dans la tâche Microblogs des années 2011 et 2012. Nos différentes contributions ont également fait l'objet de participations aux trois tâches de Microblogs de TREC (2011, 2012 et 2013).

Abstract

This work deals with the context of social information retrieval (IR), more particularly the retrieval of microblogs. Microblogs are messages of short length. They contain information on various topics :opinions, events, articles... Microblogs represent a significant part of the information generated on the Web. In the case of Twitter, the most popular platform, the number of microblogs can reach 500 million per day. Microblogs have a different form from traditional documents. Their length is reduced compared to traditional blogs and articles on the web (only 140 characters in the case of Twitter). Moreover, microblogs can have specific syntax such as #hashtags, @mentions or shortened URLs... Microblogging platforms are a social network model different from other social networks. Relationships between users are not necessarily reciprocal and subscriptions are unrestricted between microbloggers. Users of microblogging platforms do not only produce but they also search for information. The motivations of this research are diverse. Some are inspired from Web search (e.g. the search for news) and others are specific to the search for microblogs (e.g. real-time search or social information). In Twitter, 1.6 billion queries are issued every day. Though, the IR models must adapt to the specificities of microblogs : freshness, social aspect and syntactic characteristics must therefore be taken into account. The aim of our work is to improve the quality of the results of adhoc information retrieval in microblogs. Our contributions are at several levels :

- In order to accurately determine the factors limiting the performance of conventional models of search in a corpus of microblogs, we conducted an analysis of failure of a conventional model search. We selected relevant microblogs. However, they are not found by the search pattern. Then, we identified the factors preventing their return. We found that the main problem is the shortness of microblogs.

- To offset the impact of the shortness of microblogs, we proposed and tested several solutions : to extend the queries by (i) exploiting news articles, (ii) using the WordNet lexical database, (iii) applying techniques of relevance feedback of the state of art which often proved effective : Rocchio to identify terms likely to bring relevance and for weighting the terms of the new query, and the natural extension mechanism queries of the BM25 model. Using Rocchio, we tested different methods of calculating the weight of expansion terms. We finally extended microblogs thanks

to the links (URLs) they contain. Our experiments have shown that the use of URLs and the expansion of the query are crucial for IR in microblogs. Most of these experiments (expansion of queries and microblogs) were performed on the basis of the vector model and the probabilistic model, as a model of restitution. This allowed us to compare the behavior of the two models on microblogs and with the two types of expansion. In general, we found that the Vector Space Model is more efficient than the probabilistic one in the selection of relevant microblogs (better recall). However, the probabilistic model puts more value on relevant microblogs returned over all returned microblogs (better precision).

- A second part of our work is concerned with the study of the features used to identify relevant microblogs. We selected the features often used in the state of art (content features, features on the importance of authors, URLs features and quality features).Then, we evaluated them. We conducted this analysis in 3 axes. In the first axis, (i) we studied the behavior of the features in the relevant documents and compared them with their behavior in non-relevant documents. In the second axis, (ii) we analyzed the impact of the combination of the features scores with the content's score, calculated with a model of conventional IR. In the third axis, (iii) we used learning techniques as well as algorithms of feature selection that may be useful as input to the learning techniques. In general, we have shown that the features related to URLs posted in tweets are the most discriminating. The features related to the authors do not reflect the relevance.

- To take into account the temporal aspect when selecting relevant microblogs, we have proposed three methods that incorporate time in the calculation of relevance. However, this integration of time did not show any positive impact in our methods.

To perform our experiments, we used the corpus provided by TREC (Text Retrieval Conference) international survey in the task Microblogs for the years 2011 and 2012. Our various contributions have also been the subject of participations for the three tasks of Microblogs TREC (2011, 2012 and 2013).

Table des matières

Résumé **7**

Abstract **10**

Table des matières **13**

Table des figures **17**

Liste des tableaux **19**

Introduction **1**

1 Introduction 1
2 Contexte 1
3 Problématiques de la RI dans les microblogs 4
4 Présentation des contributions 6
5 Organisation du mémoire 7

1 RI Sociale **11**

1 Information sociale dans le web 12
1.1 Contenus générés par les utilisateurs (UGC) 12
1.2 Contenus générés par la pratique 14
2 RI : historique 14
2.1 Processus de RI 15
2.2 Modèles de RI 19
2.3 Évaluation 24
3 Utilisation des informations sociales en RI : RI sociale 28
3.1 Côté utilisateur 29
3.2 Côté documents 31
4 Conclusion 33

2 RI dans les microblogs **35**
1 Présentation et spécificités des plate-formes de microblogging : cas de Twitter . . . 36
1.1 Présentation générale de Twitter . . . 36
1.2 Spécificités des microblogs . . . 41
1.3 Spécificités des recherches dans les microblogs . . . 43
2 Accès à l'information dans les microblogs . . . 45
2.1 Recherche temps-réel de microblogs . . . 45
2.2 Recherche de microbloggeurs . . . 46
2.3 Détection d'opinions . . . 47
2.4 Classification thématique des microblogs . . . 48
2.5 Détection de tendances . . . 48
3 Recherche adhoc de microblogs . . . 49
3.1 Facteur de pertinence textuelle . . . 50
3.2 Facteur de pertinence social . . . 51
3.3 Facteur de pertinence temporelle . . . 52
3.4 Facteur de pertinence d'hypertextualité . . . 53
3.5 Autres facteurs de pertinence . . . 54
3.6 Bilan . . . 55
4 Évaluation de la RI dans les microblogs . . . 55
4.1 La tâche TREC Microblog . . . 55
4.2 Discussion sur les mesures d'évaluation . . . 57
5 Bilan et limites de l'état de l'art . . . 58

3 Analyse de défaillance des modèles de RI classique sur les microblogs **61**
1 Introduction . . . 61
2 Méthodologie . . . 61
3 Expérimentations . . . 62
3.1 Cadre expérimental . . . 62
3.2 Observations . . . 62
4 Synthèse . . . 67

4 Expansion de requêtes et de documents **71**
1 Introduction . . . 71
2 Expansion de requêtes . . . 71
2.1 Exploitation des articles d'actualités . . . 72
2.2 Exploitation de la base lexicale WordNet . . . 74
2.3 Suggestions orthographiques . . . 75
2.4 Réinjection de pertinence . . . 76

3 Expansion de microblogs 80
3.1 Expansion de hashtags dans les tweets 80
3.2 Emploi des URLs 81
4 Expansion de requêtes et de documents 82
5 Discussion 85
6 Bilan 87

5 Analyse des facteurs de pertinence 89
1 Introduction 89
2 Description des facteurs de pertinence 90
2.1 Facteurs de pertinence basés sur le contenu des tweets 90
2.2 Facteurs de pertinence basés sur l'hypertextualité 91
2.3 Facteurs de pertinence basés sur les hashtags 91
2.4 Facteurs de pertinence basés sur la popularité des auteurs 92
2.5 Facteurs de pertinence relatifs à la qualité des tweets 92
3 Méthodologie 93
3.1 Étude de la distribution des scores 93
3.2 Étude par la combinaison linéaire des scores 93
3.3 Étude avec les techniques de sélection d'attributs 94
4 Expérimentations 94
4.1 Étude par la distribution des scores 94
4.2 Étude par la combinaison linéaire des scores 98
4.3 Étude avec les techniques de sélection d'attributs 104
5 Conclusion 108

6 Prise en compte du temps dans la recherche de microblogs 111
1 Introduction 111
2 Emploi de la fraîcheur dans la restitution des microblogs 112
2.1 Favoriser des tweets récents 112
2.2 Favoriser les termes récents 113
2.3 Observations 114
3 Prise en compte de la fréquence temporelle 117
4 Analyse requête par requête 118
5 Conclusion 120

7 Conclusion générale 123
Références 126

Table des figures

1.1 Processus en U de la recherche d'information 16

1.2 Catégorisation des modèles de RI (Baeza-Yates et Ribeiro-Neto, 1999) 20

1.3 Exploitation de l'information sociale dans la RI 28

1.4 Résultats à partir du cercle social dans Google 31

1.5 Recommandation de profils expert sur le sujet recherché sur Bing . . 31

2.1 L'interface graphique utilisateur de Twitter 38

2.2 Informations des comptes utilisateurs sur Twitter 39

2.3 Exemple d'utilisation de Twitter (avec tweets, retweets, abonnements et hashtags) 40

2.4 Notification sur l'apparition de nouveaux résultats dans Twitter . . . 40

2.5 Tweet posté par @florencesantrot contenant une image et des hashtags (#Apple #iphone6cost1k). Il a été retweeté sept fois et favori une fois. 42

2.6 Suggestion de différents type de résultats dans le moteur de recherche de Twitter : des mots-clés, des hashtags, des comptes utilisateurs sont présentés. 44

2.7 Les réseaux constituables à partir des données de Twitter 52

2.8 Exemple de *topic* pour la tâche Microblog 57

3.1 Répartition des tweets pertinents restitués avec le modèle vectoriel par rapport à tous les tweets pertinents connus pour chaque requête de 2011 63

3.2 Répartition des tweets pertinents restitués avec le modèle vectoriel par rapport à tous les tweets pertinents connus pour chaque requête de 2012. 63

5.1 Distribution des scores des tweets pertinents et des tweets non pertinents (requêtes de 2011 à gauche et celles de 2012 à droite). 97

6.1 Distribution temporelle des tweets pertinents et non pertinents pour les requêtes de TREC Microblog 2012. Les rectangles représentent les tweets pertinents tandis que les losanges représentent les tweets non pertinents. 117

Liste des tableaux

2.1 Nombre de requêtes par jours (en milliard). Chiffres obtenus du site `http://statisticbrain.com` 37
2.2 Critères de pertinence 56

3.1 Récapitulatif des différents facteurs limitant l'efficacité du modèle de recherche sur les microblogs 68

4.1 Emploi des articles de type actualité pour l'expansion de requêtes (avec et sans pondération des termes d'expansion, 1500 résultats par requête). Un astérisque indique une amélioration significative par rapport à la baseline. 73
4.2 Récapitulatif des différents runs testés sans pondération des termes ajoutés aux requêtes. 74
4.3 Test de l'amélioration des performance via la correction orthographique des requêtes. 75
4.4 Expansion de la requête initiale avec Rocchio. Les poids des termes d'expansion sont calculés avec TF.IDF. Un astérisque indique une amélioration significative par rapport à la baseline. 77
4.5 Expansion de la requête initiale avec Rocchio. Les poids des termes d'expansion sont calculés avec BM25. Un astérisque indique une amélioration significative par rapport à la baseline. 77
4.6 Différentes configurations du modèle BM25. * montre une amélioration significative par rapport à configuration de base (run BM25). . . 78
4.7 Résultats après l'expansion de hashtags, avec le modèle vectoriel et le modèle BM25 (sans et avec paramétrage). 81
4.8 Apport de l'emploi des URLs avec le modèle vectoriel et le modèle BM25. * montre une amélioration significative par rapport au run précédent. 81
4.9 Emploi des tweets et des URLs et expansion de requêtes uniquement à partir des tweets.* montre une amélioration significative par rapport au run précédent. 83

4.10 Emploi des tweets et des URLs pour l'expansion et pour la restitution. * montre une amélioration significative par rapport au run précédent. 84
4.11 Résultats des meilleurs runs avec les tweets hautement pertinents. . . 86
4.12 Comparaison avec les résultats officiels de TREC 2012 86
4.13 Emploi des tweets pour l'expansion et des tweets et des URLs pour la restitution sur les topics de TREC 2011. 86
4.14 Comparaison avec les résultats officiels de TREC 2011 87

5.1 Apport de chaque facteur de pertinence par rapport au modèle vectoriel (baseline VSM). 99
5.2 Apport de chaque groupe de facteurs de pertinence et de leurs combinaisons par rapport modèle vectoriel (baseline VSM). 100
5.3 Comparaison avec les résultats officiels de TREC 2011 101
5.4 Apport des facteurs de pertinence pour le cas général. 102
5.5 Apport des groupes de facteur de pertinence et de leurs combinaisons pour le cas général. 103
5.6 Critères sélectionnés avec les techniques de sélection d'attributs . . . 106
5.7 Résultats (P@30), les scores en gras indiquent des améliorations significatives par rapport à la baseline 107

6.1 Amplification des scores de pertinence de contenu en fonction de leur fraîcheur 113
6.2 Amplification des scores des termes en fonction de leur fréquence d'apparition dans le temps 114
6.3 Prise en compte de la fréquence temporelle. 118
6.4 Requêtes améliorées sur la mesure MAP pour les 3 méthodes 119

Introduction

1 Introduction

Le web, créé au début des années 1990 et initialement composé de pages statiques reliées entre elles par des hyperliens, s'est rapidement orienté vers un cadre beaucoup plus collaboratif, dans lequel tous les internautes consultent, créent, partagent et diffusent de l'information.

Ce changement est dû à la mise à disposition des internautes de plusieurs outils collaboratifs : les blogs, les wiki (Wikipedia en 2001) et les plate-formes sociales (Facebook en 2004 et Twitter en 2006), où les internautes ne se limitent plus à la consommation, mais **contribuent également à la production des contenus**. Ces outils prennent souvent la forme de réseaux sociaux qui se caractérisent par un ensemble d'entités, telles que des individus ou des organisations, qui sont reliées par des liens, d'amitié ou d'abonnement, permettant l'interaction sociale entre elles.

Des quantités de contenus, toujours plus volumineuses, sont de ce fait créées tous les jours. Ce nouveau contexte de diffusion de l'information peut constituer un moyen efficace pour cerner les besoins en information des utilisateurs du Web, et permettre à la Recherche d'Information (RI) de mieux répondre à ces besoins. Les Systèmes de Recherche d'Information (SRI) doivent ainsi s'adapter aux nouvelles exigences et nécessités des utilisateurs, et aux **spécificités** de ces nouvelles sources d'informations.

La prise en compte de ces informations sociales dans la restitution d'informations a engendré un nouveau paradigme de recherche : la RI sociale. Elle consiste à adapter les modèles et les algorithmes de la RI classique en exploitant les informations sociales développées avec l'arrivée du web 2.0.

2 Contexte

Notre travail se situe dans le contexte de la recherche d'information sociale et s'intéresse plus particulièrement à la recherche de microblogs. Les microblogs sont des messages de faible longueur à travers lesquels les utilisateurs publient des informations sur différents sujets : des opinions, des événements, des statuts... Les micro-

bloggeurs (les internautes qui publient des microblogs) utilisent des plate-formes de microblogging pour cette pratique. Parmi les plate-formes de microblogging, Twitter [1] est sans conteste la plate-forme la plus utilisée. Ces plate-formes sont de plus en plus exploitées (Kwak et al., 2010), aussi bien par des individus à titre personnel que dans des organisations, qui génèrent à travers les messages qu'ils écrivent et les liens qu'ils mettent en place des quantités importantes d'information. Nous pouvons nous référer, pour montrer l'importance de la quantité d'information publiée sur ces plate-formes, au 3 août 2013 lors d'une diffusion du dessin animé *Castle in the Sky* de Miyazaki [2] : 143 199 tweets ont été envoyés la même seconde. Cet événement avait cependant été préparé auparavant par les fans de Miyazaki. Ce jour-la, un record de 500 millions de tweets par jour est noté sur Twitter [3].

Les approches de RI classiques, élaborées pour traiter les documents traditionnels ou des documents de type page Web et qui se basent principalement sur le contenu textuel des documents et sur des statistiques des fréquences de termes, ne sont plus adaptées aux *spécificités* de cette nouvelle forme de contenu. Pour valoriser au mieux l'ensemble des informations de cette nouvelle source, les méthodes existantes de recherche d'information doivent être adaptées ou de nouvelles méthodes doivent être proposées. Ces nouvelles approches doivent tenir compte aussi bien des spécificités de ces informations que des motivations des internautes pour chercher dans ce type de ressources.

Considérons les *spécificités* des microblogs. Tout d'abord, leur taille est réduite par rapport aux blogs et aux articles publiés sur le web. Les tweets par exemple sont limités à 140 caractères ; ils sont souvent composés d'une seule phrase, écrite en mode SMS. Cette spécificité participe à la concrétisation du facteur temps-réel des microblogs. En fait, elle encourage les microbloggeurs non seulement à partager plus fréquemment, mais à signaler tout ce qui se déroule dans leur vie en temps-réel. En outre, les plate-formes de microblogging sont aujourd'hui accessibles à travers plusieurs types de dispositifs (tablettes, smartphones...). Un microbloggeur peut publier ainsi plusieurs microblogs chaque jour, contrairement à un bloggeur dont la fréquence de publication des articles est de plusieurs jours.

De plus, afin de faciliter le suivi des sujets discutés, les plateformes de microblogging utilisent une syntaxe spécifique telle que les #hashtags et les @citation. Elles permettent également aux utilisateurs d'insérer des URLs et des images dans les microblogs.

L'engouement pour les plate-formes de microblogging tient certainement aussi à l'aspect réseau social induit par les liens possibles. Cependant, les plate-formes de microblogging représentent un type de réseau social différent des autres réseaux

1. `https://twitter.com`
2. `http://www.imdb.com/name/nm0594503/`
3. `https://blog.twitter.com/2013/new-tweets-per-second-record-and-how`

sociaux. Les relations entre les utilisateurs ne sont pas forcément réciproques et les abonnements sont sans restriction entre microbloggeurs.

D'autre part, les *motivations* pour exploiter ces sources sont particulières. Les microbloggeurs, outre la publication de tweets, réalisent des recherches sur les plateformes de microblogging. Dans Twitter, 1,6 milliards de requêtes sont émises chaque jour [4]. La recherche sur les plateformes de microblogging est spécifique, et ce pour deux raisons. Selon Teevan et al. (2011) la plate-forme de Twitter est utilisée non seulement comme une source d'information parmi d'autres du web, mais également comme une source d'information *temps-réel* qui permet d'obtenir des actualités, de les commenter et de les partager à l'instant de leur déroulement (par exemple, `guerre à Gaza`, `bouchon sur l'autoroute A7`). Pierre Guillou, dirigeant de la société IDEOSE, spécialisée dans l'accessibilité et les nouveaux usages du Web définit le web temps-réel comme :

« l'ensemble des informations envoyées sur le Web par des personnes de façon instantanée et publique. Ces informations sont envoyées dans un même temps à un groupe de destinataires, publiées sur le Web et analysables par des logiciels de traitement de l'information. »

Plusieurs outils sont ainsi apparus pour extraire tout type d'information en temps réel à partir des microblogs. Par exemple, en analysant près de 50 millions de tweets chaque jour, l'hédonomètre [5] créé par des chercheurs américains permet de connaître en temps-réel l'état d'esprit et l'humeur d'une grande partie de la population dans le monde entier. Diakopoulos et Shamma (2010) ont proposé un analyseur temps-réel permettant de visionner dynamiquement les sentiments des téléspectateurs au cours d'un débat politique.

De plus, Teevan et al. (2011) ont montré que les utilisateurs cherchent des *informations sociales* dans ces plate-formes (26 % des utilisateurs). Ils l'utilisent pour plusieurs raisons telles que la recherche de personnes qui ont des intérêts similaires, ou de ce qu'un utilisateur est en train de dire... Les utilisateurs emploient ces plate-formes également pour suivre ce que les autres disent à propos du contenu d'un microblog ou un sujet en particulier. Ils utilisent ainsi les #hashtags et les @citation pour faire ces recherches verticales.

Pour conclure, le microblogging a été conçu de manière à faciliter l'accès et la publication des informations. Par conséquent, cette source gagne de plus en plus d'intérêt, que ce soit pour partager ou pour acquérir de l'information. Les informations partagées dans ces plate-formes sont ainsi utilisées pour obtenir des opinions des consommateurs (Jansen et al., 2009a ; O'Connor et al., 2010), des convictions politiques (Tumasjan et al., 2010) et des actualités (Okazaki et Matsuo, 2010 ; Sa-

4. `http://engineering.twitter.com/2011/05/engineering-behind-twitters-new-search.html`

5. `http://hedonometer.org/`

kaki et al., 2010 ; Sankaranarayanan et al., 2009 ; Phelan et al., 2009). Jansen et al. (2009a) qualifient ces moyens de communication aujourd'hui comme la « bouche du monde ».

Il est primordial pour les modèles de RI dans les microblogs de considérer les facteurs simplifiant l'accès et la publication des informations employés par les plate-formes de microblogging. Ce sont en particulier, la fraîcheur, l'aspect social, et les spécificités syntaxiques des microblogs. C'est dans le contexte de recherche d'information dans les microblogs que se situent plus particulièrement nos travaux. Nous nous plaçons plus précisément dans le cadre de la recherche adhoc. L'objectif est de retrouver les microblogs répondant à un besoin d'information spécifié par un utilisateur.

3 Problématiques de la RI dans les microblogs

Comme nous l'avons vu précédemment, un moteur de recherche de microblogs doit prendre en compte leurs spécificités ainsi que de nouvelles exigences des utilisateurs en termes de fraîcheur, de nouveauté d'information, et d'importance dans le réseau social, par exemple.

D'un point de vue recherche d'information, si on projette les différentes spécificités des microblogs dans une tâche de recherche d'information, on pourra facilement identifier de nouvelles problématiques par rapport aux problématiques classiques de la RI, que ce soit au niveau de l'indexation ou bien au niveau de la restitution des informations, ou encore de l'évaluation des performances. Au niveau de l'indexation dans un cadre temps-réel, les microblogs arrivent avec **une fréquence très importante**, souvent par **rafales** correspondant à des événements, et doivent être indexés dès leur arrivée. Ce même index doit permettre également une lecture avec un accès rapide, afin de rendre disponible un microblog à l'instant de sa création et de satisfaire les besoins en informations des utilisateurs. Ceci s'oppose à la majorité des index ordinaires du web, qui sont souvent des index avec **des architectures statiques avec des taux de mise à jour réduits**. En outre, les moteurs de recherche usuels utilisent des robots qui se basent sur les liens hypertextes pour détecter les nouvelles pages, ce qui rend l'organisation des documents dans l'index dépendante des liens entre les pages. Cependant, pour une recherche temps-réel, les microblogs doivent être ordonnés en fonction de leur date de publication dans l'index afin de favoriser la fraîcheur des résultats au moment de la restitution. Ensuite, comme nous l'avons déjà motionné, les microbloggeurs tendent à écrire en mode SMS. Les messages peuvent contenir des termes mal-orthographiés, du jargon du net, beaucoup d'émoticônes (Bamman et al., 2012)... S'ajoute à ceci l'emploi de syntaxes spécifiques à certaines plate-formes, tels que les hashtags et les mentions. Tous ces

facteurs introduisent de nouvelles difficultés et demandent de faire des choix sur l'intérêt de les traiter au moment de l'indexation. À quel niveau le traitement est-il possible, sans ralentir l'indexation, et en respectant les conditions du temps-réel ?

Ces mêmes caractéristiques des microblogs posent également des problèmes au niveau de la recherche et de la restitution des données :

- **Quelle est l'unité d'information la plus appropriée pour répondre aux besoins en informations ?** Si un utilisateur recherche des informations concernant un sujet dans les plate-formes de microblogging, est-ce utile de restituer des microblogs, des hashtags, des synthèses de microblogs, des conversations, des profils, etc. ?
- **Quel est le modèle le plus approprié pour gérer les spécificités des microblogs et les exigences des utilisateurs ?** Les modèles de RI, qui de manière générale se basent sur des facteurs tels que la fréquence des termes dans les documents et la longueur des documents, demeurent limités par la faible longueur des microblogs où les termes n'apparaissent pas plus d'une fois. Ces facteurs ont un sens quand la taille du document est importante. Quelles sont alors les solutions pour compenser ce manque de contenu ? De plus, les modèles usuels se basent sur le vocabulaire du document pour mesurer sa pertinence vis-à-vis d'une requête, alors que la pertinence dans les microblogs demeure théoriquement couplée avec d'autres facteurs tels que la fraîcheur de l'information, la popularité de l'auteur de l'information, la qualité du langage utilisé, etc., en complément de la pertinence sur le contenu.
- **Quels sont les facteurs qui reflètent vraiment la pertinence dans une tâche de recherche de microblogs ? Quels sont les moyens permettant d'évaluer les facteurs ?** La plupart des approches de recherche de microblogs proposées dans l'état de l'art s'appuient sur différentes intuitions et définissent ainsi la pertinence comme la composition de plusieurs facteurs, en plus du facteur lié au contenu. Cependant, aucune évaluation individuelle de ces facteurs n'a été réalisée à ce jour.
- **Comment ces facteurs peuvent-ils être employés et combinés avec la pertinence du contenu ?** Efron (2011a) déclare que *les critères de pertinence reflètent certainement la pertinence. Cependant, il n'est pas toujours simple de déterminer comment les employer*. La popularité des auteurs, par exemple, peut être considérée de différentes manières : l'activité de l'auteur, le nombre de ses amis, sa centralité dans le réseau social, etc.
- **La fraîcheur, est-elle vraiment un facteur crucial de pertinence ?** Teevan et al. (2011) ont montré que l'une des motivations pour chercher les microblogs est la fraîcheur de l'information. Cependant, elle n'est pas l'unique motivation. Plusieurs recherches sur les microblogs visent des informations sociales ou des informations d'ordre général (des opinions de consommateurs par

exemple). Ainsi, est-il utile d'intégrer la fraîcheur comme facteur de pertinence quel que soit le besoin d'information ?

En recherche d'information, la troisième étape fondamentale, après l'indexation et la recherche, est l'évaluation. Cette phase permet de mesurer l'efficacité des approches et des choix faits durant les deux étapes précédentes. Depuis des décennies, le paradigme de Cranfield, qui établit l'évaluation des SRI à travers des corpus *statiques*, a dominé sur les expérimentations de la RI moderne. **Cette méthode ne pose-t-elle pas un problème lorsqu'elle est appliquée dans une tâche pour laquelle le facteur temps-réel est primordial ?**

4 Présentation des contributions

Nos travaux visent à améliorer la qualité des résultats de recherche d'information adhoc dans les microblogs et nous nous focalisons donc sur les problématiques liées à la recherche. La tâche adhoc consiste en la restitution de microblogs pertinents vis-à-vis d'un besoin en information exprimé sous forme de mots-clés formant la requête. Nos contributions se situent à plusieurs niveaux :

1. Afin de déterminer exactement les facteurs limitant les performances des modèles classiques de recherche dans un corpus de microblogs, nous avons conduit **une analyse de défaillance** d'un modèle de recherche usuel. Nous avons sélectionné les microblogs pertinents mais non retrouvés par le modèle de recherche. Ensuite, nous avons identifié les facteurs empêchant leur restitution. À l'issue de cette analyse, nous avons proposé et testé plusieurs solutions permettant d'améliorer la qualité des moteurs de recherche.
2. Afin de compenser l'impact de la concision des microblogs, nous avons introduit et testé plusieurs propositions. La première consiste à **exploiter des ressources de type actualités pour étendre les requêtes.** Ensuite, nous nous sommes basés sur la base lexicale WordNet souvent utilisée en RI comme un moyen de désambiguïsation et d'extension de requêtes. Nous avons également analysé l'impact des approches connues en RI sur ce type de ressources. Nous avons appliqué des techniques de réinjection de pertinence de l'état de l'art, telles que Rocchio (1971), pour identifier les termes susceptibles de favoriser la restitution de microblogs pertinents, ainsi que la pondération des termes de la nouvelle requête et le mécanisme naturel d'extension de requête du modèle BM25. Nous avons testé différentes méthodes pour calculer les poids des termes. D'autre part, nous avons exploité **les liens publiés dans les microblogs** pour étendre les microblogs. Ces hyperliens représentent de l'information additionnelle qui complète les contenus des microblogs. Ainsi, nous les avons considérés pour enrichir la représentation du contenu textuel

des microblogs.

3. Un troisième volet de notre travail concerne **l'étude des facteurs de pertinence utilisés pour identifier les microblogs pertinents**. Nous avons repris les facteurs de pertinence souvent utilisés dans l'état de l'art (de contenu, sur l'importance des auteurs, sur les URLs. . .) et nous les avons évalués. Cette analyse est conduite selon trois axes. Dans le premier axe, nous avons étudié le comportement des facteurs de pertinence dans les microblogs pertinents et les avons comparés avec leur comportement dans les documents non pertinents. Dans le deuxième axe, nous avons analysé l'impact de la combinaison des scores des facteurs de pertinence avec le score de pertinence du contenu, calculé avec un modèle classique de RI. Dans le troisième axe, nous avons utilisé des techniques d'apprentissage ainsi que des algorithmes de sélection d'attributs pour identifier les facteurs de pertinence utiles, en entrée des techniques d'apprentissage.
4. Afin de prendre en compte l'aspect temporel dans la restitution des microblogs pertinents vis-à-vis d'un besoin en information, nous avons proposé trois méthodes qui **intègrent le facteur temporel des microblogs dans le calcul de la pertinence**. Chaque méthode prend en compte le temps à sa manière.
   - La première favorise les documents récents en appliquant la technique Kernel (Lv et Zhai, 2009) pour mesurer la distribution temporelle des documents.
   - La deuxième privilégie les termes présents fréquemment au moment de la soumission de la requête.
   - La troisième favorise les termes qui apparaissent fréquemment au moment de la publication du microblog.

Afin d'évaluer l'apport de nos différentes contributions, nous nous sommes basés sur le corpus fourni par la campagne d'évaluation TREC (Text Retrieval Conference) pour la tâche Microblog en 2011 et 2012.

5 Organisation du mémoire

Ce mémoire est constitué de deux parties : la première présente le contexte général dans lequel se situe notre travail, à savoir la recherche d'information sociale et plus précisément la recherche d'information dans les microblogs. La seconde partie détaille notre contribution.

L'objectif de la première partie « **De la recherche d'informations classique à la recherche d'information sociale** » est de présenter les principes de la recherche d'information dans des contenus textuels, puis son application à l'environnement social. Cette partie comprend deux chapitres.

Le chapitre 1 présente les nouveaux contenus sociaux, développés avec l'apparition des technologies du Web 2.0. Ensuite, nous exposons les fondamentaux de la RI classique pour arriver aux spécificités de la RI sociale. Les différents types d'information sociale dans le web sont ainsi décrits, à savoir *les contenus générés par les utilisateurs* et *les contenus générés par les pratiques sociales*. Nous abordons ensuite les notions et les concepts de base de la RI classique. L'architecture générale d'un SRI y est présentée ainsi que les principaux modèles de recherche. Nous décrivons par la suite l'impact de l'emploi de l'information sociale sur la recherche d'information, en particulier dans la contextualisation des recherches ou bien dans l'enrichissement des ressources documentaires.

Nous nous concentrons dans le chapitre 2 sur une source d'information particulière : les microblogs. Nous présentons dans ce chapitre la recherche d'information dans les microblogs. Nous commençons ainsi par la description des spécificités de ce type de contenu et nous nous basons sur la plate-forme Twitter pour montrer les différences avec les documents traditionnels du web . Nous détaillons les spécificités du contenu des microblogs ainsi que les motivations des utilisateurs à chercher dans cette source d'information. Nous listons ensuite les approches d'accès à l'information à partir des microblogs proposées dans la littérature, en en particulier la recherche d'information adhoc dans les microblogs (notre domaine de recherche).

La seconde partie du mémoire intitulé « **étude des facteurs de pertinence pour la RI dans les microblogs** » expose nos contributions.

Le chapitre 3 décrit notre contribution à l'identification des facteurs limitant l'efficacité des modèles de RI classique dans un corpus de microblogs. Nous présentons une analyse de défaillance réalisée sur les résultats d'un modèle de recherche classique, dans une tâche de recherche de microblogs. Ce chapitre donne ainsi des pistes sur les considérations à prendre en compte pour améliorer la qualité des résultats.

Le chapitre 4 présente des solutions pour certains problèmes soulevés dans le chapitre 3. Plusieurs méthodes d'expansion de requêtes sont proposées et employées. Ces méthodes exploitent des ressources de différents type pour étendre les requêtes : les articles de type actualité, la base lexicale WordNet et un outil de suggestion d'orthographe. Le *feedback* est également utilisé à travers l'emploi des méthodes connues d'expansion de requêtes : Rocchio et le modèle de recherche BM25. D'autre part, des méthodes d'expansion de microblogs sont employées, à savoir l'expansion de hashtags et l'emploi des contenus pointés par les URLs pour améliorer la représentation des microblogs.

Le chapitre 5 présente une étude approfondie sur l'apport des facteurs de pertinence souvent utilisés dans les approches de l'état de l'art. Cette étude est réalisée en trois étapes : (i) la première étape consiste en la comparaison des distributions des scores des facteurs de pertinence entre les résultats pertinents et les résultats non pertinents. Les facteurs de pertinence ayant des comportements différents reflètent ainsi la pertinence. La deuxième étape est réalisée par la combinaison linéaire des scores des facteurs de pertinence. Les facteurs de pertinence améliorant la qualité des résultats reflètent ainsi la pertinence. Finalement, la troisième étape emploie les techniques de sélection d'attributs. Ces techniques permettent d'identifier automatiquement les meilleures combinaisons de facteurs de pertinence pour obtenir les meilleurs résultats.

Le chapitre 5 présente une étude approfondie sur un critère de pertinence particulier : la fraîcheur du microblog. Trois approches qui emploient le temps dans la restitution de microblogs sont proposées.

L'ensemble des évaluations se basent sur le corpus de tweets fourni par la campagne d'évaluation TREC (Text Retrieval Conference) dans la tâche microblogs des années 2011 et 2012.

En conclusion, nous dressons le bilan de nos travaux reliés à la recherche d'informations dans les microblogs. Nous introduisons ensuite les limites et les perspectives de ces travaux à court et à long terme.

État de l'art

Chapitre 1

RI Sociale

Satisfaire un besoin d'information a été souvent couplé avec des pratiques sociales. Ce couplage peut être perçu à plusieurs niveaux. D'une part, avant la naissance des SRI, le chercheur d'information se basait sur ses liens sociaux pour satisfaire son besoin. Le premier réflexe consistait à interroger les personnes qu'il connaissait et qui avaient des intérêts similaires. Ceci pouvait être réalisé également en interrogeant les amis, les proches ou simplement des bibliothécaires.

D'autre part, l'information est souvent produite dans des situations sociales, à travers des discussions et des collaborations entre les différents membres de groupes de personnes, partageant les mêmes objectifs et les mêmes centres d'intérêts.

L'arrivée de l'internet et en particulier les technologies du web 2.0 a complètement révolutionné ces pratiques. L'internaute aujourd'hui consulte les plus grandes bibliothèques et ressources scientifiques (Wikipédia[1]), utilise les moteurs de recherche pour trouver instantanément les informations (Google, Bing), discute avec d'autres utilisateurs ayant les mêmes centres d'intérêts (forum et blog), développe ses connaissances et relations sociales (réseaux sociaux), commente et consulte les avis des autres internautes (social bookmarking),... Les utilisateurs, en utilisant les technologies du web 2.0, génèrent directement ainsi de nouveaux contenus appelés **contenus générés par les utilisateurs (UGC)**. D'autres types d'information sont générés indirectement, comme par exemple, les liens sociaux, les profils ainsi que leurs traces de navigations. Ces données sont appelées **contenus générés par la pratique**.

L'exploitation et plus particulièrement l'accès à ces contenus, récemment générés, très spécifiques en terme de nature, de format, de structure et de volume, demande la définition de modèles de RI qui vont au-delà des modèles classiques définis dans le domaine de la RI depuis quelques années. En effet, les documents visés par les modèles de la RI classique se composent uniquement par leurs contenus textuels. Cependant, avec le web 2.0, plusieurs éléments, en plus du contenu textuel, doivent

1. `http://www.wikipedia.org/`

être considérés, tels que les informations sociales, les commentaires et les notes des internautes... Ces documents peuvent avoir un format spécifique (les microblogs par exemple font au maximum 140 caractères) et une syntaxe particulière. Par conséquent, pour chercher dans ces contenus, le modèle de recherche doit gérer toutes ces spécificités.

Dans ce chapitre, nous allons commencer par présenter les nouveaux contenus sociaux, développés avec l'apparition des technologies du web 2.0. Nous exposerons ensuite les principales bases de la RI classique pour arriver aux spécificités de la RI sociale.

1 Information sociale dans le web

L'information sociale dans le web est basée sur l'internet de plus en plus influencé par des services intelligents (présentés dans la suite), qui permettent à l'utilisateur de contribuer au développement, d'annoter et de collaborer à la production du contenu. Les utilisateurs sont passés de simples consommateurs à producteurs d'information. Leurs contributions peuvent être de différentes natures : les contenus publiés dans les plate-formes sociales telles que les blogs et les wikis, les réactions, les informations publiées par les autres utilisateurs telles que les annotations et les commentaires, etc. L'ensemble de ces informations est appelé **contenus générés par des utilisateurs** (UGC : User Generated Content).

1.1 Contenus générés par les utilisateurs (UGC)

Le terme « contenu généré par les utilisateurs » est devenu populaire en 2005 grâce au développement des moyens de production collaboratifs tels que les Wiki, les blogs, les forums, le social bookmarking, les plateformes de microblogging... Nous définissons en détail ci-après ces moyens de production :

– Wiki : un wiki est une application web permettant à ses utilisateurs de créer, modifier et supprimer des contenus de manière collaborative. L'information par conséquent est construite avec la participation de plusieurs personnes. Les wikis peuvent avoir plusieurs objectifs : outil de gestion de connaissances, outil de prise de notes, site communautaire, Intranet... Le premier wiki s'appelait Wikiwikiweb. Il a été développé par Ward Cunningham à Portland, Oregon, en 1994. L'application a été mise en ligne en 1995. Aujourd'hui, l'exemple le plus connu de wiki est Wikipedia [2], qui contient plus de 22 millions d'articles dans 278 langues différentes [3] . Un wiki se caractérise par l'encouragement à la

2. http://www.wikipedia.org/
3. http://en.wikipedia.org/wiki/WIKIPEDIA

création des liens hypertextes de sorte que chaque page soit reliée à plusieurs autres pages et chaque terme clé ou concept avec sa définition.

- Blog : Le blog est un type de site web sur lequel un internaute tient une chronique personnelle ou consacrée à un sujet particulier. Il s'agit d'un espace individuel d'expression, créé pour donner la parole à tous les internautes (particuliers, entreprises, artistes, hommes politiques, associations...), d'une part, et pour permettre à tous les visiteurs de réagir sur le sujet évoqué, en postant leurs commentaires sur les articles, créant ainsi une relation privilégiée entre l'auteur et ses lecteurs. Les plate-formes de blogs les plus connues sont Overblog [4], Blogger [5], SkyrockBlog [6] et CanalBlog [7].
- Forum : Un forum est un lieu d'échange d'informations où les internautes posent ou répondent à une question donnée. Les différentes contributions forment un fil de discussion (*thread* en anglais). Chaque forum de discussion se consacre à un thème précis. Par exemple, CFPOI World [8] se spécialise sur les animaux, alloforum [9] sur les voitures... Les messages publiés dans les forums sont archivés. Ceci permet aux internautes d'y participer d'une manière asynchrone. Contrairement aux blogs, les messages sont organisés chronologiquement, du plus ancien au plus récent.
- Social bookmarking : Le *social bookmarking* est un moyen pour stocker, classer, chercher et partager les liens favoris. Ces favoris seront ainsi accessibles à partir de n'importe quel point d'accès à l'internet, et non pas forcément à partir d'une machine personnelle. Ce principe simplifie ainsi leurs partages avec les autres internautes et permet de les récupérer même à partir de différentes machines. Un internaute a la possibilité de partager ses bookmarks, et également de regarder ce que les autres ont trouvé intéressant pour annoter. Selon Ebizmba [10], Delicious [11] est le site plus populaire de *social bookmarking*.
- Plate-forme de microblogging : Le microblogging dérive directement du concept des blogs. La différence réside principalement dans la longueur des publications. Les microbloggeurs sont souvent limités à un nombre de caractères qui est de l'ordre de 140 caractères (cas de Twitter). Toutefois, les microbloggeurs peuvent partager des images ou des liens externes dans leurs messages. Ce facteur encourage par conséquent les internautes à partager des microblogs plus fréquemment. Certaines plate-formes de microblogging se focalisent sur

4. http://www.over-blog.com/
5. http://www.blogger.com/
6. http://www.skyrock.com/blog/
7. http://www.canalblog.com/
8. http://www.animalforum.com/
9. http://ma850.alloforum.com/
10. http://www.ebizmba.com
11. http://delicious.com/

des thèmes spécifiques tels que Blipper [12] (livres, musiques, jeux, etc.) et Flixter [13] (films). Cependant, les sujet discutés dans Twitter, la plate-forme de micrblogging la plus populaire, ne sont pas contraints.

1.2 Contenus générés par la pratique

Ce deuxième type d'information sociale est produit au travers des différentes pratiques que les internautes réalisent tout au long de leurs sessions de navigations. On peut citer :

- Les traces des utilisateurs : elles comportent les différentes pages web visitées par les utilisateurs, les clics, les durées de visites... Ces données peuvent être utilisées afin de déterminer les préférences des utilisateurs et leurs thématiques de recherche.
- Les données personnelles : elles se composent des informations que l'utilisateur fournit au moment de son inscription sur les réseaux sociaux.
- Les liens sociaux : la plupart des plate-formes sociales définissent des règles de liaison entre leurs différents utilisateurs. Ces règles diffèrent d'une plate-forme à une autre. Prenons par exemple le cas de Twitter, il n'y a pas de restriction dans les liens sociaux. N'importe quel utilisateur peut s'abonner à un autre utilisateur, sans avoir forcément son accord (à moins que le compte soit privé, ce qui est rarement utilisé). Par opposition, dans le cas de Facebook, les deux utilisateurs doivent être d'accord pour partager leurs informations.

L'explosion des ressources sociales avec de nouvelles spécificités a permis l'émergence d'une nouvelle branche de la Recherche d'Information : la RI sociale. Il s'agit d'adapter les modèles et les algorithmes de la RI classique afin d'exploiter les informations sociales. Dans ce qui suit, avant de présenter les impacts de la prise en compte de l'information sociale dans la RI, nous présentons brièvement les concepts de base de la RI classique.

2 RI : historique

La recherche d'information (RI) traite de la représentation, du stockage, de l'organisation et de l'accès à l'information (Manning et al., 2008). L'objectif de la RI est de faciliter, pour un utilisateur, l'accès à l'information qui correspond à son besoin. Selon Baeza-Yates et Ribeiro-Neto (1999), un système de recherche d'information (SRI) doit fournir à l'utilisateur, d'une manière simple, l'information à laquelle il s'intéresse. Un SRI doit ainsi comprendre exactement la nature du besoin en information de l'utilisateur, sélectionner l'ensemble des documents qui traitent

12. http://blipper.com
13. http://flixter.com

de son besoin et finalement ordonner les éléments sélectionnés selon leur degré de pertinence décroissant.

2.1 Processus de RI

Satisfaire un besoin en information se traduit concrètement par la mise en correspondance d'un besoin d'information exprimé souvent sous forme de mots-clés, d'une part, et des informations disponibles dans les documents textuels d'une collection. Ce processus se déroule au sein d'un *système de recherche d'informations* (SRI). Un SRI est un ensemble logiciel assurant l'ensemble des fonctions nécessaires à la recherche d'information. Ces fonctions sont traduites via ce que l'on appelle le « processus en U » de la recherche d'information. La figure 1.1 (Belkin et Croft, 1992) en montre ces trois phases principales :

- l'indexation : crée un index à partir d'un corpus de documents. L'objectif de l'indexation est l'homogénéisation des représentations, tout en rendant l'accès rapide et efficace à l'ensemble des documents. Elle permet d'extraire les mots importants et caractéristiques d'un document.
- le requêtage : c'est l'étape durant laquelle l'utilisateur exprime son besoin d'information. Cette étape peut engendrer une reformulation de la requête initiale. La requête soumise par l'utilisateur subit les mêmes traitements que ceux réalisés sur les documents au cours de leur indexation.
- l'appariement : consiste à mesurer la similarité entre le besoin d'information et les descripteurs des documents dans l'index.

2.1.1 Indexation

Les documents à leur état brut sont difficiles à exploiter tels quels lors de la phase de recherche. Ainsi, l'objectif principal de cette étape est de fournir des représentations des documents et des requêtes facilement exploitables par la machine dans la phase de recherche. Cette représentation est souvent une liste pondérée de mots-clés significatifs que l'on nomme descripteurs du document (ou de la requête). L'indexation peut être manuelle, semi-automatique ou automatique.

- Indexation manuelle : c'est un spécialiste ou un documentaliste qui analyse le document et sélectionne par la suite les termes qu'il trouve représentatifs. L'indexation manuelle fournit une terminologie spécifique pour indexer et rechercher les documents, garantissant ainsi une meilleure représentation des documents et une meilleure qualité des résultats. Cependant, ce type d'indexation demande plus de temps et d'efforts que les autres. En outre, 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 (Furnas et al., 1987).

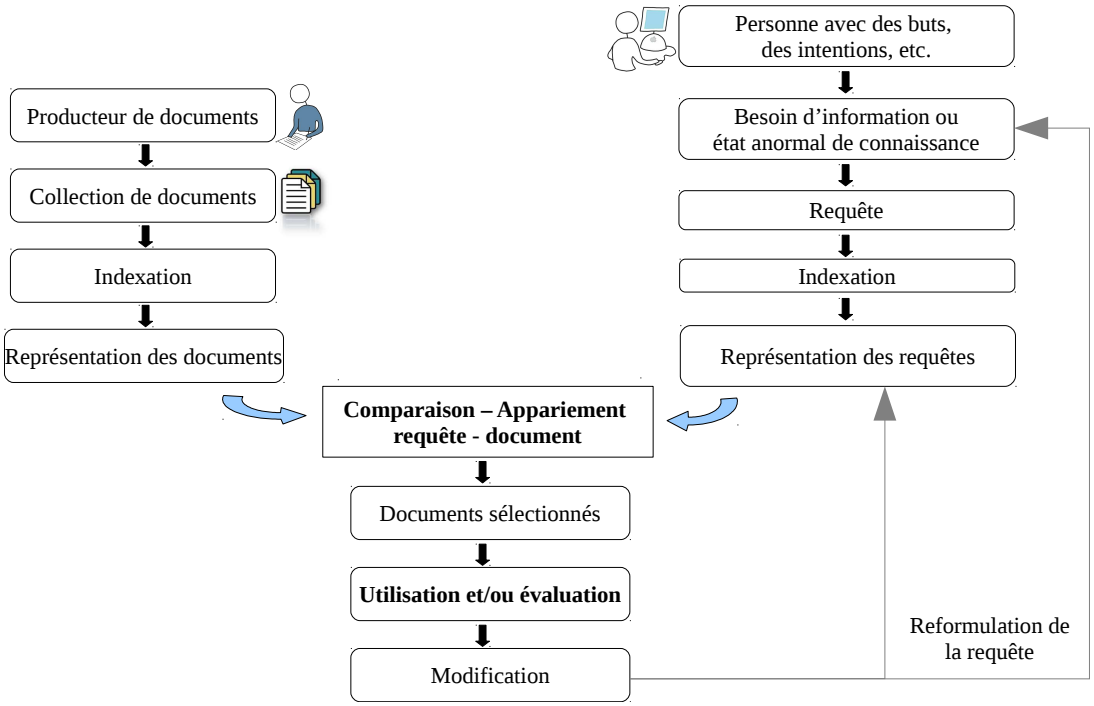


FIGURE 1.1 – Processus en U de la recherche d'information

- Indexation automatique : dans ce cas, c'est un ensemble de processus appelés robots d'indexation, qui réalisent de manière automatisée la tâche. C'est l'approche suivie par la majorité des SRI, en raison de sa rapidité et son coût réduit par rapport à l'indexation manuelle.
- Indexation semi-automatique : elle se base sur l'indexation automatique. Toutefois, une intervention humaine peut être réalisée afin d'effectuer des choix sur les termes significatifs, et pour valider la représentation finale des descripteurs. Ces choix sont souvent réalisés en utilisant un thésaurus ou une base terminologique qui est une liste organisée de descripteurs (mots-clés) liés à des règles terminologiques propres et reliés entre eux par des relations sémantiques.

D'une manière générale, l'indexation automatique comprend une chaîne de traitements automatisés. Ils sont appliqués sur les documents et également sur les requêtes. On distingue : l'extraction des mots, l'élimination des mots vides de sens, la normalisation et la pondération.

- **Extraction des mots :** cette étape consiste, dans un premier temps, à traiter chaque document afin de pouvoir extraire le texte comme une suite de caractères. Autrement dit, elle vise à résoudre les problèmes résultants des différents formats et encodages des documents, pour avoir en finalité uniquement le texte sous forme d'une séquence linéaire de caractères et de rejeter les éléments de forme. Ensuite, la séquence de caractères est découpée en une liste de termes

susceptibles d'être indexés par une analyse lexicale. Une analyse lexicale permet d'identifier les termes en reconnaissant les espaces de séparation des mots, des caractères spéciaux, des chiffres, les ponctuations, etc.

- **Élimination des mots vides :** les textes contiennent souvent des termes non significatifs appelés *mots vides* (pronoms personnels, prépositions...). Cette étape est réalisée par l'utilisation d'une liste de mots vides ou par le rejet de mots dépassant un certain nombre d'occurrences dans le document. L'élimination des termes vides a ses avantages et ses inconvénients. D'une part, pour certaines requêtes, la présence des termes vides joue un rôle très important. C'est le cas par exemple des requêtes contenant des entités nommées ou des expressions avec des prépositions (« Le Monde » qui est un journal). Cependant, leurs élimination réduit considérablement la taille des index et limite leurs effets négatifs sur les calculs statistiques.
- **Lemmatisation :** Cette étape vise à réduire un terme à une forme canonique. La lemmatisation regroupe les différentes formes que peut revêtir un mot, soit : le nom, le pluriel, le verbe à l'infinitif, etc. Par exemple, le mot « jouer », verbe à l'infinitif ni accordé, ni conjugué est un lemme. Il possède différentes flexions qui correspondent à ses formes conjuguées à diverses personnes et temps : « il jouera », « nous jouons », « ils ont joué »... Grâce à la lemmatisation, les documents contenant différentes formes d'un même terme auront les mêmes chances d'être restitués. D'autre part, ceci va éviter à l'utilisateur de saisir les différentes formes des termes lors de la recherche. Par conséquent, cette étape réduit la taille de l'index et améliore le rappel (la part des documents pertinents retrouvés par le SRI par rapport à tous les documents pertinents). Cependant, elle peut réduire la précision (la part des documents pertinents par rapport à tous les documents restitués). Par exemple, l'ensemble des termes « operate operating operates operation operative operatives operational » va devenir « oper », ce qui implique une perte de précision pour des requêtes telles que : « operational and research ; operating and system ; operative and dentistry ». On distingue quatre types principaux de lemmatisation : en consultant un dictionnaire (ex. Tree-tagger (Schmid, 1994)), en utilisant les règles de transformation (ex. Porter Stemmer (Porter, 1980)), par troncature après X caractères et la méthode de n-grammes (Mayfield et McNamee, 2003).
- **Pondération :** 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. De ce fait, un poids est associé à chaque terme. Estimer l'importance d'un terme n'est pas une tâche simple. Prenons le cas d'une collection d'un million de documents. Un terme qui existe dans tous les documents n'est pas utile dans l'index parce qu'il ne peut fournir aucune information sur le document qui pourrait intéresser un utilisateur. Cependant,

un terme qui apparaît dans 5 documents uniquement peut être de grande valeur puisqu'il permet de pointer les documents pertinents. Pour ces raisons, des mesures qualitatives sont calculées au moment de l'indexation pour chaque terme. D'une part, ces mesures permettent d'estimer le degré d'importance des termes dans les documents. D'autre part, elles permettent d'éviter un temps de calcul supplémentaire durant la phase de l'appariement. La plupart de ces mesures sont basées sur les facteurs *TF* et *IDF*, qui permettent de combiner les pondérations locales (dans le document) et globales (dans la collection) d'un terme.

TF (Term Frequency) : cette mesure est proportionnelle au nombre d'occurrences d'un terme dans un document (pondération locale). Toutefois, il existe différentes variantes de cette mesure qui dépendent de la façon dont la pertinence est mesurée. 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.

IDF (Inverse Document Frequency) : se calcule selon la formule suivante :

$$IDF_t = log\left(\frac{N}{df_t + 1}\right) \tag{1.1}$$

N est le nombre de documents dans la collection et df_t est le nombre de documents dans lesquels le terme t apparaît. Cette mesure calcule la fréquence d'un terme dans la collection (pondération globale). Comme le montre la formule 5.1, cette mesure met en valeur les termes rares et limite l'importance des termes fréquents dans la collection.

La combinaison de TF et IDF fournit une autre mesure importante :

$$TFIDF_{t,d} = TF_{t,d} * IDF_t \tag{1.2}$$

Cette mesure donne pour un terme t un score important s'il apparaît fréquemment dans peu de documents et un score faible si le terme apparaît rarement dans un même document ou dans beaucoup de documents.

2.1.2 Requêtage

Les mêmes étapes que celles réalisées sur les documents sont répétées sur les requêtes. Cependant, aucun index n'est créé.

Une fois la normalisation des termes effectuée, une représentation des termes est préparée. Cette représentation dépend de la méthode de recherche (ou modèle de

recherche) utilisée au niveau de l'appariement. Par exemple, si le modèle utilisé est le modèle vectoriel (discuté dans un prochain paragraphe), la requête va prendre la forme d'un vecteur dans un espace où chaque terme distinct du corpus représente une dimension. Si le modèle est le modèle booléen, alors le système doit créer des formules logiques avec les termes de la requête en utilisant les opérateurs AND, OR et NOT...

Les SRI modernes pratiquent également des traitements complémentaires comme *l'extension de requête* (Vechtomova et Wang, 2006).

2.1.3 Appariement

Une fois les documents indexés et la requête analysée, le SRI procède à l'appariement entre la requête et les documents. De cette mise en correspondance résulte un score de pertinence reflétant le degré de similarité entre la requête et le document. En d'autres termes, le système prédit si l'utilisateur trouvera des informations pertinentes ou non dans le document. Ce score est calculé à partir d'une valeur appelée $RSV(q, d)$ (Retrieval Status Value), où q est une requête et d un document. Cette mesure tient compte des poids des termes calculés au moment de l'indexation. Les SRI actuels calculent des scores sous forme décimale. Ceci permet d'ordonner les documents restitués. La qualité de cet ordonnancement est primordiale. En effet, l'utilisateur se contente généralement d'examiner les premiers documents renvoyés (les 10 ou 20 premiers). Si la qualité des informations présentes dans cette tranche n'est pas satisfaisante, l'utilisateur considérera le SRI comme mauvais vis-à-vis de sa requête.

Différents modèles de RI ont été proposés dans la littérature afin de formaliser la pertinence, des modèles les plus naïfs basés sur l'appariement exact jusqu'aux modèles plus élaborés basés sur l'appariement flou. Dans la suite, nous présentons les principaux modèles de la littérature.

2.2 Modèles de RI

Les modèles de RI visent à fournir un cadre théorique pour interpréter la notion de pertinence et permettent ainsi de classer les documents vis-à-vis un besoin d'information. Un modèle de recherche d'information est représenté par le quadruplet [**D**,**Q**,**F**, **R(q,d)**] :

- **D** est l'ensemble des représentations des documents dans la collection.
- **Q** est l'ensemble des représentations du besoin d'information de l'utilisateur.
- **F** représente le cadre de modélisation des documents et des requêtes, ainsi que les relations entre eux. Les relations peuvent être des relations booléennes, des vecteurs ou des distributions de probabilités des termes.

– **R(q,d)** est la fonction d'ordonnancement qui attribue un score de pertinence pour le couple composé par une représentation de la requête $q \in Q$ et d'une représentation d'un document $d \in D$.

La fonction d'ordonnancement reflète l'intuition du modèle utilisé. Par exemple, pour le modèle booléen, **F** correspond à la théorie des ensembles. Pour le modèle vectoriel, **F** repose sur un espace vectoriel de N-dimensions, des représentations de requêtes et de documents sous formes de vecteurs...

Il existe une multitude de modèles de RI. La figure 1.2 présente la classification des modèles de RI selon (Baeza-Yates et Ribeiro-Neto, 1999). Comme illustré dans cette figure, les modèles de RI peuvent être regroupés selon le type du modèle mathématique utilisé, en trois grandes classes, à savoir :

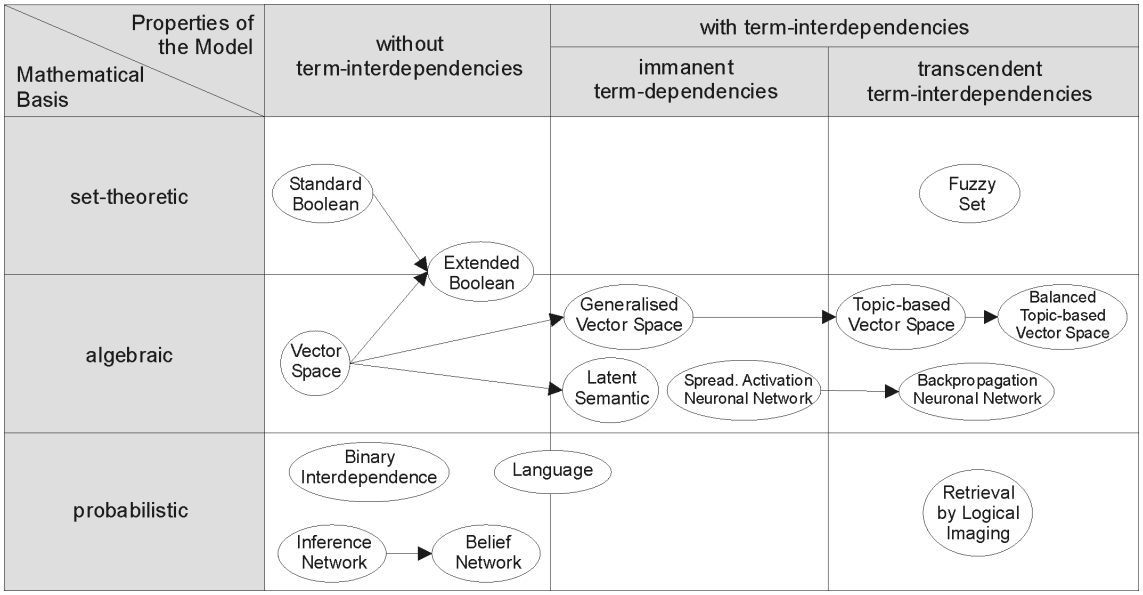


FIGURE 1.2 – Catégorisation des modèles de RI (Baeza-Yates et Ribeiro-Neto, 1999)

– Les modèles ensemblistes : ces modèles trouvent leurs fondements théoriques dans la théorie des ensembles. On distingue le modèle booléen pur (*boolean model*), le modèle boolen étendu (*extended boolean model*) et le modèle basé sur les ensembles flous (*fuzzy set model*).
– 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.
– Les modèles probabilistes, qui se basent sur les probabilités. Ils comprennent le modèle probabiliste général, le modèle de réseau de document ou d'inférence (*Document Network*) et les modèles de langue.

Dans le modèle booléen, les documents et les requêtes sont représentés sous la forme d'un ensemble de termes. Ainsi, comme suggéré dans (Gudivada et al., 1997),

il s'agit d'un modèle ensembliste. Dans le modèle vectoriel, les documents et les requêtes sont représentés sous formes de vecteurs dans un espace de N-dimensions. Pour le modèle probabiliste, le cadre de modélisation des documents et des requêtes est basé sur la théorie des probabilités.

Nous présentons dans la suite les principaux modèles issus de chacune de ces trois classes. Nous renvoyons le lecteur aux nombreux manuels introductifs à la RI (Baeza-Yates et Ribeiro-Neto, 1999 ; Manning et al., 2008) pour des présentations exhaustives des modèles de RI.

2.2.1 Modèle Booléen

Le modèle Booléen (Salton, 1968) est un modèle qui se base sur la théorie des ensembles et l'algèbre de Boole. **Le modèle Booléen** prend en compte uniquement la présence et l'absence d'un terme dans les documents : considérons le poids d'un terme i dans un document j $w_{i,j} \in \{0, 1\}$. Les poids des termes dans la matrice terme-document sont binaires. La requête q est, elle aussi, composée de termes reliés par des opérateurs logiques (ET, OU et NON). Ainsi le modèle vérifie si le document satisfait les conditions représentées par les termes de la requête. Le modèle booléen évalue si un document est pertinent ou non pertinent. Le score de chaque document sera ainsi représenté respectivement par 0 ou 1.

La décision binaire de pertinence sans aucune notion de graduation (*exact match*) réduit la qualité des résultats (notion de silence). En outre, les expressions booléennes ont une sémantique précise, ce qui rend la traduction du besoin d'information en une expression booléenne une tâche difficile. Ainsi, la majorité des expressions booléennes formulées par les utilisateurs sont simples (1 seul opérateur).

Même si la définition du besoin d'information sous forme d'une expression booléenne n'est pas toujours évidente pour les utilisateurs, le modèle booléen se caractérise par un formalisme simple et clair (représentation binaire des termes dans l'index). L'inconvénient principal est l'absence d'ordonnancement des résultats (car tous les $RSV = 1$), ce qui résulte parfois en la restitution d'un nombre très important ou très faible de documents.

2.2.2 Modèles vectoriels

Le modèle vectoriel (Salton et al., 1975) propose un cadre dans lequel la pertinence partielle est possible. Le poids des termes des documents et des requêtes n'est plus binaires. Le poids est utilisé pour mesurer la similarité entre les documents et le besoin d'information. Les documents sont ainsi ordonnés selon leur degré de similarité décroissant : du plus similaire au moins similaire ayant le score le plus faible. Le modèle vectoriel prend en compte les documents répondant partiellement

au besoin d'information. En outre, le modèle fournit une réponse plus raffinée que le modèle booléen dans le sens où il permet de sélectionner et de trier les documents.

Dans le modèle vectoriel, le document et la requête sont représentés par des vecteurs. Le degré de similarité entre un document d_j et une requête q est mesuré comme la corrélation entre les vecteurs $\vec{d_j}$ et $\vec{q}$. Cette corrélation peut être calculée par le cosinus entre les deux vecteurs.

$$\begin{aligned} sim(\vec{d_j}, \vec{q}) &= \frac{\vec{d_j}.\vec{q}}{|\vec{d_j}| \times |\vec{q}|} \\ &= \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \end{aligned} \tag{1.3}$$

D'autres fonctions de similarité ont été proposées dans la littérature, parmi lesquelles on peut citer les mesures de Jaccard et Dice (Manning et al., 2008).

Les poids des termes des requêtes et des documents dans les vecteurs sont généralement des scores basés sur $TF.IDF$. Ainsi,

$$w_{i,q} = \frac{(1 + log(tf_{i,q})) \times log(\frac{N}{n_i})}{\sqrt{\sum_k ((1 + log(tf_{k,q})) \times log(\frac{N}{n_k})^2)}} \tag{1.4}$$

$$w_{i,d_j} = \frac{(1 + log(tf_{i,d_j})) \times log(\frac{N}{n_i})}{\sqrt{\sum_k ((1 + log(tf_{k,d_j})) \times log(\frac{N}{n_k})^2)}} \tag{1.5}$$

avec TF représenté par $1 + log(tf_{i,d_j})$ et IDF représenté par $log(\frac{N}{n_i})$. Le reste de la fonction est utilisé pour la normalisation des scores.

Les avantages principaux du modèle vectoriel sont les suivants : tout d'abord, la pondération non binaire des termes favorise une meilleure qualité des résultats. Ensuite, 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.

Théoriquement, le modèle vectoriel a l'inconvénient de considérer que les termes de l'index sont tous indépendants. Cependant, en pratique, la prise en compte globale de la dépendance des termes peut faire baisser la qualité des réponses d'un système (Baeza-Yates et Ribeiro-Neto, 1999) car les dépendances sont généralement locales. C'est pour toutes ces raisons que le modèle vectoriel est encore populaire de nos jours en recherche d'information, et reste souvent utilisé comme une *baseline* (modèle de référence) lors de l'évaluation d'autres méthodes.

2.2.3 Modèle probabiliste

Le modèle probabiliste a été proposé par Robertson et Sparck Jones (1988). Il propose une solution à la problématique de la RI dans un cadre probabiliste : la

fonction de pertinence du modèle probabiliste se base sur le calcul de probabilités de pertinence des documents pour les requêtes données. Le principe de base consiste à retrouver des documents qui ont, dans le même temps, une forte probabilité d'être pertinents, et une faible probabilité d'être non pertinents. Ainsi, on distingue deux classes de documents pour une requête q_i : les pertinents (R) et les non pertinents ($\overline{R}$). Par conséquent, deux mesures de probabilité sont calculées : $P(R|d_j)$ la probabilité que le document d_j soit dans R et $P(\overline{R}|d_j)$ la probabilité que ce document soit dans $\overline{R}$. Ainsi, la pertinence entre le document d_j et la requête q est calculée par :

$$RSV(q, d_j) = \frac{P(R|d_j)}{P(\overline{R}|d_j)} \tag{1.6}$$

En appliquant la règle de Bayes et après quelques transformations, la formule précédente donne :

$$RSV(q, d_j) = \frac{P(d_j|R)}{P(d_j|\overline{R})} \tag{1.7}$$

Dans le modèle probabiliste de base, la représentation des documents est composée par des poids binaires indiquant la présence ou l'absence des termes, si on suppose que les termes sont indépendants, la formule 1.7 devient :

$$RSV(q, d_j) = \sum_{t_i \in T} x_i \cdot log \frac{p_i(1 - q_i)}{q_i(1 - p_i)} \tag{1.8}$$

avec T est l'ensemble de tous les termes, $x_i = 0$ si le terme i n'apparaît pas dans le document j ou bien $x_i = 1$ si le terme i apparaît dans le document j. $p_i = P(t_i \in D|R)$, $q_i = P(t_i \in D|\overline{R})$, $1 - p_i = P(t_i \notin D|R)$ et
$1 - q_i = P(t_i \notin D|\overline{R})$.

Lorsque des données d'apprentissage pour l'évaluation ne sont pas disponibles, on retrouve le facteur *idf* probabiliste intégré dans le modèle vectoriel :

$$RSV(q, d_j) = \sum_{t_i \in T} x_i \cdot log \left(\frac{N - R_i}{R_i} \right) \tag{1.9}$$

avec N le nombre de tous les documents et R_i est le nombre de documents contenant t_i.

Nous rappelons que, dans le modèle de base, les termes ont des poids binaires dans les documents, indiquant leur présence ou absence. La prise en compte des fréquences des termes dans les document a fait l'objet de plusieurs modèles variant du modèle de base. Par exemple, dans le modèle BM25 (Robertson et al., 1996) le calcul du poids d'un terme dans un document intègre différents aspects relatifs à la

fréquence locale des termes (tf_i), leur rareté et la longueur des documents :

$$x_i = \frac{(k_1 + 1) \cdot tf_i}{k_1 \times ((1 - b) + b \times \frac{dl}{avgdl} + tf_i} \tag{1.10}$$

avec dl est la taille du document d_j, $avgdl$ est la moyenne des tailles des documents dans la collection et k_1 ,b sont des paramètres qui dépendent de la collection ainsi que du type des requêtes.

2.3 Évaluation

L'évaluation des approches de RI est nécessaire afin d'estimer leur performance. C'est un moyen qui permet également de comparer différents systèmes et d'étudier l'impact des facteurs employés dans les approches. Un bon SRI doit satisfaire le besoin d'information de l'utilisateur. La qualité des résultats par rapport à ce besoin, la rapidité du système et la facilité d'utilisation du système représentent les principaux facteurs à évaluer pour un SRI (Mandl, 2007). Nous nous intéressons ici à celui qui nous semble le plus important : la capacité d'un système à sélectionner des documents pertinents. Le mode d'évaluation généralement utilisé aujourd'hui est basé sur celui développé dans le projet Cranfield (Cleverdon et al., 1966) communément appelé le paradigme de Cranfield. Ce paradigme définit la méthodologie d'évaluation des SRI en se basant sur 3 éléments : un corpus de documents sur lequel les recherches sont effectuées, un ensemble de requêtes de test (besoins des utilisateurs) et la liste des documents pertinents pour chacune des requêtes (la vérité terrain). 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.

2.3.1 Collection de test

Les collections de test permettent de comparer directement des résultats obtenus par des systèmes en utilisant des modèles différents. Nous détaillons ci-dessous les différentes parties de ces collections.

– Les requêtes sont un ensemble de besoins d'information utilisés pour le test. Cet ensemble est appelé également *topics* dans le jargon des campagnes d'évaluations telles que TREC, INEX... Le nombre de requêtes doit être important afin d'être le plus représentatif possible de la réalité et pour avoir une évaluation objective. Il faut au moins 25 requêtes pour garantir la qualité de l'évaluation au regard de la statistique (Buckley et Voorhees, 2000). Les requêtes sont souvent créées par les assesseurs des organismes qui organisent l'évaluation. Toutefois, elles peuvent être de vraies requêtes extraites à partir des logs des moteurs de recherche (Baeza-Yates et Ribeiro-Neto, 1999).

- Le corpus de documents est l'ensemble de documents présélectionnés. Il existe plusieurs corpus disponibles. Ces corpus diffèrent selon plusieurs critères en fonction de la tâche de recherche que l'on veut évaluer, des documents plus ou moins vulgarisés, plus ou moins spécialisés dans un domaine, dans une langue ou une autre...
- Les jugements de pertinence identifient les documents pertinents pour une requête et représentent la vérité terrain. Un score de pertinence graduel peut éventuellement être associé pour chaque couple *document/requête*. La réalisation de ces jugements est loin d'être une tâche facile. Il s'agit d'un processus long et coûteux impliquant des humains. Pour de petites collections comme celle de Cranfield, il existe des jugements de pertinence exhaustifs pour chaque paire requête-documents. Cependant, pour les grandes collections modernes, les jugements ne se font généralement que pour un sous-ensemble des documents pour chaque requête. L'approche la plus standard est celle du *pooling* (Jones et Rijsbergen, 1976), où la pertinence est évaluée sur un sous-ensemble de la collection formé à partir des premiers documents retournés par un certain nombre de systèmes différents (généralement ceux à évaluer), et parfois complété par d'autres sources telles que les résultats de recherches booléennes par mots clés ou des documents trouvés par les chercheurs experts dans un processus interactif.

De nombreux projets basés sur des corpus d'évaluation se multiplient depuis les années 1970. On peut par exemple citer la collection Cranfield ou encore la campagne CLEF (Cross Language Evaluation Forum) [14]. La campagne la plus connue est sans conteste TREC (Text REtrieval Conference) organisée annuellement depuis 1992 par le NIST [15] et la DARPA [16]. Elle a pour but d'encourager la recherche documentaire basée sur de grandes collections de test, tout en fournissant l'infrastructure nécessaire pour l'évaluation des méthodologies de recherche et de filtrage d'information. Dans ce qui suit, nous présentons les corpus les plus populaires issus de différents projets d'évaluation :

- **Conférence Text Retrieval (TREC)**. Le US National Institute of Standards and Technology (NIST) a organisé une grande série d'évaluations depuis 1992. Dans ce cadre, plusieurs tâches qui se basaient sur différentes collections d'essais ont été définies. On peut par exemple citer les collections utilisées pour la tâche adhoc entre 1992 et 1999. Au total, ces collections comprennent 6 CD contenant 1,89 millions de documents et les jugements de pertinence pour 450 besoins d'information. Les premières collections étaient composées chacune de 50 besoins d'information, évalués sur différentes ensembles de do-

14. `http://www.clef-initiative.eu/`
15. National Institue of Standards and Technology (www.nist.gov)
16. Defense Advanced Research Project Agency

cuments. TREC 6-8 fournit 150 besoins d'information sur environ 528 000 articles. Étant donné les collections de documents si grandes, il n'y a pas de jugements de pertinence exhaustifs. Au contraire, les jugements de pertinence sont disponibles uniquement pour les documents qui sont restitués parmi les premiers documents retournés pour les systèmes qui ont participé à l'évaluation (*pooling*).

- **Le projet NTCIR** a développé diverses collections d'essais de tailles similaires aux collections de TREC, en se concentrant sur les langues d'Asie de l'Est et la recherche d'information multilingue. Les requêtes sont faites dans une langue, toutefois, les collections de documents contiennent des documents dans une ou plusieurs autres langues.
- **CLEF (La campange Cross Language Evaluation Forum)** a également proposé plusieurs collections. Elle s'est concentrée sur les langues européennes et la recherche d'information multilingue.

On trouvera plus de détails sur l'évaluation à base de collections de test dans (Sanderson, 2010).

2.3.2 Mesures d'évaluation

En RI, la mise au point des modèles passe par une phase expérimentale qui suppose l'utilisation de métriques qui visent à comparer des modèles entre eux ou à mettre au point leurs paramètres. Les deux métriques de base les plus utilisées pour évaluer l'efficacité de la RI sont la précision et le rappel. Celles-ci sont définies pour le cas simple où un système renvoie un ensemble de documents vis-à-vis d'une requête (Voorhees, 2006).

La mesure de précision calcule la capacité du système à rejeter tous les documents non pertinents pour une requête. Elle est donnée par le rapport entre les documents sélectionnés pertinents et l'ensemble des documents sélectionnés :

$$\text{Précision} = \frac{|\text{Documents pertinents restitués}|}{|\text{Documents restitués}|} \in [0,1] \tag{1.11}$$

Le rappel calcule la capacité du système à restituer le maximum de documents pertinents pour une requête. Il mesure la proportion de documents pertinents restitués par le système relativement à l'ensemble des documents pertinents contenus dans la base documentaire. Il est exprimé par :

$$\text{Rappel} = \frac{|\text{Documents pertinents restitués}|}{|\text{Documents pertinents}|} \in [0,1] \tag{1.12}$$

Le rappel et la précision sont calculés indépendamment de l'ordre dans lequel les résultats sont représentés (ce sont des mesures ensemblistes). Des mesures tenant compte de l'ordre des documents sont également nécessaires. Elles permettent

par exemple d'évaluer des systèmes tels que les moteurs de recherche du web où l'ordre d'apparition des documents est crucial. À cet égard, les mesures principales proposées sont la **précision@X** et la **précision moyenne**.

La **précision@X** est la précision à différents niveaux de coupe de la liste. Cette précision mesure la proportion des documents pertinents retrouvés parmi les X premiers documents restitués par le système.

La **précision moyenne** est la moyenne des valeurs de précisions après chaque document pertinent. Elle se focalise en particulier sur les document pertinents classés dans les premiers rangs.

$$AP_q = \frac{1}{R} \sum_{i=1}^{N} p(i) \times R(i) \tag{1.13}$$

Où $R(i) = 1$ si le i$^{\text{ème}}$ document restitué est pertinent, $R(i) = 0$ si le i$^{\text{ème}}$ document restitué est non pertinent, $p(i)$ la précision à i documents restitués. R le nombre de documents pertinents pour la requête q et N le nombre de documents restitué par le système.

La **moyenne des précisions moyennes** (Mean Average Precision-MAP) est obtenue sur l'ensemble des requêtes :

$$MAP = \frac{\sum_{q \in Q} AP_q}{|Q|} \tag{1.14}$$

avec AP_q est la précision moyenne d'une requête q, Q est l'ensemble des requêtes et $|Q|$ est le nombre de requêtes. Cette mesure peut être qualifiée de globale puisqu'elle combine différents points de mesure.

Il existe plusieurs autres mesures qui peuvent servir à évaluer les SRI. Nous pouvons citer par exemple **la F-mesure, la R-précision, la BPREF** (Binary PREFerence-based measure), **la MRR** (Mean Reciprocal Rank) détaillées dans (Sanderson, 2010).

Nous avons vu dans cette section que les approches classiques de RI se basent généralement sur la fréquence des termes, que ce soit dans les documents ou dans le corpus, pour mesurer la pertinence. En outre, un document est considéré uniquement par son contenu présenté comme un sac de mots. Cependant, avec l'évolution des techniques du web 2.0, un document peut être représenté, non seulement par son contenu, mais aussi par d'autres informations sociales telles que ses liens avec les autres documents, des annotations, les commentaires des utilisateurs... Nous présentons dans la section suivante les différentes approches de RI utilisant ces informations sociales.

3 Utilisation des informations sociales en RI : RI sociale

La RI sociale consiste à adapter les modèles et les algorithmes de la RI classique en exploitant les informations sociales. Il s'agit de satisfaire les besoins d'information des utilisateurs en exploitant par exemple les connaissances des utilisateurs experts ou bien les expériences de recherche des autres utilisateurs. Cet objectif se réalise concrètement en considérant les annotations sociales (Peters et al., 2011), l'analyse des réseaux sociaux (Kazai et Milic-Frayling, 2008), les jugements de pertinence subjectifs (Xu et al., 2007) et la recherche collaborative(Karamuftuoglu, 1998) dans le processus de la RI. Comme le montre la figure 1.3, les informations sociales peuvent être exploitées au sein même du modèle de RI (modèle de document et de requête, fonction de pondération / de correspondance), ou en aval de ce modèle (reclassement de la liste des résultats) et même comme une source parmi d'autres dans le web.

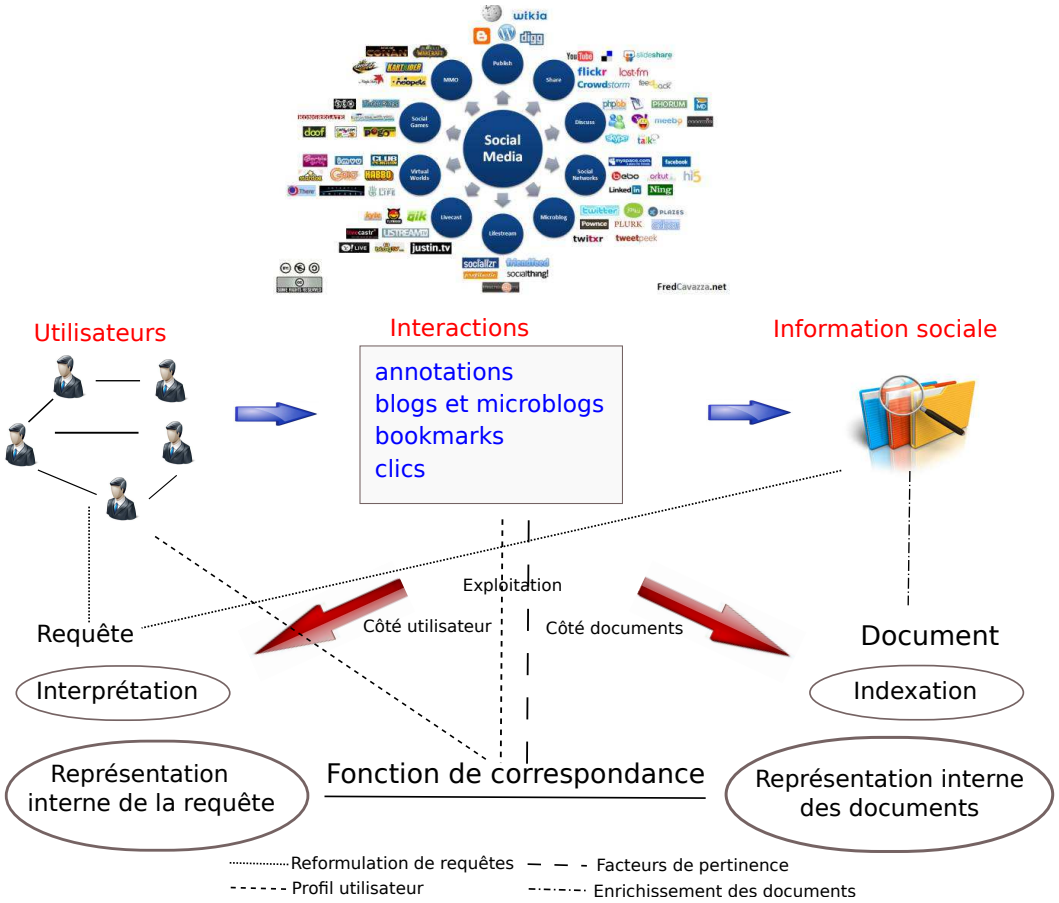


FIGURE 1.3 – Exploitation de l'information sociale dans la RI

Dans cette section, nous nous focalisons sur l'impact de l'information sociale sur le processus de RI. L'objectif étant d'améliorer la qualité des résultats, les informa-

tions sociales sont employées à plusieurs niveaux. Nous pouvons classer les approches exploitant l'information sociale en fonction du niveau de leur utilisation (côté utilisateur ou côté documents ; figure 1.3). D'une part, ces informations ont été ainsi employées du côté de l'utilisateur pour reformuler les requêtes ou bien pour définir un profil et contextualiser les résultats. D'autre part, du côté des documents, les informations sociales ont été utilisées pour enrichir la représentation des ressources documentaires.

3.1 Côté utilisateur

L'idée est d'améliorer l'efficacité des SRI en exploitant le contexte de l'utilisateur. Ceci se réalise en tenant compte des informations sur l'utilisateur (telles que son profil ou ses informations personnelles) dans le processus de recherche ou bien en améliorant la représentation de son besoin d'information, dans le but de retrouver des résultats plus spécifiques et plus raffinés. Ainsi, plusieurs travaux ont exploité l'information sociale comme moyen de reformulation de requêtes ou de création de profil pour une recherche personnalisée.

3.1.1 Information sociale pour la reformulation de requêtes

La reformulation de requêtes est vue comme un traitement pour élargir le champ de recherche pour une requête. Une requête étendue va contenir plus de termes reliés permettant d'une part de désambiguïser les mots initiaux et connaître exactement leurs sens, et d'autre part d'augmenter les chances de restituer le maximum de documents pertinents.

L'information sociale peut ainsi être utilisée pour étendre les requêtes. Koolen et al. (2009) 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 Y. Li et al. (2007) qui traite les requêtes dites « faibles ». Ces requêtes ne permettent pas de récupérer suffisamment de documents pertinents lors de la première recherche. Cette approche a montré une amélioration de qualité, en particulier sur les premiers documents renvoyés.

Bai et al. (2007) ont utilisé ODP (Open Directory Project) [17] afin de contextualiser les besoins d'information. L'idée est d'étendre les requêtes avec des ensembles de mots extraits de documents du *feedback*. L'ensemble du *feedback* est composé de documents qui sont pertinents ou pseudo-pertinents par rapport à la requête initiale, et qui sont à même de contenir des informations importantes sur le contexte

17. `http://www.dmoz.org/`

de la recherche. Les mots exprimant le plus d'information par rapport à la requête sont traités comme des concepts implicites. Ils sont alors utilisés pour reformuler la requête.

3.1.2 Information sociale pour la création de profil et la recherche personnalisée

Un profil d'utilisateur est constituée des préférences de restitution de l'utilisateur, ainsi que des contraintes sur les résultats présentés. Les informations sociales ont également été utilisées pour créer les profils des utilisateurs. Les profils sont par la suite utilisés pour définir un contexte de restitution permettant de sélectionner des résultats personnalisés. Les éléments souvent utilisés pour créer le profil d'un utilisateur sont ses relations sociales, ses annotations et ses activités dans les plate-formes sociales. Les profils à base d'informations sociales ont été utilisés, par exemple, pour faciliter la personnalisation des recherches à partir d'un environnement de marquage collaboratif. Cai et Li (2010) se sont concentrés sur l'exploration de recherches personnalisées à travers la proposition d'une approche qui permet de créer des profils d'utilisateurs basés sur les tags, ainsi que la création de profils des ressources à rechercher.

L'information sociale peut être utilisée pour personnaliser la recherche. Carmel et al. (2009), de leur côté, exploitent les relations sociales de l'utilisateur. Les résultats d'une recherche sont de nouveau classés en fonction des relations avec des personnes dans le réseau social de l'utilisateur. Les auteurs ont étudié l'impact de plusieurs types de réseaux sociaux pour la personnalisation : (1) réseau basé sur les connaissances liées à l'utilisateur via une connexion de familiarité ; (2) réseau de personnes « similaires » à l'utilisateur et qui ont des activités sociales semblables, (3) le réseau global représenté par les deux types de relations.

Les informations sociales sont également utilisées dans les moteurs de recherches du web. Google, par exemple, propose un outil pour chercher dans les information du réseau social. En choisissant « résultats personnels » (figure 1.4), un internaute est susceptible de retrouver les profils et les documents partagés par son cercle social, que ce soit celui de Google+ [18], de Twitter, de flickr [19]... Google propose également des utilisateurs qui semblent avoir les mêmes centres d'intérêts. Bing propose également son outil de recherches social *Bing social search* (figure 1.5). Cette fonctionnalité permet non seulement d'exploiter le réseau social pour valoriser les résultats du cercle, mais également de retrouver des personnes expertes et susceptibles de disposer des meilleures informations sur le sujet cherché. Bing exploite la majorité des plate-

18. `https://plus.google.com/`
19. `https://www.flickr.com/`

formes sociales tels que Facebook [20], Twitter, Klout [21] et même Google+.

FIGURE 1.4 – Résultats à partir du cercle social dans Google

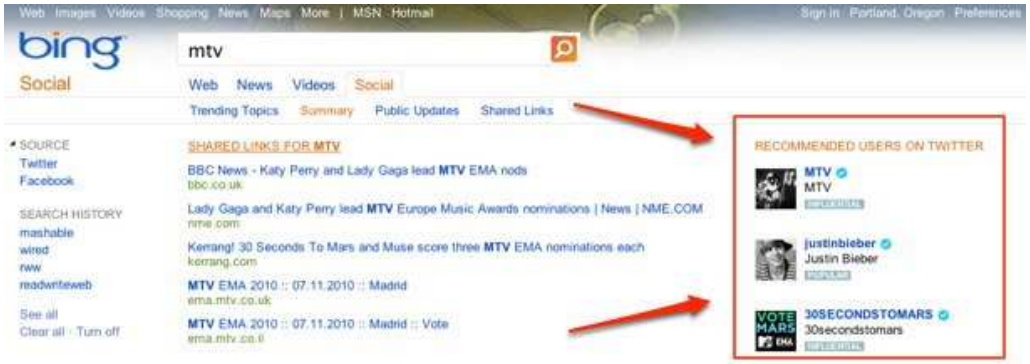


FIGURE 1.5 – Recommandation de profils expert sur le sujet recherché sur Bing

3.2 Côté documents

L'idée sous-jacente à l'utilisation des informations sociales du côté des documents est de ramener des informations supplémentaires pour enrichir la représentation des contenus recherchés ou bien pour les utiliser comme des facteurs de pertinence.

3.2.1 Information sociale pour l'enrichissement des ressources documentaires

Les commentaires et les tags réalisés par un utilisateur du web sur les contenus publiés par les autres utilisateurs dépendent fortement de ses connaissances et ses centres d'intérêts. Ainsi, ces données représentent une valeur ajoutée (des méta-données), que ce soit pour la création de profil de l'utilisateur ou bien pour enrichir le contenu et la représentation des documents. Cai et Li (2010) ont utilisé les tags pour générer des profils des ressources d'informations et des profils des utilisateurs.

20. `https://www.facebook.com/`
21. `https://www.klout.com/`

La correspondance des deux types de profils a permis d'améliorer la qualité des résultats.

Les informations sociales ont été utilisées également pour enrichir la représentation des ressources au moment de l'indexation. Attardi et Simi (2006) ont utilisé les opinions obtenues de la base lexicale « SentiWordNet » pour enrichir l'index des documents avec des étiquettes d'opinion. L'intuition est que l'utilisation des étiquettes marquant l'opinion permet de surpondérer les scores pour les documents qui ne seraient pas sélectionnées avec un simple calcul statistique sur les fréquences d'occurrence. Cet enrichissement de l'index permet d'améliorer le rappel.

3.2.2 Information sociale comme facteur de pertinence

Le contenu social a démontré son avantage pour l'amélioration et l'enrichissement des contenus. De même, ces contenus sont utilisés au niveau de la mesure de la pertinence d'un document, comme un facteur parmi d'autres.

Bao et al. (2007) ont trouvé que le *social bookmarking* peut améliorer les recherches sur le web selon deux aspects : 1) les annotations représentent généralement de bons résumés pour les pages web correspondant ; 2) le nombre d'annotations indique la popularité des pages web. Ainsi, deux nouveaux algorithmes sont proposés pour intégrer les facteurs ci-dessus dans le classement de la page : 1) SocialSimRank (SSR) calcule la similarité entre les annotations sociales et les requêtes ; 2) SocialPageRank (SPR) capte la popularité des pages web en fonction des annotations qui y sont réalisées (Bao et al., 2007).

D'autres travaux ont relié la pertinence avec l'importance de leurs auteurs. En fait, plus l'auteur est populaire, plus l'information est fiable. La popularité d'un auteur est ainsi mesurée à travers ses informations sociales. C'est l'exemple de l'approche de Macdonald et Ounis (2006), qui ont proposé un modèle de recherche d'information mesurant la pertinence en fonction de l'expertise de son auteur par rapport au besoin d'information. Ils supposent que chaque document représente un vote pour chaque personne qui le cite. C'est le cas également de l'approche de Korfiatis et al. (2006) qui ont évalué les documents de Wikipédia à travers la popularité de leurs auteurs. Pour ce faire, ils ont construit un modèle du réseau social de Wikipédia et ont défini des mesures de qualité telles que la centralité des auteurs. Les auteurs ont trouvé que cette méthode d'évaluation est prometteuse, particulièrement avec les articles traitant de sujets susceptibles d'exposer différents points de vus, tels que les sujet politiques. Kazai et Milic-Frayling (2008) ont défini la notion de confiance accordée à un auteur. Cette confiance reflète la pertinence du document publié par l'auteur. Elle se calcule à travers la centralité du nœud du sous-graphe d'un auteur. Ce sous-graphe est obtenu à partir du graphe composé par plusieurs acteurs (auteurs, éditeurs et consommateurs), ainsi que des liens de données (publi-

cations) et des différentes relations sociales (tels que les collaborations, les citations et les annotations entre les différents acteurs).

Outre l'expertise, la popularité et la confiance, les informations sociales sont utilisées comme facteur de pertinence relié à la fraîcheur de l'information. Dong, Zhang, et al. (2010) ; Dong, Chang, et al. (2010) ont proposé d'utiliser les informations publiées sur les plate-formes de microblogging pour détecter les nouvelles URLs qui sont susceptibles de ne pas être encore indexées par les moteurs de recherche. Les auteurs ont également utilisé les informations sociales obtenues à partir des plate-formes de microblogging comme des mesures de pertinence et de qualité des documents pointés par les URLs.

4 Conclusion

Nous avons présenté dans ce chapitre l'information sociale dans le Web, développée avec l'évolution des technologies du Web 2.0. Nous avons ensuite décrit les concepts de base de la RI classique et, en particulier, ceux que nous utilisons dans nos travaux. Enfin, nous avons discuté l'impact de l'évolution de ces informations sociales sur le processus de RI, ainsi que leur emploi dans le but d'améliorer l'efficacité des SRI.

Outre l'amélioration des résultats de la RI, l'information sociale s'est imposée comme une source d'information parmi d'autres dans le Web [22]. La forte demande en égard à cette source d'information réclame l'adaptation des approches de RI dans les différentes tâches (par exemple, la détection d'opinion, la recherche d'expert, la recherche adhoc...) sur les informations sociales. Dans le chapitre suivant, nous présentons un aperçu des différentes tâches de RI sur l'information sociale, ainsi qu'un aperçu des approches de l'état de l'art. Nous nous focalisons uniquement sur les informations publiées sur la plate-formes de microblogging Twitter, celle-ci constituant le cadre applicatif de notre travail.

22. Par exemple, il y a en moyenne 2 milliards requêtes soumises sur Twitter par jour (contre 5 milliards sur Google) : `http://www.statisticbrain.com/`

Chapitre 2

RI dans les microblogs

Nous présentons dans ce chapitre la recherche d'information dans les microblogs, et en particulier, la recherche adhoc de microblogs. Les microblogs sont une forme réduite des blogs. Ils représentent une source d'information récente. Les utilisateurs emploient des plate-formes de microblogging pour partager et accéder à des microblogs. Ces plate-formes prennent la forme de réseaux sociaux qui se distinguent par des interactions sociales intenses et une diversité dans les sujet discutés, par rapport aux autres sources d'information.

Il existe plusieurs plate-formes de microblogging. Les 5 plate-formes les plus utilisées [1] sont Twitter, FriendFeed [2], Tumblr [3], Posterous [4] et Identi.ca [5]. Parmi elles, Twitter est sans conteste la plus utilisée. Cette plate-forme compte plus de 650 millions d'utilisateurs, publiant en moyenne 58 millions de tweets par jour. Twitter est utilisé également comme source d'information. En moyenne, 2,1 milliards de requêtes sont soumises chaque jour sur son moteur de recherche.

La RI dans les microblogs est différente de la recherche dans le Web. Ceci est dû aux différences de forme des microblogs par rapport aux documents du web, à la spécificité de leur contenu et également aux motivations des recherches (information fraîches...). Les travaux de la littérature qui portent sur la RI dans les microblogs peuvent être regroupés en deux catégories. La première porte sur l'étude des caractéristiques et l'analyse statistique des *microblogs*. Kwak et al. (2010), par exemple, ont étudié les spécificités linguistiques, démographiques, topographiques et spatio-temporelles des microblogs. La seconde porte sur les tâches de recherche d'information au sens large (accès à l'information) dans les *microblogs*. Notre travail s'inscrivant dans la seconde catégorie, nous détaillerons plus particulièrement dans ce chapitre les tâches de RI dans les microblogs. Auparavant, nous commençons

1. `http://www.gurugrounds.com/uncategorized/top-10-microblogging-sites/`
2. `http://friendfeed.com/`
3. `https://www.tumblr.com/`
4. `http://www.posterous.com/`
5. `https://identi.ca/`

par présenter les spécificités des plate-formes de microblogging et des microblogs, en s'attachant au cas de Twitter. Nous considérons cette plate-forme étant donné qu'elle représente le cadre applicatif de tous les travaux de la RI dans les microblogs de l'état de l'art, d'une part, et, d'autre part, parce qu'elle est la plus utilisée dans le monde réel. Les autres plate-formes de microblogging ont pratiquement les mêmes spécificités et le même principe de fonctionnement.

1 Présentation et spécificités des plate-formes de microblogging : cas de Twitter

1.1 Présentation générale de Twitter

Twitter est l'exemple le plus populaire des plate-formes de microblogging. Ces plate-formes sont les réseaux sociaux les plus récents du Web 2.0. Elles sont considérées comme une nouvelle forme de blogs, où les informations diffusées sont courtes et publiées plus rapidement. Ces informations concernent différents sujets. Les utilisateurs parlent de leur quotidien, des événements, des tendances... parfois à la mode SMS et en partageant des messages de faible longueur (par exemple 140 caractère au plus dans le cas de Twitter).

Twitter a connu une croissance exponentielle durant ces dernières années. Nous présentons ci-dessous les principales spécificités de cette plate-forme, ainsi que l'information qui y est produite.

1.1.1 Lancement et évolution

L'idée de base de Twitter est de permettre aux amis, aux familles et aux collaborateurs de communiquer et de rester connectés en partageant des réponses rapides et fréquentes (tweets) à la question : Quoi de neuf ? Plusieurs études (Java et al., 2007 ; Mischaud, 2007) ont cependant montré que les utilisateurs de Twitter dépassent ce premier objectif, en documentant dans leurs messages leur vie quotidienne, en partageant des hyperliens et en commentant des événements. Ces pratiques ont transformé le microblogging. C'est désormais un moyen de partager son état d'esprit personnels, mais aussi de publier des histoires et des nouvelles, pour exprimer ses opinions, pour discuter sur différents sujets dans des contextes sociaux, économiques et même politiques...

Lancée en octobre 2006, la plate-forme comptait 94,000 utilisateurs en avril 2007 [6] pour atteindre 200 millions en 2012 [7]. Au début de 2014, Twitter compte plus de

6. http://usatoday.com/tech/webguide/2007-05-28-social-sites_N.htm
7. http://www.bbc.co.uk/news/business-12889048

645 millions d'utilisateurs actifs [8].

Concernant le trafic, le nombre de tweets publiés croît tous les jours. En mars 2007, en moyenne, les microbloggeurs publient 20 000 tweets par jour [9]. Ce nombre a évolué pour atteindre 50 millions en janvier 2010. Aujourd'hui, **le nombre de tweets par jour peut atteindre 500 millions** [10].

Outre la publication de tweets, les microbloggeurs exploitent cette plate-forme pour chercher des informations récentes sur des sujets particuliers. En 2011, le nombre de requêtes soumises au moteur de recherche de Twitter était de l'ordre de 1,6 millions de requête par jour [11]. Ce nombre a évolué pour atteindre 2,1 milliards de requêtes [12] par jour en 2013.

Avec cette évolution, Twitter s'est rapidement positionné parmi les premières sources d'information utilisées sur le Web. Le tableau 2.1 liste le nombre de requêtes soumises à Google, Twitter et Facebook chaque jour. Le nombre de requêtes soumise à Twitter correspond à 42 % des requêtes soumises à google. Ce chiffre montre l'importance de Twitter en tant que source d'informations et la dépendance des utilisateurs à cette source d'information.

Source	Année	Nb de requêtes
Google	2013	5,1
Twitter	2013	2,1
Facebook	2012	1,0

Tableau 2.1 – Nombre de requêtes par jours (en milliard). Chiffres obtenus du site `http://statisticbrain.com`

1.1.2 Concepts et fonctionnement des plate-formes de microblogging

La figure 2.1 montre l'interface de Twitter. L'interface est composée de plusieurs sections. Dans la section *Tweets* appelée également *Timeline*, un utilisateur peut voir le flux de ses tweets ainsi que ceux de ses amis, triés par ordre chronologique inverse. On peut remarquer également une section de tendances qui contient les 10 sujets les plus populaires dans Twitter à un moment donné. L'utilisateur peut consulter les tendances du monde entier, comme il peut se focaliser sur un endroit

8. `http://www.statisticbrain.com/twitter-statistics/`
9. `http://www.begeek.fr/twitter-90-millions-de-tweets-par-jours-21210`
10. `http://www.blogdumoderateur.com/statistiques-twitter-entree-en-bourse/`
11. `http://engineering.twitter.com/2011/05/engineering-behind-twitters-new-search.html`
12. `http://www.statisticbrain.com/twitter-statistics/`

plus spécifique. La plate-forme suggère également des utilisateurs qui ont des centres d'intérêts similaire à l'utilisateur courant dans la section *suggestions*.

FIGURE 2.1 – L'interface graphique utilisateur de Twitter

En s'inscrivant sur une plate-forme de microblogging, un utilisateur fournit plusieurs informations telles que sa photo, sa localisation, son site Web et une courte bibliographie (figure 2.2). Dans la bibliographie, les utilisateurs décrivent généralement leurs activités et leurs centres d'intérêt. Ces informations sont ensuite probablement utilisées par les plate-formes dans la recommandation des utilisateurs.

La figure 2.3 donne un exemple d'utilisation d'une plate-forme de microblogging. Un utilisateur A peut suivre le flux de microblogs envoyés par un utilisateur C sans lui demander la permission (sauf pour les comptes privés que nous ne détaillons pas ici). Les relations entre utilisateurs des réseaux sociaux sont appelées des abonnements. Si A est abonné à C, alors A est appelé abonné (*follower*) de C (*followee*) et reçoit automatiquement toutes les publications de C dans sa *timeline*. Les relations d'abonnement peuvent être unilatérales (dans un seul sens), mais également bilatérales (dans les deux sens) si C s'abonne à son tour à A. On parle dans ce cas d'une relation d'amitié. Si un microbloggeur diffuse un message, tous ses abonnés

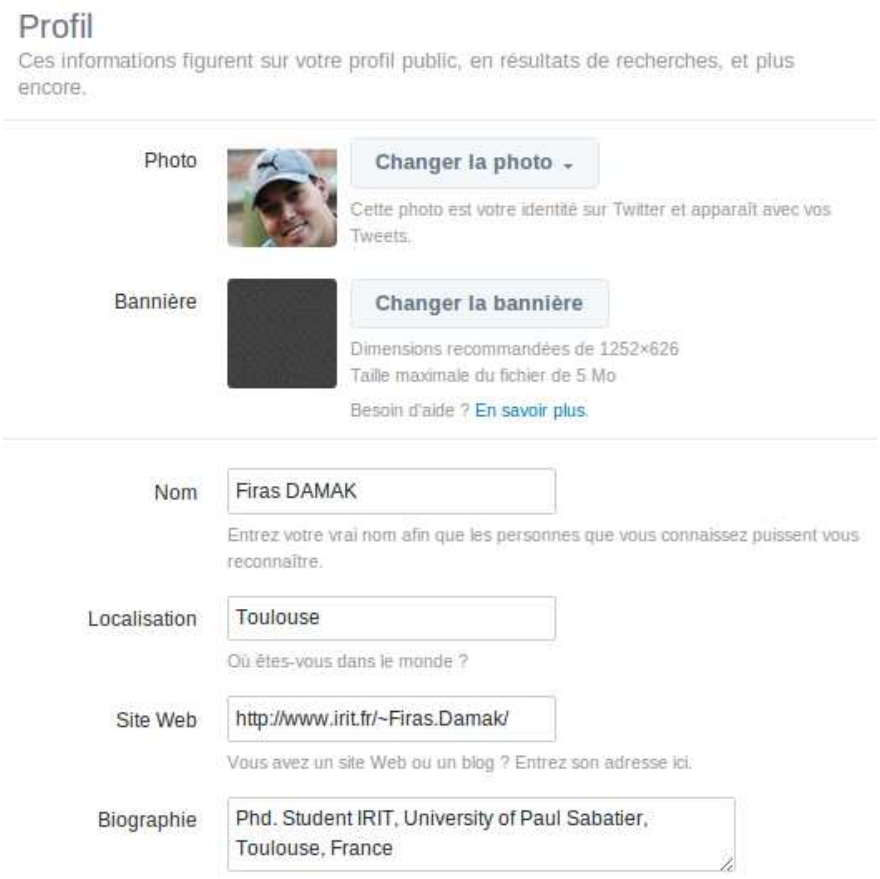


FIGURE 2.2 – Informations des comptes utilisateurs sur Twitter

le reçoivent. Un microbloggeur peut également envoyer un message direct et privé à l'un de ses amis (*direct message*). Si le microbloggeur partage un message pour la première fois, le message sera un *tweet*, sinon, s'il le rediffuse, le message sera un *retweet* et il va contenir dans ce cas la mention **RT**. En rediffusant un microblog, un microbloggeur peut y ajouter de l'information complémentaire. Finalement, et comme indiqué plus tôt, un utilisateur peut en mentionner un autre dans un message (@mention).

Les individus ne sont pas les seuls propriétaires de comptes. Les entreprises ou encore les sites d'information sont aujourd'hui très présents sur les plate-formes de microblogging.

1.1.3 Système temps-réel

L'une des spécificités fondamentale des plate-formes de microblogging est leur nature temps-réel : la présentation des publications (*timeline*), la présentation des résultats de recherches, le traitement du contenu publié...

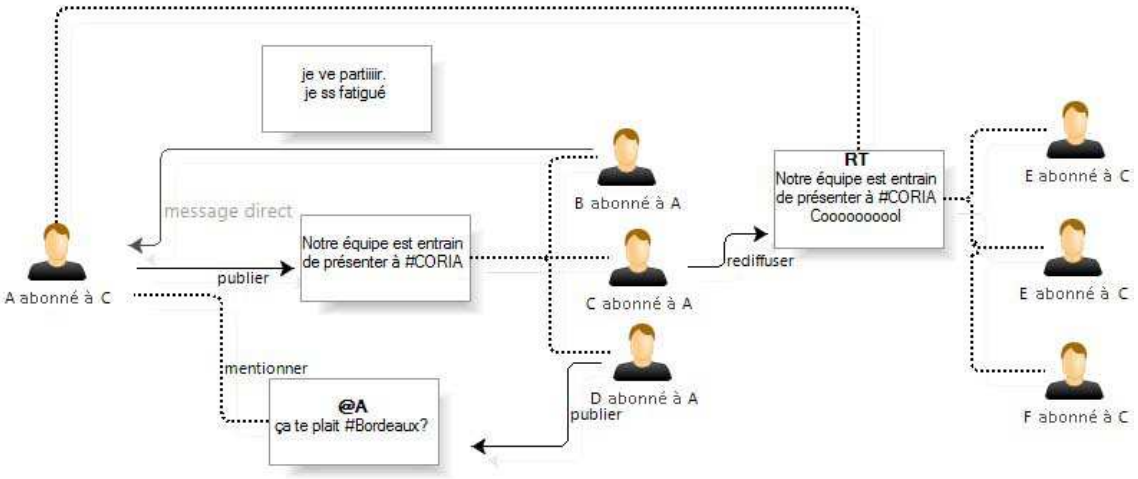


FIGURE 2.3 – Exemple d'utilisation de Twitter (avec tweets, retweets, abonnements et hashtags)

1.1.3.1 ***Timeline*** Twitter, comme les autres plate-formes de microblogging, est un système temps-réel par excellence dont la fraîcheur est la spécificité la plus importance. Cette spécificité peut être aperçue à plusieurs niveaux :

- Un utilisateur, en accédant à sa page, reçoit en temps-réel les microblogs de ses abonnés. Ces microblogs défilent sur sa page et le plus récent s'affiche au début de la file.
- Pour répondre à un besoin d'information, le moteur de recherche de Twitter affiche les tweets-résultats par ordre chronologique inverse (des plus récents aux plus anciens). Si à un moment donné un nouveau microblog pertinent est publié, l'utilisateur reçoit une notification pour l'afficher (figure 2.4).

FIGURE 2.4 – Notification sur l'apparition de nouveaux résultats dans Twitter

- En dépit de la quantité de microblogs publiée chaque seconde, un système de microblogging indexe ces contenus et les rend disponibles à tous les utilisa-

teurs à l'instant même de leur publication. Ceci représente une révolution par rapport aux autres sources d'information du Web. Google, par exemple, met jusqu'à une semaine pour indexer une page Web [13]. Wikipédia met jusqu'à une année pour inclure des modifications sur ses pages (Frank et al., 2013, 2012).

1.1.3.2 Usage temps-réel Alors que les blogueurs mettent à jour leurs blogs une fois tous les quelques jours, les microbloggeurs postent généralement plusieurs microblogs chaque jour (Java et al., 2007), en particulier pour décrire des événements qui se déroulent au moment de l'écriture du microblog. Ainsi, les microblogueurs peuvent savoir à tout moment ce que les autres microblogueurs sont en train de faire ou à quoi ils sont en train de penser.

Un grand nombre de tweets publiés sur Twitter ont rapport à des événements. Il peut s'agir d'événements sociaux tels que des fêtes, des compétitions sportives et des campagnes présidentielles. Il peut également s'agir de catastrophes telles que des tempêtes, des incendies, des émeutes, des fortes pluies et des tremblements de terre, ou bien tout simplement d'informations sur des embouteillages (Endarnoto et al., 2011). Twitter est un outil de notification temps-réel de tous se qui se passe dans le monde. C'est un moyen rapide et fiable pour transmettre les informations dans des situations critiques nécessitant des interventions d'urgence (incendies par exemple). Twitter a été ainsi utilisé par les victimes des incendies en Californie [14] et en Australie en 2009, pour décrire exactement la situation et aider les autres victimes en transmettant les informations utiles aux secours. Twitter peut également être utilisé pour faire du reportage temps-réel, comme cela a été le cas lors des conflits produits à la suite des élections présidentielles en Iran en 2009, malgré le contrôle imposé aux médias traditionnels par les autorités iraniennes [15].

Outre sa fonction de moyen de diffusion de l'information, la quantité gigantesque d'information publiée dans Twitter est utilisée aussi comme ressource statistique pour détecter, de manière continue, les tendances, l'humeur des gens, les opinions des consommateurs (Jansen et al., 2009a ; O'Connor et al., 2010) et même leurs convictions politiques (Tumasjan et al., 2010).

1.2 Spécificités des microblogs

Jansen et al. (2009b) ont réalisé une étude linguistique sur Twitter. Ils ont trouvé qu'un tweet contient en moyenne 15 mots. Ce chiffre est extrêmement faible comparé aux autres sources d'information du Web. Les articles de Wikipédia, par exemple,

13. `http://referencement-alsace.fr/`

14. A. Bloxham, "Facebook more effective than emergency services in a disaster," The Daily Telegraph, December 20, 2008.

15. M. Musgrove, "Twitter is a player in Iran's drama" The Washington Post, July 09, 2009.

possèdent en moyenne 320 termes par article [16]. Cette particularité représente un défi pour les techniques de recherche d'information classiques qui se basent principalement sur les fréquences des termes dans les documents.

FIGURE 2.5 – Tweet posté par @florencesantrot contenant une image et des hashtags (#Apple #iphone6cost1k). Il a été retweeté sept fois et favori une fois.

Un microbloggeur peut inclure différents types de signes dans un tweet, en plus du contenu textuel. Ces pratiques ont peu à peu évoluées pour devenir des « normes de balisage » :

- @ suivi du nom d'utilisateur permet d'indiquer qu'on mentionne ou s'adresse à une personne particulière (représenté par son compte),
- # suivi par un mot est un hashtag. Un hashtag indique un mot important que le système peut utiliser pour permettre une recherche par navigation (figure 2.5). Les hashtags permettent de catégoriser les microblogs selon un contexte (événement, lieu, etc.) : par exemple, certaines émissions télévisées définissent des hashtags spécifiques à utiliser par les microbloggeurs souhaitant exprimer leurs avis sur le sujet de l'émission. Les conférences scientifiques définissent également des hashtags permettant, d'une part, aux participants de partager leurs remarques et, d'autres part, aux gens de l'extérieur de suivre ce qui se passe dans la conférence en temps-réel.
- Les microblogs peuvent également contenir des URL. Ces hyperliens prennent

16. http://en.wikipedia.org/wiki/Wikipedia:Words_per_article

une forme courte en raison du nombre limité de caractères autorisés par microblog. Il existe deux services très connus pour créer la forme réduite des URL : bit.ly et tinyurl.com. Dans le cas où l'URL correspond à une image, Twitter affiche un aperçu de cette image dans l'interface de l'utilisateur comme le montre la figure 2.1.

– Les internautes peuvent mettre des photos dans leurs microblogs (figure 2.5). En cliquant dessus, l'utilisateur pourra voir la photo en taille normale.

Outre les données postées explicitement par les microbloggeurs, les microblogs contiennent également des méta-données de différentes natures :

– de géolocalisation : les microblogs publiés à travers les terminaux mobiles équipés de GPS fournissent des informations de géolocalisation. Ces informations permettent de localiser l'endroit duquel lequel le microblog a été publié.
– d'horodatage : chaque microblog est caractérisé par sa date de publication. Cette information est utilisée pour mesurer sa fraîcheur s'il fait partie d'une liste de résultats d'une recherche.
– d'auteur : Les plate-formes de microblogging stockent le compte depuis lequel est publié chaque microblog. Ceci permet aux utilisateurs de trouver les microblogs d'un auteur en particulier.
– de favoris : on peut savoir, pour chaque microblog, combien de fois il a été choisi dans les listes de favoris des autres utilisateurs (figure 2.5) ainsi que l'ensemble des utilisateurs qui l'ont sélectionné.
– de rediffusion : RT indique que le message est rediffusé. Le mécanisme de rediffusion permet aux utilisateurs de partager de nouveau des microblogs qu'ils trouvent intéressants parmi les microblogs publiés par leurs amis (par exemple, RT @mashable Top 10 Twitter Trends This Week http ://on.mash.to/eA2jY5). Dans Twitter, on peut connaître le nombre de fois qu'un tweet a été retweeté (figure 2.5). On peut également accéder à la liste des utilisateurs qui ont retweeté un tweet donné.

1.3 Spécificités des recherches dans les microblogs

Le moteur de recherche de microblogs est spécifique au niveau des données en entrée ou des résultats. D'une part, outre des mots-clés, un utilisateur peut mélanger des comptes utilisateurs, des hashtags et même des URLs, dans sa recherche. La figure 2.6 montre les suggestions de différents types de données de recherche de Twitter.

D'autre part, les résultats affichés diffèrent en fonction du type de données utilisées : si l'utilisateur sélectionne un compte utilisateur parmi la liste des suggestions, l'interface affichera le profil de ce compte (ses informations et ses tweets). Dans les autres cas, l'interface affichera une liste de microblogs contenant les termes, le hash-

FIGURE 2.6 – Suggestion de différents type de résultats dans le moteur de recherche de Twitter : des mots-clés, des hashtags, des comptes utilisateurs sont présentés.

tag ou l'URL recherchée. Les résultats sont présentés par défaut dans l'ordre chronologique inverse. Cependant, l'utilisateur peut choisir d'afficher tous les résultats, comme le montre la figure 2.4. Les microblogs sont alors triés selon toute probabilité de pertinence telle que leur popularité (fréquence de favoris et de retweets).

Teevan et al. (2011) ont étudié les motivations des utilisateurs pour chercher les informations sur Twitter. Ils ont également identifié les pratiques de recherche des microblogueurs. En observant les pratiques de 54 utilisateurs actifs de Twitter, ils ont constaté que les internautes cherchent dans Twitter pour avoir :

- Des informations **récentes** : 49 % des participants ont cherché des informations sur les actualités, les sujets « tendance », les événement récents, le trafic routier, les accidents du quartier. . . .
- Des information **sociales** : 26 % des participants ont cherché des informations sur d'autres utilisateurs, tels que ceux qui ont des intérêts similaires, ou même ce que dit un utilisateur en particulier.
- Des information sur des sujets, qui s'apparentent aux recherches souvent effectuées sur les moteurs de recherche du Web. 36 % des participants ont cherché des sujets spécifiques.

Les auteurs ont également analysé les logs de moteurs de recherche pour identifier

les différences entre les recherches effectuées sur Twitter et celles effectuées sur les moteurs de recherche du Web. Les différences se manifestent à plusieurs niveaux :

- au niveau des requêtes (Twitter/Web) : sur la longueur des requêtes (1,6/3 mots), sur la présence de noms de célébrités (15 %/3 %), ou de « # » (21 %/0,1 %).
- au niveau de l'importance des requêtes : en moyenne, chaque requête est soumise 2 fois sur le web, et 3 fois dans Twitter. Ceci peut être dû aux tendances présentées par la plate-forme sous forme de liens permettant d'obtenir les tweets récents sur les sujet tendances.
- au niveau des sessions de recherches de Twitter qui sont plus courtes que celles réalisées sur le Web, que ce soit sur le temps ou sur le nombre de requêtes. Dans Twitter, une session consiste souvent en la surveillance des tweets sur une requête particulière, en actualisant les résultats sur une période de temps. En d'autre termes, les utilisateurs ont tendance à actualiser les résultats pour avoir l'information récente, sans attendre les notifications de la plate-forme.

Pour conclure, les plate-formes de *microblogging* (Twitter en particulier), représentent un nouveau type de source d'information en pleine évolution grâce à un ensemble de caractéristiques spécifiques :

- de fonctionnalité, telles que le partage d'information temps-réel, les abonnements sans restriction, etc. Ces nouvelles fonctionnalités ont popularisé de nouvelles pratiques comme le suivi de l'actualité de célébrités, la réalisation de campagnes électorales, l'analyse de l'humeur et des avis des gens en temps-réel, la participation à distance à des conférences, etc.
- de forme, telles que la faible longueur des messages, l'utilisation du jargon du net, une syntaxe spécifique, etc.

La quantité et la nature des tweets ont suscité de nouveaux usages tant de la part des individus que des organisations. La section suivante synthétise les travaux de littérature traitant de l'accès à l'information dans les microblogs.

2 Accès à l'information dans les microblogs

Dans ce paragraphe, nous listons les travaux de l'état de l'art sur la problématique de l'accès à l'information via Twitter. Nous classons ces travaux en fonction du type d'information recherché.

2.1 Recherche temps-réel de microblogs

Pour cette tâche, l'utilisateur souhaite obtenir de l'information pertinente la plus fraîche possible vis-à-vis d'un besoin en information (Ounis et al., 2011). Générale-

ment, un certain temps s'écoule avant que cette information soit disponible sur le web et qu'elle soit indexée par les moteurs de recherche (Dong, Zhang, et al., 2010). Dans la RI temps-réel, la date de publication d'un document est considérée comme un facteur de pertinence très important, si ce n'est pas le plus pertinent. Une interprétation possible de cette tâche consiste à trier anti-chronologiquement tous les documents publiés avant la date de soumission de la requête, et ensuite, à écarter les documents non pertinents (Ounis et al., 2011). La tâche se réduit donc à l'identification des caractéristiques des documents pertinents à restituer. Plusieurs travaux ont proposé des critères utilisés comme facteurs de pertinence supplémentaires à la pertinence textuelle : la fraîcheur (Magnani et al., 2012 ; Vosecky et al., 2012), la popularité de l'auteur (Zhao et al., 2011 ; Massoudi et al., 2011), la présence d'URLs (Vosecky et al., 2012)... Des études empiriques ont montré que ces critères reflètent la pertinence lorsqu'ils sont employés en plus de la pertinence textuelle (Damak et al., 2013). Nous présenterons dans la section suivante un état de l'art des différentes approches de recherche de microblogs et des approches qui ont utilisé les critères de pertinence supplémentaires. Nous détaillerons également les différentes manières avec lesquelles ces critères de pertinences ont été employés.

2.2 Recherche de microbloggeurs

La recherche de microbloggeurs s'apparente à la tâche de recherche d'experts de la RI classique. Les objectifs sont l'identification des utilisateurs les plus populaires, ceux qui ont les mêmes centres d'intérêts que l'utilisateur courant, ou bien les experts dans des domaines spécifiques.

Plusieurs travaux se sont focalisés sur l'identification des utilisateurs les plus populaires dans les plate-formes de microblogging. Ils se basent sur des méthodes telles que la centralité calculée au travers du graphe social. À titre d'exemple, TwitterRank (Weng et al., 2010) est une approche inspirée de l'algorithme PageRank (Brin et Page, 1998) qui mesure l'influence des utilisateurs sur Twitter. Le score de chaque utilisateur est mesuré en fonction des scores de ses abonnés. Cette approche prend en compte les similarités des sujets discutés entre les utilisateurs, ainsi que la structure des liens d'abonnements. Ben Jabeur, Tamine, et Boughanem (2012) ont mesuré la popularité d'un auteur en proposant un algorithme semblable à PageRank. Cet algorithme mesure la popularité d'un auteur dans un réseau formé par les retweets, les mentions et les réponses. Tunkelang [17] a proposé un modèle qui se base également sur l'algorithme PageRank. Cependant, il a introduit le facteur de renvoi des

17. http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank/

messages par les abonnés d'un utilisateur :

$$Popularité(u) = \sum_{v \in abonnés(u)} \frac{1 + p * popularité(v)}{||Abonnements(v)||} \tag{2.1}$$

avec $abonnés(u)$ est l'ensemble des utilisateurs abonnés à u, $Abonnements(v)$ est l'ensemble des utilisateurs auxquels v est abonné et p est la probabilité que l'utilisateur v va retweeter les tweets de u.

D'autres approches mesurent l'importance des utilisateurs autrement. En analysant les habitudes de diffusion d'information dans Twitter, Lee et al. (2010) ont découvert que la diffusion d'information atteint son maximum à son apparition. Le pic dans le taux de publication est observé au moment de l'apparition de l'information. Ensuite, ce taux diminue en avançant dans le temps. Par conséquent, ils ont proposé une approche considérant l'ordre temporel de diffusion de l'information pour détecter le meilleur diffuseur d'information. L'utilisateur le plus important est celui qui diffuse souvent les informations en premier.

La majorité des approches proposées prend en compte un ordonnancement statique de l'importance des utilisateurs. Cependant, Cappelletti et Sastry (2012) considèrent que, dans un environnement temps-réel, l'importance d'un utilisateur doit être évolutive. Ainsi, ils ont proposé un modèle qui se base sur le potentiel d'un utilisateur à amplifier la diffusion d'une information. Cette importance varie avec l'évolution du réseau social de l'utilisateur. Un utilisateur est d'autant plus important que l'information qu'il partage est susceptible d'atteindre un grand nombre d'utilisateurs. Concrètement, ceci est calculé en fonction de deux facteurs : le premier calcule à quel degré un utilisateur peut être retweeté ou cité par ses abonnés et le deuxième mesure la taille de l'audience des retweets et des citations de l'utilisateur.

2.3 Détection d'opinions

La détection d'opinion a été souvent étudiée en recherche d'information, particulièrement dans la recherche de blogs (Pang et Lee, 2008 ; Missen et al., 2009). L'objectif est de retrouver les documents exprimant des opinions sur le sujet de la requête. La majorité des approches proposées se basent sur des ressources lexicales comportant les termes d'opinions telles qu'*opinionFinder* (Wilson et al., 2005), *General Inquirer lexicon* (Hatzivassiloglou et McKeown, 1997) ou *SentiWordnet* (Baccianella et Sebastiani, 2010). La plupart des approches spécifient des critères (présence de termes et leurs fréquences, *Parts of speech*, de syntaxe, de négation...) et exploitent des techniques d'apprentissage automatique. Les mêmes principes ont été ainsi utilisées sur les microblogs.

Comme les blogs, les microblogs expriment des opinions (Jansen et al., 2009a). Shamma et al. (2009) ont montré que la plupart des tweets ont un ton négatif, et

que les microblogs permettent d'obtenir des opinions immédiates et des réactions sur des produits. Ils ont trouvé également que les tweets peuvent être utilisés pour annoter les débats politiques avec les opinions des téléspectateurs. Plus précisément, ils ont constaté que le taux de messages contenant des opinions dans Twitter peut servir comme un prédicateur de l'évolution des sujets dans l'événement médiatisé.

Bollen et al. (2009) ont modélisé les phénomènes socio-économiques à travers l'analyse des opinions dans les tweets. Leur liste de phénomènes est composée de vingt événements de la vie quotidienne, en intégrant le comportement des marchés boursiers correspondant à l'indice Dow Jones Industrial Average et les indices des prix du pétrole de West Texas Intermediate. Ils ont trouvé que l'humeur globale des gens est corrélée avec ces événements. Par exemple, à Thanksgiving, il y avait plutôt une humeur de joie et rarement des sentiments de fatigue. Durant les élections présidentielles aux États Unis, il y avait au début beaucoup de doute avant les élections (sentiments de confusion et de dépression), suivi de sentiments de joie et de bonheur après la publication des résultats. Un dernier exemple est celui de la baisse de l'indice de Dow Jones qui coïncide avec les sentiments de dépression.

2.4 Classification thématique des microblogs

L'objectif de la classification thématique de microblogs est de créer des filtres thématiques sur les flux d'information. Ceci est réalisé en identifiant les sujets discutés dans les microblogs. La classification thématique des microblogs nous permettra, par extension, de classer les utilisateurs en fonction de leurs centres d'intérêts.

Une première solution pour ce type de problème est de regrouper les *microblogs* en fonction des hashtags qu'ils contiennent (Efron, 2010). Ramage et al. (2010) ont utilisé une implémentation étiquetée et évolutive de Latent Dirichlet Allocation (Blei et al., 2003) afin d'extraire les tags et de les utiliser pour caractériser les utilisateurs et les *microblogs*. Song et al. (2010) se sont basés sur des informations spatio-temporelles afin d'identifier des tags corrélés. Ces tags sont utilisés par la suite pour regrouper les tweets et les classifier. Enfin, Bernstein et al. (2010) ont proposé un algorithme pour détecter précisément les sujets des *microblogs*. Ce dernier consiste à détecter les entités nommées dans un microblog et les soumettre à un moteur de recherche. Le sujet du microblog correspondra alors au terme le plus important dans les résultats, calculé à travers un algorithme de pondération ($TFIDF$ (Robertson, 2004)).

2.5 Détection de tendances

La détection de tendances vise à identifier automatiquement les thèmes émergeant qui apparaissent dans le flux de microblogs en temps-réel (R. Li et al., 2012).

Les tendances sont généralement des événements émergents, les dernières nouvelles et les sujets qui attirent l'attention des utilisateurs. La détection des tendances revêt donc une grande utilité pour les journalistes et les analystes, car elle leur permet d'être rapidement actifs sur les sujets « tendances ». Par exemple, lors de l'annonce de la mort de Michael Jackson le 25 juin 2009, Twitter a été immédiatement inondé par un énorme volume de commentaires. La détection de tendances est également importante pour les professionnels du marketing en ligne et les sociétés de suivi d'opinion, puisque les tendances indiquent des sujets qui captent l'attention du public. Plusieurs applications ont été développées pour détecter les tendances à partir de Twitter : Trendsmap [18], What The Trend [19], Twinitor [20] et Twendr [21]. D'autres travaux ont même utilisé Twitter comme un système préventif aux catastrophes. Par exemple, Sakaki et al. (2010) se sont basés sur Twitter pour créer un système d'avertissement des tremblements de terre et Lampos et Cristianini (2010) ont utilisé les tweets pour suivre la propagation des épidémies.

3 Recherche adhoc de microblogs

Le principe de la recherche adhoc de microblogs est similaire à la RI adhoc classique. Il s'agit de répondre à une requête via un index de microblogs et sélectionner ceux qui sont pertinents (Efron, 2011a). La différence entre la RI adhoc dans les tweets et la RI adhoc dans les documents du Web réside dans la nature de l'information traitée et des sessions de recherches. Ces différences sont principalement dues aux spécificités des microblogs par rapport aux autres sources d'information et les motivations des utilisateurs pour chercher dans cette source d'information.

Efron (2011a) a posé la question : *quels sont les facteurs reflétant la pertinence dans la recherche de microblogs ? Les facteurs tels que la popularité de l'auteur et l'horodatage ont probablement leur importance pour juger l'utilité d'un microblog par rapport à un autre. Cependant, la manière de considérer ces qualités n'est pas évidente.*

Ainsi, il existe plusieurs facteurs de pertinence à prendre en compte dans la conception des approches de recherche de microblogs, en plus de la pertinence textuelle : facteurs sociaux, facteurs de popularité des auteurs, facteurs de fraîcheur, facteurs liées aux URLs... Nous présentons dans cette section les principaux facteurs de pertinence employés dans la recherche de microblogs ainsi que leurs différents objectifs.

18. `http://trendsmap.com/`
19. `http://whatthetrend.com/`
20. `http://twinitor.com/`
21. `http://twendr.com/`

3.1 Facteur de pertinence textuelle

Le problème principal de la pertinence textuelle dans la recherche de microblogs réside dans leur faible longueur. Les modèles de RI classiques qui, de manière générale, se basent sur des facteurs tels que la fréquence des termes dans les documents et la longueur des documents, sont limités par la faible longueur des microblogs, où les termes n'apparaissent pas plus d'une fois.

La majorité des approches de RI dans les microblogs ne tiennent ainsi pas compte des facteurs de normalisation et de fréquence utilisés dans les modèles de RI classique : par exemple Che Alhadi et al. (2011) emploient le modèle vectoriel en éliminant le facteur de la normalisation de la longueur. Massoudi et al. (2011) de leur coté utilisent uniquement la présence ou l'absence du terme dans le modèle de langue (LM) à la place de sa fréquence dans le document.

Ferguson et al. (2012) ont étudié l'impact des fréquences et leur normalisation dans la mesure de la pertinence avec le modèle BM25. Ils ont trouvé que ces facteurs sont non seulement inefficaces, mais dégradent aussi les résultats d'une tâche de recherche de microblogs.

Certains travaux ont proposé des méthodes plus sophistiquées pour résoudre le problème de fréquences et de normalisation. Lin et al. (2012) emploient une méthode qui se base sur la co-occurrence des termes. Ils construisent un graphe pondéré dont les nœuds représentent les termes et les liens représentent leurs co-occurrences dans les tweets de la collection. Ainsi, le score de chaque terme de la requête dans un microblog est calculé en fonction des poids des liens de ce terme avec les termes du tweet.

Au lieu d'ignorer les facteurs de fréquences, certaines approches ont essayé d'améliorer la représentation des termes, que ce soit des requêtes ou des microblogs afin de réduire l'effet de leur faible taille. Plusieurs techniques d'expansion de requêtes ont été proposées. Kumar et Carterette (2013) ont étendu les requêtes avec les termes les plus fréquents dans les résultats de la requête initiale. D'autres approches ont exploitées des critères temporels dans le choix des termes d'extension (Efron, 2011b ; Miyanishi et al., 2013). Du côté des microblogs, Efron et al. (2012) ont proposé deux approches pour améliorer leur représentation. La première consiste à enrichir chaque microblog dans l'index avec les microblogs ayant des contenus similaires. La deuxième approche exploite les microblogs similaires à chaque microblog pour lui créer un profil temporel. Ce profil sera utilisé au moment de la restitution des résultats. McCreadie et Macdonald (2013) et Ben Jabeur et al. (2013), quant à eux, ont fusionné le contenu du microblog avec le contenu de l'URL, s'il existe.

3.2 Facteur de pertinence social

Étant donné que le microblogging est une forme de réseau social, il est ainsi possible de traiter le problème de tri des microblogs en exploitant un critère particulier qui n'est pas (aussi facilement) disponible dans la recherche sur le Web traditionnel, à savoir le réseau social sous-jacent aux plate-formes. Cette catégorie d'approches considère que la pertinence est liée à la crédibilité de la source d'information.

La plupart des approches exploitant le réseau social ont défini des critères de pertinence reflétant l'importance des utilisateurs. 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... Si certains travaux ont combiné ces critères linéairement (Nagmoti et al., 2010 ; Zhao et al., 2011 ; Damak et al., 2011), d'autres ont utilisé des techniques d'apprentissage : SVM (Joachims, 2005) et Linear Regression dans l'approche de Duan et al. (2010) et RankSVM dans l'approche de Cheng et al. (2013).

Dans une deuxième catégorie d'approches, des graphes représentant les liens sociaux ont été générés à partir des plate-formes. Ces graphes représentent différents types de liens comme le montre la figure 2.7 : utilisateur × utilisateur et dans ce cas les liens sont les relations d'amitiés (abonnements ou abonnés ou citation), utilisateur × tweet et dans ce cas les liens représentent les statuts des utilisateurs, tweet × tweet et dans ce cas les liens représentent les retweets... L'approche présentée dans (Yamaguchi et al., 2010) utilise, par exemple, l'algorithme PageRank (Brin et Page, 1998) pour mesurer l'importance d'un microbloggeur dans un graphe composé par les utilisateurs et les tweets. Jabeur et al. (2012) 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. Ravikumar et al. (2012), quant à eux, représentent les URLs publiées dans les tweets et les liens d'hypertextualité entre elles en plus des tweets et des utilisateurs.

Les approches de la deuxième catégorie ont exploité des liens sociaux, de tout genre, mais en relation avec le tweet lui même. Une troisième catégorie d'approches exploite les informations sociales de celui qui cherche l'information en plus des informations sociales reliées aux tweets. L'idée ainsi est de comparer les informations sociales des deux côtés afin de restituer des résultats personnalisés. Uysal et Croft (2011) ont mesuré la distance entre l'auteur du tweet et le chercheur d'information à travers plusieurs critères tels que : l'existence d'une relation directe entre eux, l'existence d'un retweet ou d'une citation de l'un à l'autre, l'emploi de mêmes hashtags, la publication de mêmes URLs dans leurs tweets... C'est le principe aussi de l'approche proposée dans (Feng et Wang, 2013). Les auteurs ont utilisé des critères comme la similarité entre les abonnés de l'auteur du tweet et celui qui le cherche, puis la similarité entre la date de la dernière interaction entre eux.

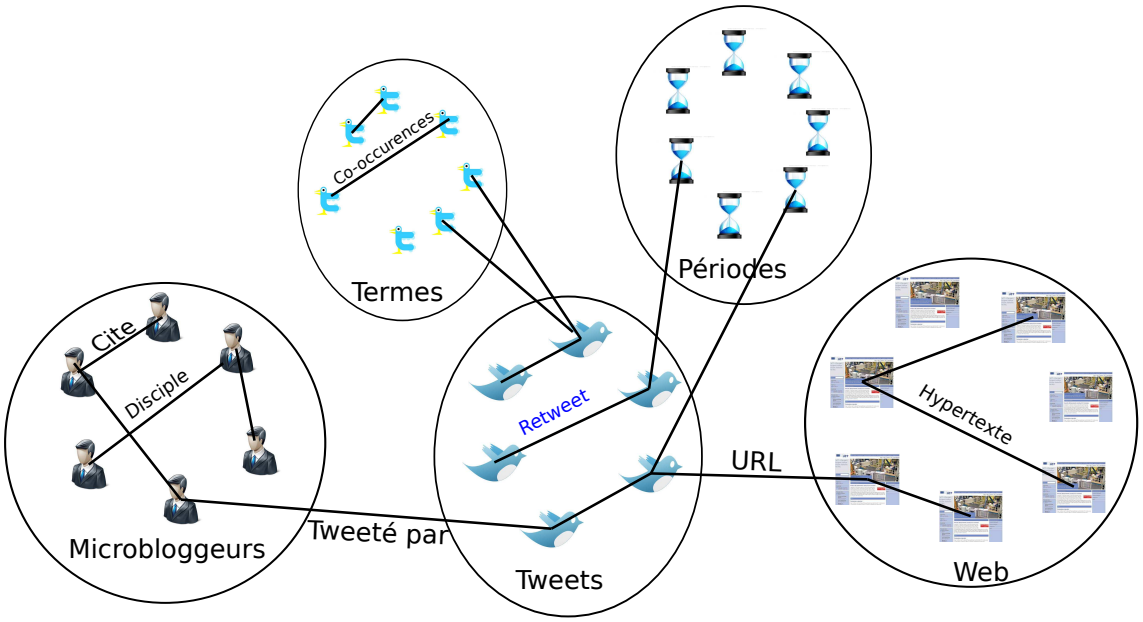


FIGURE 2.7 – Les réseaux constituables à partir des données de Twitter

Les intuitions diffèrent d'une méthode à une autre. Cependant, peu de travaux ont essayé de comparer les différentes approches. Kwak et al. (2010) ont comparé deux approches pour mesurer l'importance des utilisateurs. Dans la première approche, ils ont appliqué l'algorithme PageRank sur le réseau composé par les liens d'amitié. La deuxième approche estime l'importance d'un utilisateur en fonction de la fréquence des messages d'un utilisateur retweetés par ces abonnés. L'analyse a montré un désaccord total entre les résultats de ces deux approches, ce qui montre que la définition de l'importance d'un utilisateur, dans le cas des microblogs, nécessite encore beaucoup de recherche.

3.3 Facteur de pertinence temporelle

Pour les requêtes qui concernent les événements et les « buzz », il est crucial de prendre en compte la fraîcheur des résultats dans la mesure pertinence. Le microblogging, système temps-réel par excellence, incite les utilisateurs à exprimer leurs opinions et discuter en temps-réel. Ainsi, la prise en compte du temps est primordiale dans la recherche de microblogs.

La caractéristique temporelle des microblogs a été employée de différentes manières et avec différentes intuitions. Les premiers travaux ont tout simplement essayé de favoriser les microblogs récents. Plus un microblog est proche de la requête, plus il est susceptible d'être pertinent. Cette intuition est concrétisée en calculant un score de fraîcheur du document, en termes de différence temporelle entre la date de la soumission de la requête et la date de publication du document. Ce score est ensuite intégré dans le modèle de recherche (Massoudi et al., 2011) ou bien utilisé

comme un attribut parmi d'autres dans un modèle d'apprentissage (Cheng et al., 2013).

Le facteur temporel a été employé également dans les modèles de RI classiques. Par exemple, Efron et Golovchinsky (2011) se sont basés sur les microblogs récents pour lisser les scores dans le modèle de langue : le degré de lissage des résultats les plus éloignés temporellement était plus élevé, afin de favoriser les résultats récents. Enfin, les résultats récents ont été utilisés pour sélectionner les microblogs représentant le modèle de pertinence (Efron et Golovchinsky, 2011 ; Kumar et Carterette, 2013).

Outre pour le calcul de pertinence, le temps a été employé dans l'extension des requêtes. Plus précisément, il est utilisé pour sélectionner le fragment de temps contenant les documents susceptibles d'avoir des termes utiles pour l'extension. La meilleure période contenant les documents les plus prometteurs pour l'extension a été choisie en fonction du taux de tweets publiés sur le sujet de la requête (Efron, 2011b), du taux des tweets retweetés sur le sujet de la requête (Choi et Croft, 2012), ou bien des tweets qui sont proches temporellement de la requête (Massoudi et al., 2011). D'autres travaux sont allés plus loin en analysant les variations temporelles dans la liste des résultats initiaux (Miyanishi et al., 2013). Ainsi, le nombre de résultats utilisés pour étendre les requêtes de chaque période est proportionnel au taux de tweets publiés dans cette période sur le sujet de la requête.

Finalement, le temps a été utilisé pour enrichir la représentation et extraire de l'information supplémentaire des microblogs et des requêtes. En considérant les dates de publication des microblogs similaires à un microblog, Efron et al. (2012) génèrent le profil temporel d'un microblog. Ce profil permet de mesurer l'implication du microblog à des événements qui ont été discutés à différents moments dans le temps. Ensuite, cette mesure va être comparée avec la distribution temporelle de la requête en tant que mesure de similarité.

3.4 Facteur de pertinence d'hypertextualité

Les microbloggeurs peuvent partager plusieurs URLs dans leurs microblogs. En fait, les microbloggeurs partagent également des URLs dans leurs statuts pour attirer l'attention de leurs amis sur une nouvelle information contenue dans une page web, souvent pas encore indexée par les moteurs de recherche classiques. Ces pages représentent ainsi de l'information enrichissante par rapport au seul contenu du tweet.

Les URLs ont souvent été utilisées dans la restitution des microblogs en réponse à une requête. La finalité est d'améliorer la qualité des résultats, certes, mais les manières d'intégrer ces URLs diffèrent d'une approche à une autre. Les URLs ont été employées dans un premier temps comme facteur de pertinence. En réalité, c'est

leur présence qui reflète la pertinence du tweet pour (Cheng et al., 2013). D'autres approches ont raffiné le critère en calculant la fréquence plutôt que la présence d'une URL (Zhao et al., 2011 ; Duan et al., 2010). Ces deux critères ont été employés avec d'autres, que ce soit dans des combinaisons linéaires ou dans des algorithmes d'apprentissage. Malgré leur simplicité, ces critères ont montré un fort impact dans l'amélioration de la qualité des résultats (Damak et al., 2013).

D'autre part, les URLs ont été utilisées comme des éléments parmi d'autres pour caractériser l'écosystème des plate-formes de microblogging. Le réseau formé par ces éléments est utilisé pour mesurer la centralité des tweets, ainsi que leur fiabilité (Ravikumar et al., 2012).

Enfin, le contenu des URLs est utilisé pour enrichir le vocabulaire des tweets, limités en longueur. Certaines approches ont utilisé le contenu dans la définition du modèle du document avec le modèle de langue (Zhongyuan et al., 2012). D'autres, comme McCreadie et Macdonald (2013), ont représenté chaque microblog comme une composition multidimensionnelle dont les dimensions sont le contenu du microblog et le contenu des URLs si elles existent.... Généralement, quelle que soit la manière avec laquelle les URLs sont exploitées, elles améliorent remarquablement la qualité des résultats.

3.5 Autres facteurs de pertinence

D'autres facteurs peuvent être utilisés pour la recherche de microblogs. Les facteurs de qualité des microblogs sont indépendants de la requête. Avec les particularité des microblogs (qualité du langage, longueur faible ...), ces critères sont essentiels pour estimer la qualité d'un microblog. Voici les critères les plus populaires dans la littérature :

- **Longueur du microblog** : nombre de termes dans le microblog. La longueur d'une phrase reflète la quantité d'information qu'elle véhicule (Zhao et al., 2011 ; Magnani et al., 2012 ; Metzler et Cai, 2011 ; Duan et al., 2010).
- **Fréquence de Retweets** : nombre de fois qu'un tweet a été retweeté. Si un utilisateur repartage un tweet, alors ceci suggère qu'il a trouvé son contenu intéressant (Zhao et al., 2011 ; Magnani et al., 2012 ; Vosecky et al., 2012 ; Duan et al., 2010).
- **Fréquence de hashtags** : nombre de hashtags dans un tweet. Les hashtags sont utilisés pour définir un topic pour le tweet, ou bien pour s'intégrer à une conversation (Duan et al., 2010).
- **Réponse** : indique que le microblog est une réponse à un autre. Ceci montre qu'il ne s'agit pas d'un message isolé et sans interaction (Vosecky et al., 2012 ; Metzler et Cai, 2011 ; Duan et al., 2010).
- **Qualité du langage** : les microbloggeurs ne font pas en général très attention

en écrivant les tweets. Il peuvent également abréger certains mots à cause de la contrainte liée à la longueur restreinte des tweets. Ce critère calcule le ratio des termes correctement orthographiés par rapport à tous les termes du microblogs (Metzler et Cai, 2011). Han et Baldwin (2011) ont proposé d'améliorer la qualité des microblogs en corrigeant les termes mal-orthographiés. Ils tiennent compte du contexte du tweet pour proposer les corrections convenables pour les termes erronés.

– **Sentiment** : les microblogs reflétant des sentiments sont pertinents lorsqu'un utilisateur cherche des avis sur des produits ou des événements. Ce critère est mesuré en calculant le ratio des termes exprimant des sentiments par rapport à la longueur du tweet (Cheng et al., 2013).

3.6 Bilan

Le tableau 2.2 résume la majorité des critères de pertinence que nous venons de décrire, souvent utilisés en complément de la pertinence textuelle. Certaines approches qui les emploient les combinent linéairement (Zhao et al., 2011 ; Massoudi et al., 2011). D'autres approches ont employé des techniques d'apprentissage pour les considérer dans la restitution (Duan et al., 2010 ; Cheng et al., 2013 ; Uysal et Croft, 2011).

4 Évaluation de la RI dans les microblogs

Comme nous l'avons vu au chapitre 1, l'évaluation en RI se fait principalement à travers les collections de tests, souvent construites dans le cadre de campagnes d'évaluation. La RI dans les microblogs ne déroge pas à cette règle, avec la mise en place de la tâche Microblog dans la campagne d'évaluation TREC.

4.1 La tâche TREC Microblog

Il s'agit, pour un moteur de recherche, de fournir les tweets dont le contenu satisfait un besoin en information exprimé sous forme de mots clés (tâche adhoc). Les systèmes proposés doivent retrouver les résultats pertinents, mais aussi les plus récents, par rapport à la date de soumission de la requête (*real-time retrieval*). Les résultats doivent être publiés avant la date de la soumission de la requête. Depuis 2011, trois versions de cette tâche ont été mises en œuvre (2011, 2012 et 2013).

La collection de test Tweets2011 comprend :

– 16 millions de tweets (0,5 Go) exprimés dans diverses langues et publiés sur Twitter entre le 23 janvier 2011 et le 8 février 2011,

Tableau 2.2 – Critères de pertinence

Critère	Références
Popularité du tweet dans la liste de résultats	(Duan et al., 2010 ; Ben Jabeur, Damak, et al., 2012)
Nombre de termes en commun entre le tweet et la requête	(Damak et al., 2011)
Nombre de fois que le tweet à été retweeté	(Zhao et al., 2011 ; Magnani et al., 2012 ; Vosecky et al., 2012 ; Duan et al., 2010)
Nombre de hashtags dans le tweet	(Duan et al., 2010)
Présence de hashtags dans le tweet	(Vosecky et al., 2012 ; Metzler et Cai, 2011)
Popularité des hashtags dans la collection	(Vosecky et al., 2012)
Longueur du tweet	(Zhao et al., 2011 ; Magnani et al., 2012 ; Metzler et Cai, 2011 ; Duan et al., 2010)
Présence d'URLs dans le tweet	(Vosecky et al., 2012 ; Massoudi et al., 2011 ; Metzler et Cai, 2011 ; Duan et al., 2010)
Nombre D'URLs dans le tweet	(Zhao et al., 2011)
Popularité de l'URL dans la collection	(Vosecky et al., 2012)
Le tweet est-il une réponse ?	(Vosecky et al., 2012 ; Metzler et Cai, 2011 ; Duan et al., 2010)
Nombre de tweets de l'auteur	(Zhao et al., 2011)
Nombre d'abonnés de l'auteur	(Magnani et al., 2012 ; Massoudi et al., 2011 ; Duan et al., 2010 ; Zhao et al., 2011)
Nombre de mentions de l'auteur	(Vosecky et al., 2012 ; Duan et al., 2010)
Différence temporelle entre le tweet et la requête	(Magnani et al., 2012 ; Vosecky et al., 2012 ; Metzler et Cai, 2011)
Qualité du language du tweet	(Metzler et Cai, 2011)
Sentiment positif/négatif dans le tweet	(Cheng et al., 2013)

– 49 *topics* dont on trouvera un exemple en figure 2.8. La balise title décrit le besoin exprimé à un moment donné (querytime). Ce moment correspond concrètement à la date de publication du tweet le plus récent de la requête,

– les jugements de pertinence (*qrels*) associées aux 49 *topics*. La pertinence de chaque tweet est ternaire : non pertinent, moyennement pertinent et hautement pertinent. Tout tweet exprimé dans une langue autre que l'anglais est non pertinent. Il en est de même pour les retweets et les tweets identifiés comme *spam* par les assesseurs.

```
<top>
<num>  Number: MB007</num>
<title>  Pakistan diplomat arrest murder</title>
<querytime>  Tue Feb 08 22:56:33 +0000 2011</querytime>
<querytweettime>  35109758973255680 </querytweettime>
</top>
```

FIGURE 2.8 – Exemple de *topic* pour la tâche Microblog

La collection de test Tweets2012 comprend :
– le même corpus de tweets que celui de 2011,
– 60 nouvelles requêtes avec leurs jugements de pertinence. Seuls les tweets hautement pertinents ont été considérés dans l'évaluation des systèmes.

La collection de test Tweets2013 comprend :
– une nouvelle collection de 240 millions de tweets (70 Go), publiés dans la période du 1er février 2013 au 31 mars 2013. Cette collection est accessible uniquement à travers une API (contrairement à l'ancienne collection).
– 60 nouvelles requêtes avec les jugements de pertinence associés.

En 2012, une deuxième évaluation a été introduite, *real-time filtering*. L'objectif est d'évaluer la capacité des systèmes à indexer le flux des tweets en temps reél et d'en extraire les tweets pertinents pour un besoin en information. Cette tâche n'entrant pas dans notre problématique de recherche, nous ne la détaillons pas davantage.

4.2 Discussion sur les mesures d'évaluation

De façon usuelle, les moteurs de recherche trient les résultats en fonction du score de pertinence des documents. Ce n'est pas le cas dans la tâche Microblog de TREC, qui promeut la recherche temps réel (*real-time search*). Cela se traduit par une préférence pour les tweets les plus proches temporellement de la requête. Au niveau de la procédure d'évaluation en 2011, cette contrainte est mise en œuvre en réordonnant les résultats (*runs*) d'un moteur de recherche en fonction de l'attribut querytweettime des tweets (le champ sim – score de similarité – du *run* est recalculé en fonction). Cette prise en compte a suscité une ambiguïté dans l'interprétation des scores des participants : il n'y a pas de moyen pour identifier les systèmes qui ont considéré la fraîcheur dans la mesure de pertinence. Nous notons, à titre indicatif, que les meilleurs systèmes de cette édition sont les systèmes qui se basent sur la

pertinence textuelle en réalisant une coupure (*cut-off*) agressive (càd. X tweets). Cette prise en compte temporelle a été écarté à partir de l'édition de 2012.

Deux mesures officielles ont été considérées dans les trois versions de la tâche : la précision à 30 documents (P@30) et la précision moyenne (AP). Notons que ces mesures ont été calculées en considérant tous les tweets pertinents (*all-rel*) en 2011 et 2013 ou uniquement les tweets hautement pertinents (*high-rel*) en 2012. Les valeurs de ces mesures, pour chaque requête, sont moyennées pour obtenir le score global d'un système (P@30 moyennée et MAP). Le classement des systèmes a été réalisé sur la P@30 moyennée, la MAP étant uniquement donnée à titre indicatif. En 2012, les courbes ROC ont été également données à titre indicatif. Notons également que les systèmes ont des caractéristiques différentes : intervention manuelle ou pas (run automatique), utilisation de sources externes ou pas, utilisation de sources futures (dont la publication est postérieure à la date de la requête) ou pas. Bien évidemment, les résultats sont à apprécier en groupant au préalable les systèmes possédant des caractéristiques similaires.

5 Bilan et limites de l'état de l'art

Le microblogging est une nouvelle source d'information en pleine croissance, fortement exploitée par les utilisateurs pour partager et trouver de l'information. Plusieurs chercheurs se sont focalisés sur l'accès à l'information à partir de cette source. Les travaux réalisés extraient différents types d'informations (personnes, tendance, opinion...). Dans cette thèse, nous nous concentrons uniquement sur la recherche adhoc de microblogs. Pour ce type d'information, plusieurs approches avec différentes intuitions ont été proposées. La grande majorité des travaux ont défini des facteurs de pertinence supplémentaires par rapport à celui de la seule pertinence textuelle. Cependant, les chercheurs n'ont pas examiné de près les problèmes des approches de la RI classique.

C'est pourquoi, dans nos travaux, nous avons commencé dans un premier temps par (i) la réalisation d'une analyse de défaillance des modèles de RI classiques afin d'identifier les facteurs principaux limitant leur efficacité sur ce type de contenu (chapitre 3). Nous avons trouvé que la majorité des problèmes sont dus au vocabulaire limité induit par la faible longueur des tweets. C'est pourquoi (ii) nous avons compensé ce problème en appliquant des techniques d'expansion de requêtes et de microblogs (chapitre 4).

Nous avons montré dans l'état de l'art que la majorité des approches emploient une multitude de facteurs de pertinence en plus de la pertinence textuelle. Cependant, peu de travaux ont essayé d'évaluer leurs impact réel dans la restitution. Nous avons ainsi (iii) réalisé une étude des critères souvent utilisés dans les travaux afin

de déterminer ceux qui reflètent vraiment la pertinence (chapitre 5). Par définition, la recherche d'information dans les microblogs implique automatiquement la prise en compte de la fraîcheur dans la mesure de la pertinence. Le dernier chapitre de notre contribution (chapitre 6) (iv) traite particulièrement ce facteur et l'impact de son emploi sur la qualité des résultats.

Contribution

Chapitre 3

Analyse de défaillance des modèles de RI classique sur les microblogs

1 Introduction

La majorité des approches présentées dans la littérature pour la recherche de microblogs emploient différents facteurs de pertinence en plus de la pertinence textuelle comme, par exemple, la popularité de l'auteur du microblog, la qualité du langage utilisé, la fraîcheur, etc. Toutefois, la pertinence textuelle est toujours considérée comme le facteur principal de pertinence. Cette pertinence textuelle est généralement calculée avec des modèles de RI classiques (Ounis et al., 2011, 2012). Ces modèles se basent principalement sur les fréquences des termes et les longueurs des documents (modèles sac de mots). Cependant, dans le cas des microblogs, le nombre de termes par microblog est en moyenne égal à 15 et chaque terme n'apparaît qu'une seule fois.

Dans ce chapitre, nous présentons une analyse de défaillance réalisée pour déterminer le comportement des modèles de RI classiques sur les microblogs. Les observations tirées de cette analyse nous permettront d'identifier les pistes à exploiter pour gérer cette forme de contenu de façon plus pertinente.

2 Méthodologie

Notre analyse a pour but de déterminer les facteurs pénalisant les modèles de RI classiques dans la restitution de microblogs. Pour ce faire, nous avons analysé les microblogs pertinents mais non restitués avec un modèle de RI classique. Nous nous sommes basés, dans notre analyse, sur la collection TREC Microblog et sur les requêtes des tâches de 2011 et 2012. Nous avons employé le modèle vectoriel comme modèle de RI classique, et ce pour deux raisons : d'une part, ce modèle est souvent utilisé en RI et a toujours prouvé son efficacité (Baeza-Yates et Ribeiro-Neto, 1999).

En outre, ce modèle est considéré comme baseline dans les éditions 2011 et 2012 de la tâche Microblog de TREC.

La question de recherche liée à cette analyse est la suivante : les facteurs limitant les modèles de recherche classiques sont-ils dus :

- à la taille réduite des microblogs ?
- au vocabulaire limité des microblogs ?
- à la syntaxe (@mention et #hashtag) fréquemment utilisée dans les microblogs ?
- à la qualité du langage utilisé par les utilisateurs ?

Dans un deuxième temps, nous avons examiné le contenu pointé par les URLs accompagnant les tweets. L'objectif est d'avoir une idée de l'impact de leur prise en compte dans la restitution et de leur potentiel d'enrichissement du contenu des tweets.

3 Expérimentations

3.1 Cadre expérimental

Nos expérimentations ont reposé sur le moteur de recherche *open source* Lucene [1], qui implémente une version modifiée du modèle vectoriel présentée dans (Cohen et al., 2007). La version de Lucene que nous utilisons intègre le lemmatiseur Porter (1980) et utilise une liste de mots vides. Nous avons modifié cette version de sorte que la recherche ne tienne compte que des tweets publiés avant le `querytime` de chaque topic, que ce soit dans la mesure de la pertinence ou bien au niveau de la restitution des résultats. En effet, dans la recherche de microblogs et afin de respecter la contrainte de recherche en temps-réel, nous devons nous positionner à l'instant où la requête est soumise. Dans un contexte réaliste d'emploi du moteur de recherche de microblogs, les tweets publiés après le `querytime` de la requête ne sont évidemment pas connus !

Pour nos analyses, nous avons conservé les 1500 premiers tweets restitués par Lucene pour chaque requête.

3.2 Observations

Lucene, dans sa configuration décrite ci-dessus, obtient un rappel moyen de 0,7188 avec les requêtes de 2011 et de 0,6340 avec les requêtes de 2012. Même si le modèle vectoriel arrive à restituer une bonne proportion des documents pertinents, le nombre des documents pertinents non restitués varie d'une requête à une autre.

1. `http://lucene.apache.org`

Les deux figures 3.1 et 3.2 montrent les proportions des tweets pertinents restitués par le modèle vectoriel par rapport à tous les tweets pertinents pour les requêtes des éditions de 2011 et de 2012.

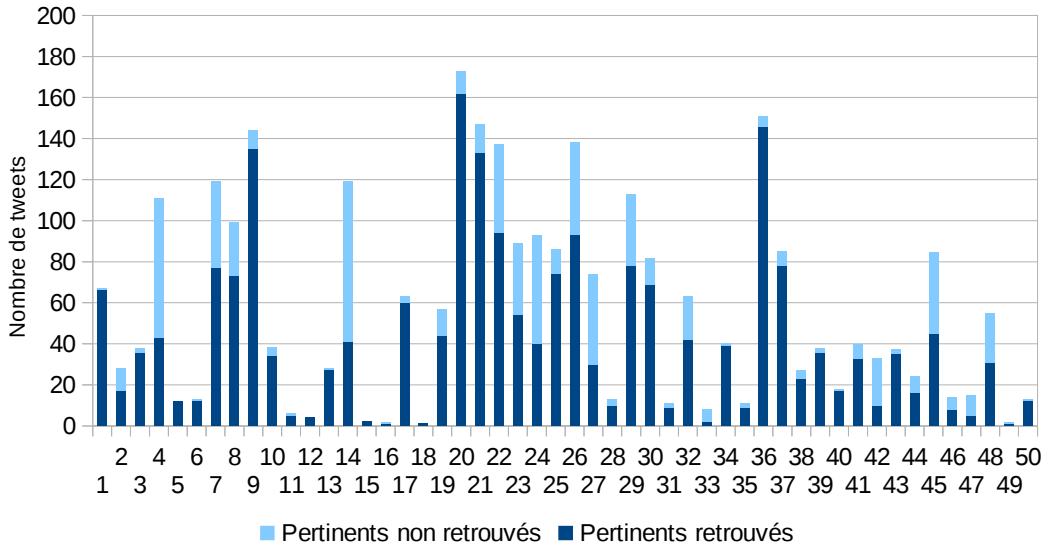


FIGURE 3.1 – Répartition des tweets pertinents restitués avec le modèle vectoriel par rapport à tous les tweets pertinents connus pour chaque requête de 2011

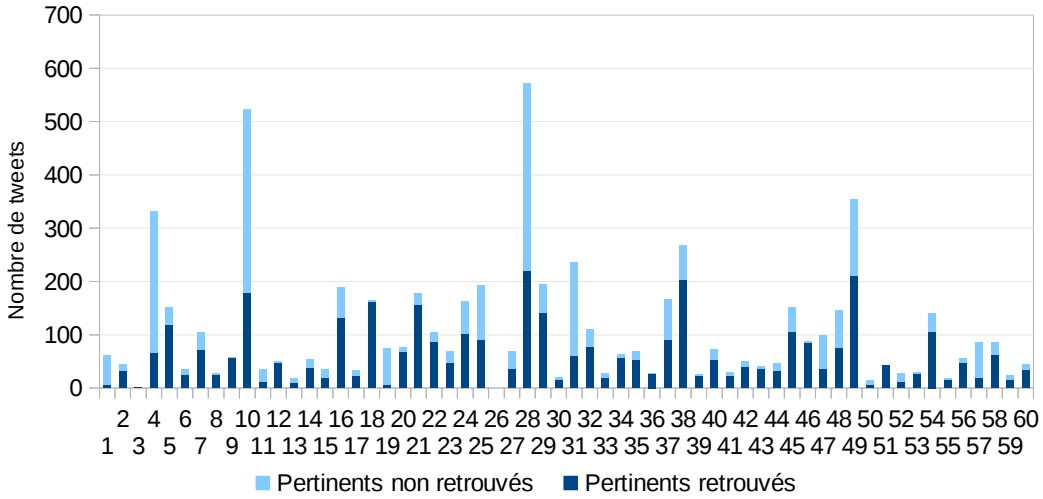


FIGURE 3.2 – Répartition des tweets pertinents restitués avec le modèle vectoriel par rapport à tous les tweets pertinents connus pour chaque requête de 2012.

Dans la suite, on note chaque requête par « son numéro »/« édition de TREC ». Sur l'ensemble des 109 requêtes de 2011 et 2012, le modèle vectoriel restitue tous les tweets pertinents de 22 requêtes. Pour 30 requêtes, moins de 5 documents pertinents

sont manquants. C'est le cas par exemple des requêtes `BBC World Service staff cuts (1/2011)`, `MSNBC Rachel Maddow (34/2011)`, `release of "Known and Unknown" (14/2011)`, `Starbucks Trenta cup (12/2012)` et `Tea Party caucus (53/2012)`.

Pour les autres requêtes, le nombre de documents pertinents non retrouvés varie d'une requête à une autre. Par exemple, sept documents pour `Giffords' recovery (37/2011)`, 23 pour `Holland Iran envoy recall (42/2011)`, 68 pour `Mexico drug war (4/2011)`, 78 pour `release of "The Rite" (14/2011)` , 179 pour `smartphone success (31/2012)` et 345 pour `fishing guidebooks (10/2012)` qui représentent le nombre le plus grand de tweets non restitués pour une requête.

Nous notons également que le nombre de tweets non restitués n'est pas proportionnel avec le nombre de tweets pertinents de la requête. Par exemple, le modèle vectoriel n'a pas restitué uniquement 5 tweets pertinents pour la requête `Moscow airport bombing (36/2011)` ayant pourtant 151 documents pertinents et 44 documents pour la requête `reduce energy consumption (27/2011)` ayant pourtant 74 documents pertinents en totalité. Ou encore, pour la requête `farmers markets opinion (21/2012)`, le modèle vectoriel identifie 68 tweets pertinents sur 76, soit 90 % de rappel. Cependant, pour le topic `The daily (4/2012)` 66 tweets pertinents parmi les 266 ont été restitués, soit un rappel de 25 %.

Nous avons analysé les résultats requête par requête pour identifier les problèmes auxquels les modèles de RI sont confrontés et qui pénalisent notamment le rappel. Le problème le plus remarquable observé à l'issue de notre analyse est *la différence de vocabulaire* (*vocabulary mismatch*) entre la requête et les tweets pertinents. Ce problème est bien connu en recherche d'information (Furnas et al., 1988). Dans notre cas, on le rencontre sous plusieurs formes.

1. **Absence totale des termes de la requête dans les documents pertinents**. Nous avons observé qu'un nombre important de tweets traite du sujet de la requête sans avoir, pour autant, aucun terme en commun avec cette dernière. C'est le cas par exemple de la requête `Amtrak train service (23/2011)`. Ce phénomène concerne 29 documents pertinents parmi 35 non retrouvés. Ces documents traitent des fonds réservés pour construire une nouvelles ligne de train ou relatent les difficultés des voyageurs. C'est le cas également de la requête `Obama birth certificate (41/2011)`. Certains documents évoquent des confusions sur la nationalité du président. Nous pouvons également citer le topic `British Government Cuts (1/2012)`, pour lequel ont été jugés pertinents des tweets qui traitent des licenciements dans le secteur public, de la baisse des salaires des employés dans certains secteurs, des coupes de budgets consacrés aux Jeux Olympiques, etc.

   Ce phénomène est présent pour 58 requêtes sur 109 (53 %), à hauteur de 40 %

des tweets pertinents non restitués. Plus précisément, ce problème apparaît pour au moins 1 800 tweets pertinents non restitués parmi les 4 448 tweets pertinents non restitués que nous avons au total sur toutes les requêtes.

2. **Problèmes des noms propres et des entités nommées**. Une première remarque concerne les noms propres orthographiés de différentes manières. Par exemple, pour le topic `Glen Beck (9/2012)`, dans certains tweets pertinents les utilisateurs emploient `Glenn` plutôt que `Glen`. Également, pour le topic `Bieber and Stewart trading places (13/2012)`, les utilisateurs emploient les prénoms `justin` et `jon`. Les entités nommées peuvent également être écrites de différentes manières : dans le topic `anti-bullying (40/2012)` les tweets non restitués contiennent `cyberbulling` plutôt que `bullying`. C'est le cas également du topic `Superbowl commercials (49/2012)`, pour lequel les auteurs utilisaient généralement `super bowl` en deux termes, ou encore `Bed bug` au lieu de `bedbug (2/2012)`.

   D'autre part, nous avons remarqué que certaines requêtes contiennent des entités nommées contenant des prépositions. Cependant, Lucene prend en compte les prépositions comme des termes vides et les élimine, ce qui modifie le sens de la requête. C'est le cas par exemple de la requête `release of "the Rite" (14/2011)`. C'est la cas également de la requête `the daily`, le moteur de recherche a extrait des tweets contenant le terme `daily` plutôt que des tweets traitant du journal `the daily` (266 tweets pertinents non retrouvés parmi les 332 pertinents du topic). Ceci résulte de notre utilisation d'une liste de mots vides. Cette requête aurait probablement conduit à de meilleurs résultats si elle avait été traitée sous forme d'expression.

   Les problèmes liés aux entités nommées sont présents dans 7 topics sur 119 (5 %), à hauteur de 50 % des tweets pertinents non restitués. Plus précisément, au moins 546 tweets pertinents non restitués sur les 4 448 présentent ce phénomène.

3. **Problèmes de lemmatisation**. Une première remarque est que des termes différents ne sont pas appariés, alors qu'ils relèvent d'une même racine. Par exemple, pour la requête `somalian piracy (57/2012)` étaient présents dans les tweets jugés pertinents les termes `pirates` ou `pirate`. La requête `global warming and weather (29/2011)` contient le terme « warmism » et non pas « warming ».

   Nous avons constaté également ce problème avec les requêtes contenant des termes qui reflètent la nationalité ou des noms de pays. Les documents pour ces requêtes contiennent les noms des pays et non pas les nationalités telles qu'elles apparaissent dans les requêtes ou l'inverse. C'est le cas par exemple de la requête `Mexico drug wars (4/2011)`. Les documents non restitués de cette

requête contiennent souvent le terme « Mexican ». C'est le cas également de la requête `Pakistan diplomat arrest (7/2011)` où les documents contiennent le terme « pakistani ».

D'autre part, nous avons remarqué l'apparition des termes de la requête concaténés sous forme de #hashtags ou de @citation. Par exemple, dans le topic `texting and driving (54/2012)` tous les tweets pertinents non restitués contiennent les termes de la requête mais concaténés en un hashtag (`#donttextanddrive`). C'est le cas de la requête `BBC World Service staff cuts (1/2011)` le document non restitué contient le hashtag `#BBCWorldService`. Pour la requête `Taco Bell filling lawsuit (20/2011)` certains documents non restitués contiennent le hashtag `#TacoBell` ou la citation `@TacoBell`...

Les lemmatiseurs utilisés par les moteurs de recherche — Porter (1980) dans notre cas — sont incapables de résoudre ce type de problème, ce qui explique l'impossibilité de Lucene à restituer ces tweets.

Ce phénomène est clairement présent dans 13 topics sur 109. Plus précisément, cela représente au moins 210 tweets pertinents non retrouvés sur les 4 448.

4. **Acronymes écrits de différentes manières**. C'est le cas du topic `FDA approval of drugs (8/2012)`, pour lequel les tweets pertinents contenaient également l'acronyme `USFDA`. Nous avons également constaté que pour le topic `NCIS (27/2012)` plusieurs tweets pertinents contenaient la signification de l'acronyme : `Naval Criminal Investigative Service`. Ce phénomène est présent clairement dans deux topics sur 109, à hauteur de 50 % des tweets pertinents non restitués. Cela représente au moins 50 tweets pertinents non restitués sur les 4 448.

Outre *la différence de vocabulaire*, nous avons remarqué que tous les **termes des requêtes n'apparaissent pas avec la même importance dans les requêtes**. Certains termes des requêtes n'aident ainsi pas à sélectionner les tweets pertinents.

Ce phénomène apparaît de trois manières : (i) des requêtes contenant des termes qui n'apparaissent pas dans la majorité des documents pertinents non restitués, (ii) des requêtes contenant des termes qui apparaissent toujours, même dans les document non pertinents retrouvés et (iii) des requêtes composées uniquement de termes concernés par (i) ou (ii).

Par exemple du premier cas (i), les documents non retrouvés de la requête `2022 FIFA soccer (2/2011)` ne contiennent jamais le terme "soccer". Les documents non retrouvés de la requête `phone hacking British politicians (7/2011)` ne contiennent jamais le terme "politicians". Dans la requête `fishing guidebooks (10/2012)`, 345 tweets pertinents non retrouvés parmi les 524 pertinents du topic

ne contiennent pas le terme `guidebooks`, ni un dérivé de ce terme.

Comme exemple du deuxième cas, les documents non retrouvés de la requête `Super Bowl, seats (24/2011)` où `Super bowl` apparaît dans tous les documents restitués et les documents pertinents non restitués. Toutefois, ce phénomène n'a été observé que pour cette requête. Pour le troisième cas, les documents non restitués de la requête `Emanuel residency court rulings (21/2011)` ne contiennent jamais les termes `Emanuel`, `residency` et `ruling`, mais contiennent tous le terme `court`. C'est également le cas de la requête `reduce energy consumption (27/2011)` pour lequel les documents non restitués ne contiennent jamais les termes `reduce` et `consumption` mais contiennent toujours le terme `energy`. Ce phénomène a été observé dans 17 requêtes sur les 50 et a empêché la restitution d'au moins 200 documents pertinents.

De façon identique, nous avons constaté que, dans le cas des requêtes contenant des entités nommées, les tweets pertinents non retrouvés contiennent ces entités nommées, mais sans les autres termes des requêtes. Par exemple, pour le topic `McDonalds food (28/2012)`, tous les tweets non restitués contiennent seulement le terme `McDonalds` parmi tous les termes de la requête (350 tweets pertinents non restitués parmi 572 tweets pertinents de cette requête).

Suite à ces observations, nous avons voulu savoir pour combien de tweets la prise en compte des contenus des URL qu'ils contiennent permettrait de régler ce problème de vocabulaire. En d'autres termes, nous avons voulu savoir si les termes des requêtes non présents dans des tweets pertinents étaient présents dans les documents pointés par les URL. Nous avons donc analysé le contenu des URL des tweets pertinents non restitués et nous avons constaté que leur prise en compte aiderait à retrouver des tweets pertinents dans 41 topics sur 109 (37 %). Cela représente au moins 800 tweets pertinents non restitués.

4 Synthèse

Le tableau 3.1 résume les différentes observations de notre analyse, qui sont des problèmes classiques de la RI. Leurs effets sont cependant amplifiés avec les microblogs à cause de leur faible longueur, ce qui implique un vocabulaire limité. Quel que soit le modèle de RI utilisé, mesurer la similarité entre une requête qui ne dépasse souvent pas quatre termes et un microblog composé en moyenne de 15 termes revient à une présence ou absence des termes de la requête dans un microblog, dans la majorité des cas.

Au niveau des analyses des facteurs limitant l'efficacité du modèle de recherche sur les microblogs, nous avons montré que le problème principal, comme attendu,

Cause de la défaillance	Pourcentage de tweets non restitués sur 4 448 pertinents au total	Pourcentage de requêtes concernées sur 109 requêtes
Absence totale des termes des topics dans les tweets pertinents	$\simeq$ 40,46 %	51,21 %
Termes des requêtes avec des importances différentes	$\simeq$ 20,12 %	16,51 %
Termes des requêtes à traiter sous forme d'expression et/ou sans liste de mots vides	$\simeq$ 7,77 %	2,75 %
Noms propres et entités nommées orthographiés de différentes manières	$\simeq$ 4,49 %	4,58 %
Termes non appariés mais dérivant d'une même racine	$\simeq$ 4,04 %	8,25 %
Termes de la requête concaténés sous forme de hashtag ou de citation	$\simeq$ 1,79 %	6,42 %
Acronymes écrits de différentes manières	$\simeq$ 1,12 %	1,83 %

Tableau 3.1 – Récapitulatif des différents facteurs limitant l'efficacité du modèle de recherche sur les microblogs

provient de la concision des microblogs. Cette concision engendre une correspondance limitée entre les termes des microblogs et les termes des requêtes, même s'ils sont sémantiquement similaires. Ce fait est apparu de différentes manières : absence totale des termes de certaines requêtes dans les tweets pertinents, noms propres et entités nommés orthographiés de différentes manières... Nous avons fréquemment identifié des problèmes de lemmatisation : termes non appariés même si dérivant d'une même racine ou des termes concaténés pour former des hashtags ou des citations. Outre les problèmes de vocabulaire, nous avons remarqué que, pour certaines requêtes, les termes n'ont pas un caractère discriminant. Ces termes sont fréquemment présents dans les documents pertinents et les documents non pertinents ou bien ils n'apparaissent jamais.

De manière générale, les problèmes soulevés avec les requêtes de 2011 sont les mêmes pour les requêtes de 2012. Cependant, les requêtes de 2012 contiennent un nombre plus important de tweets pertinents, ce qui les rend plus difficiles (2 864

tweets pertinents pour les 49 requêtes de 2011 / 6 286 tweets pertinents pour les 60 requêtes de 2012).

Nous avons montré que la plupart des problèmes de la recherche d'informations dans les microblogs ne dépend pas du modèle de recherche. Ces problèmes ne concernent pas les fréquences des termes dans les microblogs, ou bien la distribution des termes dans les microblogs. Ce sont plutôt des problèmes de vocabulaire et des problème de lemmatisation. Le problème de vocabulaire, avec ses différentes formes observées, ou bien les problèmes de lemmatisation, peuvent affecter l'efficacité de n'importe quel modèle qui se base uniquement sur le contenu textuel brut des microblogs et avec les requêtes avec leurs descriptions initiales. Cependant, un problème, tel que *les termes de requêtes qui ont des importances différentes*, regarde exactement le fonctionnement des modèles de recherche, et sa gravité dépend fortement de la manière avec laquelle un modèle calcule les scores de pertinence. La prise en compte de la fréquence du terme dans la collection (IDF) joue ainsi un rôle très important ici.

Le problème de vocabulaire semble surmontable avec l'expansion de requêtes et de documents. Les termes à caractère non discriminant peuvent être pondérés en exploitant le *feedback*. C'est pourquoi, dans le chapitre suivant, nous présenterons les différentes méthodes d'expansion de requêtes et de documents que nous avons appliquées pour surmonter le problème du vocabulaire souvent rencontré dans la recherche d'information dans les microblogs.

Chapitre 4

Expansion de requêtes et de documents pour la recherche de microblogs

1 Introduction

À travers l'analyse de défaillance conduite et présentée dans le chapitre précédent, nous avons montré que le vocabulaire limité lié à la taille réduite des microblogs est le facteur empêchant le plus les SRI de restituer des microblogs pertinents.

Notre objectif, dans ce chapitre, est de proposer des éléments de solutions pour surpasser ces limites : **absence de termes en commun entre les requêtes et les microblogs, termes des requêtes n'ayant aucun caractère discriminant, entités-nommées orthographiées de différentes manières, problèmes de lemmatisation, termes concaténés...**

Une solution au problème du vocabulaire est l'expansion de requêtes ou de documents (technique connue en RI). Nous proposons ici d'**améliorer la représentation des requêtes**. Dans un premier temps, nous exploitons des ressources externes pour étendre les requêtes. Ces ressources comprennent des articles d'actualité ainsi que la base lexicale WordNet. Nous testons également l'impact de méthodes de ré-injection de pertinence (telles que Rocchio et BM25). Nous proposons également d'**améliorer la représentation des microblogs**. Nous testons quelques méthodes pour améliorer la représentation des microblogs, telles que l'expansion des hashtags et l'exploitation des contenus des URLs publiées dans les microblogs.

2 Expansion de requêtes

Pour améliorer la représentation des requêtes, nous avons considéré différentes ressources. Certaines sont externes par rapport à la collection de tweets. L'infor-

mation dans les tweets étant très dépendante du temps, nous avons employé des ressources sensibles au temps pour étendre les requêtes, telles que les articles des actualités publiés dans les journaux les plus populaires dans le monde. D'autre part, nous avons exploité la base lexicale WordNet pour trouver les différents aspects des requêtes et l'API de correction orthographique du moteur de recherche Bing pour trouver les différentes formes des entités nommées. En outre, nous avons étendu les requêtes à partir des tweets en appliquant des techniques de ré-injection de pertinence (*relevance feedback*).

Pour réaliser les expérimentations qui suivent , nous nous sommes basés sur les 60 requêtes de TREC Microblog 2012. Pour chaque requête, nous avons considéré les 1500 premiers tweets restitués avec le modèle vectoriel. La validation des améliorations ou des dégradations est réalisée selon le test t de Student pairé et bilatéral avec $p < 0{,}05$. Nous nous sommes basés sur le run obtenu avec le modèle vectoriel implémenté dans Lucene comme baseline.

2.1 Exploitation des articles d'actualités

La première source que nous avons considérée est constituée des actualités publiées de façon concomitante aux requêtes. Nous avons en effet remarqué que la majorité des topics des requêtes portent sur des actualités (50 % des topics). Pour cette raison, nous proposons d'étendre les requêtes avec des mots-clés extraits à partir des articles de presse publiés sur le sujet du topic. Les API du NYTimes [1] et du Guardian [2] permettent d'obtenir des articles de ces journaux en fonction d'une requête. Étant donné que les articles restitués sont classés selon leur pertinence décroissante, nous avons considéré les cinq premiers articles restitués par chaque source et publiés avant la date du topic pour produire un méga-document (Klas et Fuhr, 2000). Ensuite, nous avons utilisé l'API Alchemy [3] pour extraire les mots-clés représentatifs de ce méga-document. L'API Alchemy réalise une analyse linguistique, un traitement du langage naturel et un apprentissage automatique pour analyser le contenu et en extraire des mots-clés. Nous avons évalué l'extension de la requête avec trois (3Act) ou sept(7Act) termes renvoyés par Alchemy. Les nouvelles requêtes sont formées par les termes initiaux des requêtes et les termes d'expansion. Dans un premier temps, nous ne pondérons pas les termes ajoutés dans la requête (tous les termes de la requête étendue ont un poids égal à 1). Dans un second temps, nous pondérons uniquement les termes ajoutés aux requêtes (3Act(pond) et 7Act(pond)) avec un poids de 0,8 (choix arbitraire pour ces premières expérimentations). Les résultats sont présentés dans le tableau 4.1. La colonne Run contient le nom des runs.

1. `http://developer.nytimes.com/`
2. `http://www.guardian.co.uk/open-platform/`
3. `http://www.alchemyapi.com/`

Ils sont présenté sous la forme Modèle-Requête-Champ utilisé. Le champ utilisé spécifie le contenu employé pour la restitution. À ce niveau, nous utilisons uniquement le contenu textuel des tweets (*Tweets*) dans la restitution. Notons dès à présent que dans les sections suivantes, nous exploiterons d'autres contenus pour la restitution des résultats, hormis le contenu textuel des tweets.

Run	Modèle	Requête étendue	Champ utilisé	P@30	Rappel	MAP
Baseline	VSM	—	Tweets	0,2842	**0,6340**	0,1871
VSM-3Act-Tweets	VSM	3Act	Tweets	0,2689	0,5691	0,1699
VSM-7Act-Tweets	VSM	7Act	Tweets	0,3040*	0,5985	0,1923*
VSM-3Act(pond)-Tweets	VSM	3Act(pond)	Tweets	0,2785	0,5923	0,1806
VSM-7Act(pond)-Tweets	VSM	7Act(pond)	Tweets	**0,3079***	0,6156	**0,1962***

Tableau 4.1 – Emploi des articles de type actualité pour l'expansion de requêtes (avec et sans pondération des termes d'expansion, 1500 résultats par requête). Un astérisque indique une amélioration significative par rapport à la baseline.

Concernant le rappel, nous constatons que la pondération améliore les résultats par rapport à la non pondération (runs Modèle-X(pond)-Champ par rapport aux runs Modèle-X-Champ). Par exemple le run « VSM-3Act(pond)-Tweets » améliore le run « VSM-3Act-Tweets » de 3,91 %. Cependant, aucune amélioration significative n'est à remarquer par rapport au run Baseline utilisant la requête originale.

Concernant la P@30 et la MAP, le fait d'étendre les requêtes avec trois termes uniquement dégrade les résultats, que ce soit avec ou sans pondération. Cependant, en étendant les requêtes avec sept termes, nous observons des améliorations significatives par rapport à la « baseline ». Cette amélioration est légèrement plus importante en pondérant les termes d'expansion. Le run « VSM-7Act(pond)-Tweets » est celui qui a obtenu la meilleure amélioration par rapport à la baseline : 8,33 % en p@30 et 4,86 % en MAP.

De manière générale, l'emploi des articles d'actualités comme source pour étendre les requêtes a amélioré la précision et a dégradé le rappel. En d'autre termes, cette approche a amélioré les rangs des premiers microblogs pertinents (P@30 améliorée de 8,33 %), sans pouvoir retrouver autant de nouveaux tweets pertinents. Nous avons comparé les tweets pertinent des runs « VSM-7Act(pond)-Tweets » et « Baseline » : sont également présent 91 % des tweets pertinents du run « Req7ActPondTweet » dans le run « Baseline ». Nous avons également remarqué que cette méthode d'expansion a renforcé la pertinence d'une partie des documents pertinents. Cette partie se compose des microblogs contenant totalement ou partiellement les termes initiaux des requêtes. Cependant, elle n'a pas aidé à restituer de nouveaux tweets pertinents, en particulier ceux qui ne contiennent pas les termes des requêtes.

La dernière observation peut être expliquée par le fait que nous nous sommes basés sur les premiers articles d'actualités résultant d'une recherche avec les termes

des requêtes sur les deux API (NYTimes et Guardian). Ces APIs utilisent leurs moteurs de recherche pour trouver des articles en fonction des termes de nos requêtes. Ceci implique que les termes les plus importants retrouvés à partir de ces articles correspondent en premier lieu aux termes initiaux des requêtes, ou bien aux termes fortement dépendant des termes des requêtes (les termes présentant l'aspect sémantique général des requêtes). Ainsi, cette méthode permet de mieux représenter les requêtes initiales, sans donner d'autres aspects sémantiques des requêtes, permettant ainsi de restituer les microblogs pertinents et portant sur les sujets des requêtes, tout en n'ayant aucun terme en commun avec elles. Ceci explique ainsi la dégradation du rappel et l'amélioration de la précision.

Afin de retrouver des termes d'expansion représentant d'autres aspects des requêtes, nous avons testé l'expansion avec la base lexicale WordNet.

2.2 Exploitation de la base lexicale WordNet

La base de données lexicale WordNet a été souvent utilisée en RI comme un moyen de désambiguïsation et d'extension de requêtes. Nous avons testé cette stratégie en étendant chaque terme de la requête avec le premier synset retrouvé. Chaque requête étendue va ainsi être composée des termes de la requête initiale et des termes d'expansion. De la même manière que dans le paragraphe précédent, dans un premier temps, nous ne pondérons pas les termes ajoutés dans la requête. Tous les termes ont un poids égal à 1. Dans un second temps, nous avons pondéré (WN(pond)) uniquement les termes ajoutés aux termes initiaux des requêtes avec un poids de 0,8 (choix arbitraire pour observer l'impact de la pondération). Les résultats sont présentés dans le tableau 4.2.

Run	Modèle	Requête étendue	Champ utilisé	P@30	Rappel	MAP
Baseline	VSM	—	Tweets	0,2842	0,6340	0,1871
VSM-WN-Tweets	VSM	WN	Tweets	0,2797	0,6305	0,1854
VSM-WN(Pond)-Tweets	VSM	WN(pond)	Tweets	**0,2881**	**0,6362**	**0,1878**

Tableau 4.2 – Récapitulatif des différents runs testés sans pondération des termes ajoutés aux requêtes.

L'expansion avec WordNet n'améliore non plus pas les résultats par rapport à la baseline (0,37 % d'amélioration sur la MAP, 1,37 % sur la P@30 et 0,34 % sur le rappel entre le run « Baseline » et « VSM-WN(pond)-Tweets »). En outre, les améliorations sur les trois mesures ne sont pas significatives. En fait, au niveau des tweets pertinents restitués, 59/60 des requêtes ont renvoyé exactement les mêmes tweets pertinents. La différence est uniquement présente au niveau de la requête *somalian piracy*. Avec l'expansion, cette requête s'est transformée en *somalian somali piracy*. En réalité, WordNet a compensé une faiblesse de Porter. Dans l'analyse de

défaillance, nous avons signalé ce problème : plusieurs tweets pertinents de cette requête contiennent le terme *somalia* ou *somalis*. Porter n'arrive pas à traiter et correspondre ces variances. Ainsi, l'ajout du terme *somali* dans la requête a permis la restitution de 42 nouveaux tweets pertinents. D'où la faible amélioration globale.

Concernant la pondération, nous avons observé le même impact que celui observé dans le paragraphe précédent. Elle améliore les résultats par rapport à la non pondération. Nous avons comparé les runs « VSM-WN-Tweets » et « VSM-WN(pond)-Tweets » et nous avons trouvé que 100 % des tweets pertinents du premier run apparaissent dans le deuxième run. En contre partie, le deuxième run a restitué uniquement 14 nouveaux tweets pertinents par rapport au premier run.

De manière générale, l'emploi de WordNet n'a servi à améliorer ni le rappel ni la précision. Pour 59 requêtes parmi 60, aucun nouveau tweet pertinents n'a été observé. Nous avons également testé l'emploi de plusieurs termes d'expansion mais ceci n'a fait que dégrader les résultats.

2.3 Suggestions orthographiques

Nous avons remarqué, dans certaines requêtes, des entités nommées orthographiées de manières différentes à celles dans les tweets pertinents. C'est pourquoi nous avons testé l'outil « Bing spelling suggestions[4] ». Cette API permet de corriger les termes mal orthographiés et de retrouver les autres écritures des entités nommées. Pour chaque terme d'une requête, nous avons ajouté ses autres formes d'écriture dans la requête initiale (sans pondération). Cependant, seules deux requêtes parmi les 60 ont été modifiées. Ce sont la requête « Bedbug epidemic » qui est devenue « Bedbug epidemic bed bug » et la requête « Glen Beck » qui est devenue « Glen Beck Glenn ». Les résultats de ces deux requêtes ont un rappel plus élevé que celui de la baseline (19,91 % et 2,08 % respectivement). En considérant toutes les requêtes, nous avons obtenu une amélioration du rappel de 0,28 % (tableau 4.3). Cependant, ni cette amélioration, ni les améliorations des autres mesures ne sont significatives.

Run	Modèle	Requête étendue	Champ utilisé	P@30	Rappel	MAP
Baseline	VSM	—	Tweets	0,2842	0,6340	0,1871
VSM-ReqBing-Tweet	VSM	ReqBing	Tweets	**0,2893**	**0,6358**	**0,1884**

Tableau 4.3 – Test de l'amélioration des performance via la correction orthographique des requêtes.

4. `http://www.bing.com/developers/`

2.4 Réinjection de pertinence

Une source typique pour étendre les requêtes est constituée de l'ensemble des termes présents dans les premiers documents restitués en réponse aux requêtes initiales. Cette technique s'appelle *la réinjection de pertinence* (Relevance Feedback). Nous avons testé et analysé l'impact de deux approches classiques de la RI afin de voir leur efficacité sur ce genre de documents : Rocchio et le modèle BM25.

2.4.1 Expansion de requêtes avec Rocchio

Nous avons utilisé la version améliorée (Salton et Buckley, 1997) de la formule originale de Rocchio (1971). Cette version prend en compte uniquement les documents qui ont obtenu les meilleures scores dans la reformulation. La formule est la suivante :

$$Q_{nouv} = \alpha.Q_{orig} + \frac{\beta}{|R|}.\sum_{\vec{r}\in R}\vec{r} \tag{4.1}$$

Q_{nouv} est le vecteur des termes pondérés de la requête étendue. Q_{orig} est le vecteur de termes pondérés de la requête initiale. R est l'ensemble des documents pertinents. $\vec{r}$ est le vecteur des termes obtenus de R pour l'expansion. Nous avons gardé les valeurs par défaut des paramètres : $\alpha = 1$ and $\beta = 0,75$. La taille de R est fixée à 10. Ce choix est consistant à la vu des expérimentations réalisées sur les collections de TREC (Carpineto et al., 2001). Le nombre de termes ajoutés est également limité à 10. Ce choix correspond au résultat d'une étude sur l'expansion de requête à partir du *feedback*, pour la recherche de microblogs. Cette étude est réalisée par Aboulnaga et Clarke (2012).

L'objectif de l'emploi de Rocchio est double : d'une part, il permet de résoudre le problème de vocabulaire en améliorant la représentation des requêtes avec un vocabulaire plus riche. D'autre part, il permet, au travers des meilleurs résultats de la première restitution, de pondérer les termes des requêtes. Ceci pourrait résoudre le problème noté dans l'analyse de défaillance : les termes de la requête n'ont pas tous la même importance.

Dans un premier temps, le poids des termes d'expansion dans le vecteur $\vec{r}$ ont été calculés avec TF-IDF (Rocch(TF.IDF)). Les résultats sont présentés dans le tableau 4.4.

Nous remarquons que la technique de Rocchio améliore significativement les résultats par rapport à la baseline, que ce soit au niveau du Rappel, de la P@30 ou la MAP : respectivement 8,00 %, 13,72 % et 18,17 %. Nous avons comparé les tweets pertinents du run « VSM-Rocch(TF.IDF)-Tweets » avec les tweets pertinents du run « Baseline ». Nous avons trouvé que 14 % des tweets pertinents du run « VSM-Rocch(TF.IDF)-Tweets » n'existaient pas dans le run « Baseline ». Ceci correspond

Run	Modèle	Requête étendue	champ utilisé	P@30	Rappel	MAP
Baseline	VSM	—	Tweets	0,2842	0,6340	0,1871
VSM-Rocch(TF.IDF)-Tweets	VSM	Rocch(TF.IDF)	Tweets	**0,3232***	**0,6822***	**0,2211***

Tableau 4.4 – Expansion de la requête initiale avec Rocchio. Les poids des termes d'expansion sont calculés avec TF.IDF. Un astérisque indique une amélioration significative par rapport à la baseline.

à 589 nouveaux tweets pertinents. Ces nouveaux tweets pertinents sont répartis sur 42 requêtes parmi les 60. Ce sont souvent des tweets contenant un seul terme de la requête initiale et certains termes d'expansion. Cependant, l'expansion a ignoré 77 tweets pertinents qui existaient déjà dans le run « Baseline ». Ces tweets sont répartis sur toutes les requêtes avec un ou deux tweets non retrouvés pour chacune.

Dans un deuxième temps, les poids des termes d'expansion sont calculés avec le modèle BM25.

Run	Modèle	Requête étendue	champ utilisé	P@30	Rappel	MAP
Baseline	VSM	—	Tweets	0,2842	0,6340	0,1871
VSM-Rocch(TF.IDF)-Tweets	VSM	Rocch(TF.IDF)	Tweets	0,3232*	**0,6822***	0,2211*
VSM-Rocch(BM25)-Tweets	VSM	Rocch(BM25)	Tweets	**0,3311***	0,6764*	**0,2304***

Tableau 4.5 – Expansion de la requête initiale avec Rocchio. Les poids des termes d'expansion sont calculés avec BM25. Un astérisque indique une amélioration significative par rapport à la baseline.

Le tableau 4.5 présente les nouveaux résultats. Par rapport à la « baseline », nous avons obtenu des améliorations significatives sur les trois mesures : 6,70 %, 16,50 % et 23,14 % respectivement sur le rappel, la P@30 et la MAP. Par rapport au run qui emploie TF.IDF pour pondérer les termes d'expansion (« VSM-Rocch(TF.IDF)-Tweets »), nous remarquons des améliorations significatives uniquement sur la P@30 et la MAP : 2,44 % et 4,20 % respectivement. Les tweets pertinents du run « VSM-Rocch(BM25)-Tweets » sont à 99 % ceux du run « VSM-Rocch(TF.IDF)-Tweets ». Ce sont également les mêmes termes d'expansion qui ont été sélectionnés et ajoutés dans les requêtes initiales en calculant les poids avec BM25, que ceux sélectionnés avec TF.IDF. Toutefois, la pondération des termes d'expansion avec les scores de BM25 à permis de mieux classer les tweets pertinents, ce qui a amélioré la précision et la MAP.

2.4.2 Expansion de requêtes via le modèle BM25

Une des méthodes classiques de réinjection de pertinence est le mécanisme « naturel » du modèle BM25. Naturellement, le facteur approximatif de IDF dans le

modèle BM25 est :

$$IDF^{**} = log\left(\frac{r + 0.5/n - r + 0.5}{R - r + 0.5/N - R - n + r + 0.5}\right) \tag{4.2}$$

avec r est le nombre de documents pertinents contenant le terme t, R est le nombre de tous les documents contenant le terme t, n est le nombre de documents pertinents et N est la taille de la collection. En absence d'information de pertinence au préalable, ce facteur devient :

$$IDF^{*} = log\left(\frac{N - R}{R}\right) \tag{4.3}$$

L'emploi de IDF^{**} nécessite une connaissance préalable des documents pertinents. Ainsi, l'idée est de considérer les premiers résultats de la première restitution réalisée en considérant IDF^{*}. Cet ensemble est supposé être l'ensemble de pertinence (feedback). Ensuite, on réalise une deuxième restitution, mais toujours avec la requête initiale, en considérant IDF^{**} et le *feedback* pour le calcul des scores. En se basant sur des expérimentations réalisées sur les collections de TREC (Carpineto et al., 2001), le *feedback* est constitué des 10 premiers tweets restitués avec la requête initiale.

Pour étudier l'impact de l'expansion de requêtes avec le modèle BM25, il est évident de comparer les différentes propositions (*emploi du feedback et expansion*) avec les résultats du modèle BM25 de base (équation 5.12). Ceci nous a permis également de comparer les résultats des modèles BM25 et vectoriel.

Le tableau 4.6 montre les résultats. BMX25fb indique que le modèle emploie le *feedback* dans le calcul du score de pertinence.

La première remarque est que le modèle BM25 (run « BM25- — -Tweets ») a obtenu des résultats plus faibles que le modèle vectoriel.

Run	Modèle	Requête étendue	champ utilisé	P@30	Rappel	MAP
Baseline	VSM	—	Tweets	0,2842	0,6340	0,1871
BM25- — -Tweets	BM25	—	Tweets	0,2836	0,6043	0,1654
BM25fb- — -Tweets	BM25fb	—	Tweets	0,2655	0,5940	0,1604
BMX25- — -Tweets	BMX25	—	Tweets	0,3186*	**0,6643***	0,2170*
BMX25fb- — -Tweets	BMX25fb	—	Tweets	0,3135	0,6364	0,2163
BMX25fb-Reqexp-Tweets	BMX25fb	Reqexp	Tweets	**0,3571***	0,6369	**0,2300***

Tableau 4.6 – Différentes configurations du modèle BM25. * montre une amélioration significative par rapport à configuration de base (run BM25).

Motivés par le travail de Ferguson et al. (2012), nous avons modifié les paramètres initiaux du modèle BM25 afin de limiter au maximum la prise en compte des facteurs de normalisation et la fréquence des termes dans le calcul du score. En fait,

comme nous l'avons déjà mentionné, dans la recherche de microblogs, la fréquence des termes n'améliore vraiment pas les résultats. De plus, la normalisation de la longueur des documents dégrade les résultats[5]. Nous avons ainsi paramétré $k1 = 0{,}1$ et $b = 0$. BMX25 indique la prise en compte de ces paramètres dans le modèle BM25. « BMX25- — -Tweets » et « BMX25fb- — -Tweets » représentent respectivement les runs sans et avec l'emploi du *feedback*, mais avec les nouveaux paramètres. On peut remarquer, dans un premier temps, que le nouveau paramétrage améliore considérablement les résultats : 10,0 %, 12,3 % et 31,2 % d'amélioration respectivement pour le rappel, la p@30 et la MAP, entre le run « BM25- — -Tweets » et le run « BMX25- — -Tweets ». Le run « BMX25- — -Tweets » est également meilleur que le run « Baseline ». Les améliorations sont respectivement de 4,8 %, 12,1 % et de 16,0 % dans le rappel, la p@30 et la MAP. Le run « BMX25- — -Tweets » contient 707 nouveaux tweets pertinents par rapport au run « Baseline » et contient 92,0 % des tweets pertinents du run « Baseline ». Ceci correspond à 390 tweets pertinents non retrouvés. Ces tweets se caractérisent de manière générale par leur longueur très réduite (un ou deux termes et une URL). La différence au niveau du nombre de tweets restitués entre les run « BMX25- — -Tweets » et « Baseline » correspond approximativement au nombre de tweets non restitués à la cause des différences dans les importance des termes des requêtes (695 sur les requêtes de 2012), observé dans le chapitre précédent. Le fait de se baser principalement sur le facteur IDF dans la restitution de microblog a résolu ce problème.

Le run « BM25fb- — -Tweets » est celui qui emploie le *feedback* avec le modèle BM25 de base. À ce niveau, aucune amélioration n'a été constatée. Concernant le run « BMX25fb- — -Tweets », les résultats montrent que, encore une fois, le *feedback* n'améliore pas les résultats. Nous avons comparé les tweets pertinents des runs « BMX25- — -Tweets » et « BMX25fb- — -Tweets ». 99% des tweets pertinents du run « BMX25fb- — -Tweets » existaient dans le run « BMX25- — -Tweets » (13 nouveaux tweets). Cependant, l'emploi du *feedback* a négligé 161 tweets pertinents. En réalité, 110 de ces tweets non restitués avaient un rang supérieur à 1500. Pour cette raison, ils n'ont pas été considérés dans le rappel. Nous pouvons ainsi constater que l'emploi du feedback ne permet pas de restituer de nouveaux tweets pertinents et ne résout pas le problème de vocabulaire.

Au lieu de fournir simplement une méthode de pondération des termes de la requête d'un utilisateur, la réinjection de pertinence peut également impliquer l'expansion de la requête avec certains termes (dans ce cas dix termes pour les raisons expliquées dans le paragraphe précédent) à partir du *feedback* (dix premiers tweets de la première restitution). Ces termes sont choisis par le facteur de pertinence de

5. Ceci coïncide avec les résultats du chapitre suivant où nous allons montrer que la longueur des microblogs est un facteur de pertinence dans la recherche de microblogs.

l'équation 4.2. Le run réalisant l'expansion et le *feedback* est « BMX25fb-Reqexp-Tweets ». Au niveau du rappel, aucune amélioration n'a été observée par rapport au run « BMX25- — -Tweets ». Cependant, la p@30 a progressé de 12,0 % et la MAP de 6,3 %. Nous avons comparé les tweets pertients des deux runs « BMX25fb-Reqexp-Tweets » et « BMX25- — -Tweets ». Même si le rappel s'est dégradé de manière significative, le run « BMX25fb-Reqexp-Tweets » contient 467 nouveaux tweets pertinents (13 %). En contrepartie, il a négligé 813 tweets pertinents qui existaient dans le run « BMX25- — -Tweets ».

Nous avons comparé également les runs « BMX25fb-Reqexp-Tweets » et « BMX25fb- — -Tweets ». Même si ces deux runs ont pratiquement le même nombre de tweets pertinents, ils diffèrent d'un ensemble considérable de tweets pertinents (de l'ordre de 470 tweets pertinents). L'expansion améliore considérablement le rang des tweets pertinents.

Ainsi, pour le modèle BM25 employé dans le cas de recherche de microblogs, nous pouvons conclure que le feedback dégrade le rappel. En outre, il n'améliore ni la MAP ni la précision, tant qu'il n'est pas accompagné d'une expansion de requêtes. L'expansion de requêtes améliore les rangs des tweets pertinents et réduit partiellement l'effet négatif du feedback au niveau du rappel.

3 Expansion de microblogs

Outre l'expansion des requêtes, nous avons évalué l'impact de l'expansion de microblogs, et ce de plusieurs façons : expansion de hashtags et emploi des URLs.

3.1 Expansion de hashtags dans les tweets

Dans l'analyse de défaillances du chapitre 2, nous avons constaté qu'un nombre important de tweets pertinents non restitués contient les termes de la requête collés ensemble sous forme de hashtags (par exemple, #TextAndDrive). Nous avons mis l'index à jour en étendant chaque hashtag composé avec les termes qui le composent. Nous avons remarqué que les auteurs mettaient parfois le premier caractère de chaque terme composant en majuscule. Ainsi, nous nous sommes basés sur cette observation pour étendre les hashtags composés. Pour chaque tweet contenant un hashtag composé, nous avons ajouté les termes composants au tweet (champ utilisée : *TweetsHashExp*). Une légère amélioration mais non significative dans le rappel (tableau 4.7) est constatée.

Run	Modèle	Requête étendu	champ utilisée	P@30	Rappel	MAP
BMX25- — -Tweets	BMX25	—	Tweets	0,3186	0,6643	**0,2170**
BMX25- — -TweetsHashExp	BMX25	—	TweetsHashExp	**0,3198**	**0,6681**	0,2166
Baseline	VSM	—	Tweets	0,2825	0,6340	0,1871
VSM- — -TweetsHashExp	VSM	—	TweetsHashExp	0,2785	0,6361	0,1859

Tableau 4.7 – Résultats après l'expansion de hashtags, avec le modèle vectoriel et le modèle BM25 (sans et avec paramétrage).

3.2 Emploi des URLs

À l'issue de notre analyse de défaillances et plus particulièrement de l'analyse des URLs publiées dans les tweets pertinents, nous avons remarqué que la prise en compte des pages web pointées par les URLs en complément des contenus des tweets pourrait améliorer la restitution des tweets pertinents. Le contenu des URLs présente souvent les termes des requêtes, même si le tweet ne les contient pas. Une première proposition consiste alors à la prise en compte d'un tweet selon 1) son contenu (champ utilisée : *Tweets*) ainsi que 2) le contenu des documents pointés par les URLs (champ utilisée : *Tweets+URL*) présentes dans le tweet (2 646 611 tweets contiennent une URL dans la collection). Nous avons commencé par considérer les deux champs (*Tweets+URL*) dans la recherche avec les requêtes originales. Le tableau 4.8 montre que l'emploi les URLs dans la restitution améliore significativement les résultats, que ce soit avec le modèle vectoriel ou bien BM25.

Run	Modèle	Requête étendue	Champ utilisé	P@30	Rappel	MAP
Baseline	VSM	—	Tweets	0,2825	0,6340	0,1869
VSM- — - Tweets+URL	VSM	—	Tweets+URL	0,3814*	**0,7171***	**0,2593***
BM25- — -Tweets	BM25	—	Tweets	0,2836	0,6043	0,1654
BM25- — -Tweets+URL	BM25	—	Tweets+URL	0,3816*	0,6686*	0,2267*
BMXx25- — -Tweets+URL	BMXx25	—	Tweets+URL	**0,3944***	0,6879*	0,2360*

Tableau 4.8 – Apport de l'emploi des URLs avec le modèle vectoriel et le modèle BM25. * montre une amélioration significative par rapport au run précédent.

Dans le cas du modèle vectoriel, le run « VSM- — - Tweets+URL » a eu des améliorations de 13,1 %, 35,0 % et 38,7 % sur le rappel, la P@30 et la MAP. En comparant les tweets pertinents des runs « VSM- — - Tweets+URL » et « baseline », nous avons remarqué que l'effet des URLs n'était pas totalement positif, en particulier au niveau de la sélection des tweets pertinents. Le run « VSM- — - Tweets+URL » contient 1013 (22,85 %) nouveaux tweets pertinents par rapport au run « Baseline ». Cependant, 275 tweets pertinents du run « Baseline » n'ont pas été de nouveau restitués.

Les mêmes améliorations sont constatées avec le modèle BM25 : 10,6 %, 34,6 % et 37,0 % respectivement sur le rappel, la p@30 et la MAP. Encore une fois, nous pouvons affirmer que l'effet des URLs n'est pas totalement positif sur la sélection des tweets pertinents. Cette observation est plus claire avec le modèle BM25. Le run « BM25- — -Tweets+URL » contient 1039 nouveaux tweets pertinents par rapport au run « BM25- — -Tweets ». Cependant, 670 tweets pertinents du run « BM25- — -Tweets » n'ont pas été de nouveau restitués.

Le double effet de l'emploi des URLs revient au fait qu'une quantité importante de tweets non pertinents contient les termes des requêtes dans les contenus des URLs. Nous avons remarqué cette observation même au niveau des contenus des tweets : plusieurs tweets non pertinents contiennent les termes des requêtes et traitent les sujet des requêtes...

Finalement, nous avons testé une configuration qui définit les paramètres du modèle BM25 en fonction du champ recherché. Les paramètres k_1 et b sont initialisés respectivement à 1,2 et 0,75 lorsque la recherche des termes d'une requête est effectuée sur le champ *UrlText* (*BMXx25*). Ils ont été initialisés à 0,1 et 0 lorsque la recherche est effectuée sur le champ *Tweets*. Le run avec cette configuration est « BMXx25- — -Tweets+URL ». Nous pouvons observer des améliorations de 2,9 %, 3,3 % et de 4,1 % respectivement sur le rappel, la P@30 et la MAP, par rapport au run « BM25- — -Tweets+URL ». Le parametrage a permis de restituer 215 nouveaux tweets pertinent. Cependant, 166 tweets pertinents du run « BM25- — -Tweets+URL » n'ont pas été de nouveau restitués.

De manière générale, nous pouvons remarquer que le modèle BM25 est plus performant au niveau de la précision. En d'autre termes, les rangs des tweets pertinents avec le modèle BM25 sont meilleurs (plus proche de la tête de liste) que les rangs des tweets pertinents avec le modèle vectoriel. En contrepartie, le modèle vectoriel restitue une quantité plus importante de tweets pertinents : il est meilleur au niveau du rappel. Concernant l'emploi des URLs, les résultats montrent qu'elles ont un rôle très important et améliorent les résultats de manière remarquable, même si elles sont la cause de la perte d'une quantité non négligeable de tweets pertinents.

Nous avons montré dans la section 2 que l'expansion des requêtes améliore les performances, et dans cette section, que l'emploi des URLs améliore les résultats. Dans la section suivante, nous présenterons les résultats de la combinaison de ces deux facteurs.

4 Expansion de requêtes et de documents

À ce niveau, nous avions le choix entre l'expansion des requêtes avec le *feedback* composé uniquement par le contenu des premiers tweets restitués ou bien avec le

feedback composé par le contenu des tweets et des URLs ensemble.

Le tableau 4.9 montre les résultats de l'emploi du contenu des tweets uniquement dans l'expansion et du contenu des tweets et des URLs dans la restitution (première de nos possibilités). Nous avons testé trois configurations : les deux premières se basent sur le modèle vectoriel comme modèle de restitution. La différence réside au niveau de l'expansion. (i) Dans un premier temps nous calculons les poids des termes avec TF.IDF (« VSM-Rocch(TF.IDF)-Tweets+URL ») et (ii) dans un deuxième temps avec BM25 (« VSM-Rocch(BM25)-Tweets+URL »). (iii) La troisième configuration emploie le modèle BM25 dans la restitution (« BMXx25fb-Reqexp-Tweets+URL »). Pour le run « BMXx25fb-Reqexp-Tweets+URL », nous avons initialisé les paramètres en fonction du champ de restitution comme expliqué dans le paragraphe précédent.

Run	Modèle	Requête étendue	Champ utilisé	P@30	Rappel	MAP
VSM-Rocch(TF.IDF)-Tweets	VSM	Rocch(TF.IDF)	Tweets	0,3232	0,6822	0,2211
VSM-Rocch(TF.IDF)-Tweets+URL	VSM	Rocch(TF.IDF)	Tweets+URL	**0,3894***	**0,7506***	**0,2777***
VSM-Rocch(BM25)-Tweets	VSM	Rocch(BM25)	Tweets	0,3311	0,6764	0,2304
VSM-Rocch(BM25)-Tweets+URL	VSM	Rocchio(BM25)	Tweets+URL	**0,3960***	**0,7524***	**0,2869***
BMX25fb-Reqexp-Tweets	BMX25fb	Reqexp	Tweets	**0,3571**	0,6369	0,2300
BMXx25fb-Reqexp-Tweets+URL	BMXx25fb	Reqexp	Tweets+URL	0.3712*	**0,6294**	**0.2333**

Tableau 4.9 – Emploi des tweets et des URLs et expansion de requêtes uniquement à partir des tweets.* montre une amélioration significative par rapport au run précédent.

La première observation que nous pouvons tirer est que les runs considérant les URLs et les tweets en plus de l'expansion de requêtes sont meilleurs que les runs considérant les tweets. Toutefois, l'intensité de cette amélioration dépend du modèle de restitution. Nous pouvons remarquer des améliorations importantes avec le modèle vectoriel sur les trois mesures. Cependant, les améliorations avec le modèle BM25 sont moins importantes (notons même une dégradation du rappel).

Lorsqu'on utilise le modèle vectoriel pour la restitution, nous remarquons encore une fois que la pondération des termes d'expansion avec BM25 donne de meilleurs résultats qu'avec TF.IDF. Nous avons comparé les tweets pertinents des deux runs « VSM-Rocch(BM25)-Tweets+URL » et « VSM-Rocch(BM25)-Tweets ». L'emploi des URLs a résulté des améliorations de 11,2 %, 19,6 % et de 24,5 % respectivement sur le rappel, la P@30 et la MAP. Le run « VSM-Rocch(BM25)-Tweets+URL » contient 809 nouveaux tweets pertinents (17 %) et a échoué à restituer 259 (6 %) tweets qui existaient dans « VSM-Rocch(BM25)-Tweets ». Nous remarquons ainsi de nouveau le double effet de l'emploi des URLs pour les mêmes raisons précédemment expliquées. Toutefois, la quantité de nouveaux tweets pertinents dépasse la quantité

des tweets non restitués.

Concernant le modèle BM25, l'emploi des URLs a amélioré de manière significative uniquement la P@30 (4,0 %). Les deux runs « BMXx25fb-Reqexp-Tweets+URL » et « BMX25fb-Reqexp-Tweets » contiennent pratiquement le même nombre de tweets pertinents. Cependant, ces deux runs diffèrent d'un certain nombre de tweets pertinents (de l'ordre de 650 tweets). Nous pouvons ainsi conclure que, avec le modèle BM25, le double effet de l'emploi des URLs est plus important. Il n'y a pas ainsi d'effet positif sur le rappel. Cependant, ce facteur améliore considérablement le rang des documents pertinents (effet positif sur la précision).

Finalement, nous avons voulu tester l'impact de l'emploi des URLs même dans l'expansion de requêtes (deuxième de nos propositions citées au début de la section 4). En d'autres termes, les termes d'expansion seront sélectionnés à partir du contenu des tweets et des URLs des résultats formant le *feedback*. Le tableau 4.10 montre les résultats de l'emploi du contenu des tweets et des URLs dans l'expansion de requêtes avec le modèle vectoriel (Rocchio(BM25)(T+U)) et avec le modèle BM25 (Reqexp(T+U)). (T+U) indique l'emploi de Tweets et des URLs dans l'expansion.

Run	Modèle	Requête étendue	Champ utilisé	P@30	Rappel	MAP
VSM-Rocchio(BM25)-Tweets+URL	VSM	Rocchio(BM25)	Tweets+URL	**0,3960**	**0,7524**	**0,2869**
VSM-Rocchio(BM25)(T+U)-Tweets+URL	VSM	Rocchio(BM25)(T+U)	Tweets+URL	0,2633	0,5892	0,1841
BMXx25fb-Reqexp-Tweets+URL	BMXx25fb	Reqexp	Tweets+URL	0,3712	**0,6294**	**0,2333**
BMXx25fb-Reqexp(T+U)-Tweets+URL	BMXx25fb	Reqexp(T+U)	Tweets+URL	**0,3966***	0,5208	0,2143

Tableau 4.10 – Emploi des tweets et des URLs pour l'expansion et pour la restitution. * montre une amélioration significative par rapport au run précédent.

Pour le modèle vectoriel, nous avons remarqué une dégradation remarquable en employant les URLs dans l'expansion. Cependant, avec le modèle BM25, l'emploi des URLs conduit à des effets différents. D'une part, le rappel et la MAP se sont dégradés considérablement. D'autre part, la P@30 s'est améliorée pour atteindre le meilleur score parmi toutes nos configurations précédentes. Ces observations sont expliquées ainsi : le fait de considérer les URLs dans l'expansion a dévié le sens des requêtes et généré des dégradations. Cependant, cette dégradation s'est transformée en amélioration, en particulier avec le modèle BM25, étant donné que ce modèle exploite le *feedback* (composé par les tweets et les contenus des URLs) dans la nouvelle restitution. Ceci a conduit, d'une part, à une perte importante dans le nombre de tweets pertinents restitués, mais, d'autre part, à une mise en valeur maximale des tweets pertinents restitués (reclassement vers la tête de la liste).

5 Discussion

La conclusion principale des expérimentations de ce chapitre est que l'expansion de requêtes et **la prise en compte des contenus des URL dans la restitution** paraissent indispensables pour la recherche des microblogs, que ce soit au niveau du rappel ou la précision. Les URLs permettent non seulement de fournir des informations supplémentaires pour les internautes, mais présentent également un vocabulaire très utile pour les moteurs de recherche, qui sera utilisé pour mesurer la pertinence du microblog vis-à-vis d'un besoin en information. **L'expansion de requêtes** permet de mieux représenter les besoins d'information (améliore le rappel), et de mettre en valeur les tweets pertinents (améliore la précision). **La pondération des termes de la requête**, elle aussi, joue un rôle très important dans l'amélioration des résultats. Elle permet de mettre en valeur les tweets pertinents en relation avec les termes importants des requêtes (améliore la précision). Ceci est aperçu, d'une part, en regardant les runs utilisant l'expansion de requêtes avec les articles des actualités (tableau 4.1), avec et sans pondération, ou en comparant les runs se basant sur TF.IDF avec les runs se basant sur BM25 pour pondérer les termes d'expansion (tableau 4.5).

Concernant le modèle de restitution, la supériorité d'un modèle par rapport à un autre dépend des facteurs supplémentaires utilisés et aussi des résultats à analyser (rappel ou précision). De manière générale, **BM25 obtient de meilleures précisions et VSM obtient les meilleurs rappels**.

Le paramétrage est crucial pour le modèle BM25. Le fait d'initialiser k_1 à 0,1 et b à 0 (on ne prend pas en compte la normalisation par la longueur) lui permet de prendre un avantage par rapport au VSM. Cependant, l'emploi du *feedback* dégrade ses résultats, que ce soit avec ou sans paramétrage, à moins que ce *feedback* soit accompagné d'une expansion de requêtes (tableau 4.6). Dans ce cas, nous arrivons à obtenir les meilleures P@30. Cette dernière observation reste valide que ce soit avec ou sans l'emploi des URLs.

Les améliorations avec le modèle vectoriel sont plus équilibrées. En employant des facteurs supplémentaires (Rocchio ou URLs), nous apercevons des améliorations sur le rappel ou bien sur la précision. Concernant l'expansion, la pondération des termes avec BM25 ou avec TF.IDF fait ressortir, dans la plupart des cas, les mêmes termes d'expansion. Cependant, la pondération de ces termes avec BM25 permet de restituer plus de tweets pertinents et de les ranger de façon plus pertinente qu'avec TF.IDF.

Concernant l'emploi des URLs dans l'appariement, l'impact de ce facteur dépend du modèle de restitution. De manière générale ce facteur améliore toutes les mesures avec le modèle vectoriel. Cependant, il améliore uniquement la P@30 avec le modèle probabiliste (tableau 4.9). L'emploi des URLs, en plus des

tweets, dans l'expansion de requêtes n'a pas montré un effet positif avec le modèle vectoriel. Cependant, il a permis d'avoir la meilleure P@30 parmi toutes nos expérimentations, même s'il a dégradé considérablement le rappel (tableau 4.10).

Nous avons comparé les deux meilleurs runs au niveau de la P@30 « VSM-Rocch(BM25)-Tweets+URL » et « BMXx25fb-Reqexp(T+U)-Tweets+URL », avec les résultats officiels de la tâche Microblog de TREC 2012 (Ounis et al., 2012). L'évaluation des résultats officiels des participants de la tâche microblogs 2012 est réalisée en considérant uniquement les tweets hautement pertinents. Le tableau 4.11 montre les résultats de ces deux runs considérant les tweets hautement pertinents. D'ailleurs, ces deux runs ont conservé leur avantage par rapport aux autres runs, même avec cette considération.

Run	Modèle	Requête étendue	Champ utilisé	P@30	Rappel	MAP
VSM-Rocch(BM25)-Tweets+URL	VSM	Rocchio(BM25)	Tweets+URL	**0,2531**	**0,7722**	**0,2264**
BMXx25fb-Reqexp(T+U)-Tweets+URL	BMXx25fb	Reqexp(T+U)	Tweets+URL	**0,2531**	0,6087	0,2113

Tableau 4.11 – Résultats des meilleurs runs avec les tweets hautement pertinents.

Le tableau 4.12 montre les résultats des deux meilleurs runs officiels des participants de 2012. Chacun de nos deux runs nous aurait permis de nous placer à la 2ème position des participants selon la P@30. Nous n'avons bien évidemment considéré que les runs automatiques dans cette comparaison.

Groupe	Run	P@30	MAP
HIT MTLAB	hitURLrun3	**0.2701**	**0.2642**
IRIT	VSM-Rocch(BM25)-Tweets+URL	0.2531	0.2264
IRIT	BMXx25fb-Reqexp(T+U)-Tweets+URL	0.2531	0.2113
HIT MTLAB	hitLRrun1	0.2446	0.2411

Tableau 4.12 – Comparaison avec les résultats officiels de TREC 2012

Run	Modèle	Requête étendue	Champ utilisé	P@30	Rappel	MAP
VSM-Rocch(BM25)-Tweets+URL	VSM	Rocchio(BM25)	Tweets+URL	0.4701	0.8752	0.4700

Tableau 4.13 – Emploi des tweets pour l'expansion et des tweets et des URLs pour la restitution sur les topics de TREC 2011.

Le tableau 4.13 montre les résultats du run « VSM-Rocch(BM25)-Tweets+URL » sur les requêtes de 2011. Les jugements des runs officiels de la tâche de 2011 sont réalisés en considérant tous les tweets pertinents. Ce run nous aurait permis de nous

placer à la 1ère position des participants selon la P@30 (tableau 4.14). Notons la présence de notre run officiel (Damak et al., 2011).

Groupe	Run	P@30	MAP
IRIT	VSM-Rocch(BM25)-Tweets+URL coupé à 30	**0,4701**	**0,2966**
isi	isiFDL	0,4551	0,1923
FUB	DFReeKLIM30	0,4401	0,2348
CLARITY_DCU	clarity1	0,4211	0,2139
Purdue_IR	myrun2	0,3993	0,2003
IRIT	Run officiel (Damak et al., 2011)	0,2565	0,1940

Tableau 4.14 – Comparaison avec les résultats officiels de TREC 2011

6 Bilan

Dans ce chapitre nous avons proposé quelques méthodes pour améliorer la qualité des résultats d'une tâche de recherche de micrblogs. Nous avons exploité les articles des actualités et la base lexicale WordNet pour étendre les requêtes. En outre, nous avons analysé l'impact des techniques de RI classique sur ce nouveau type de document. Nous avons particulièrement testé le modèle vectoriel et le modèle probabiliste. Avec le modèle vectoriel, nous avons étendu les requêtes avec la technique de Rocchio. Avec le modèle BM25, nous avons utilisé son mécanisme naturel de feedback et d'expansion. Ensuite, nous avons testé l'effet de l'emploi du contenu des URLs en complément du contenu des tweets. L'emploi des URLs était avantageux uniquement pour la restitution (et non pour l'expansion). De manière générale, le modèle BM25 fournit de meilleures précisions. Le modèle vectoriel arrive à restituer plus de tweets pertinents. L'expansion de requêtes améliore le rappel et la précision avec le modèle vectoriel. Cependant, avec le modèle BM25, son effet positif est clair surtout sur la P@30. L'emploi des URLs pour la restitution est primordial. En contrepartie, elles n'ont pas montré d'intérêt pour l'expansion des requêtes.

Chapitre 5

Analyse des facteurs de pertinence de l'état de l'art

1 Introduction

Nous avons montré dans l'état de l'art que les approches de RI dans les microblogs emploient une multitude de critères de pertinence : critères de fraîcheur, critères sur les auteurs, critères du réseau social, des différentes données présentes dans microblogs (hashtags, URLs)..., en plus de la pertinence textuelle. Ces différents critères sont concrètement pris en compte dans les modèles de recherche proposés en combinant *des facteurs de pertinence* pour mesurer la pertinence des *microblogs* vis-à-vis d'un besoin en information. Par exemple, en considérant le critère importance de l'auteur, les facteurs de pertinence associés pourraient être le nombre de microblogs de l'auteur et le nombre de ses abonnés (Nagmoti et al., 2010). Nous pouvons également considérer le nombre de fois qu'un utilisateur a été mentionné ou bien le score de l'auteur selon un algorithme semblable à PageRank basé sur des relations de rediffusion des messages (Ben Jabeur et al., 2011).

Même si les intuitions justifiant l'emploi de ces facteurs de pertinence encouragent leur exploitation, la valeur réelle de ces facteurs de pertinence n'a jamais été démontrée. En outre, nous avons montré dans le chapitre 3 (analyse de défaillances) que, dans la recherche de microblogs, la plupart des problèmes remontés par les modèles de recherche sont des problèmes de vocabulaire (Damak, 2013), problèmes pour lesquels nous avons proposé des solutions dans le chapitre 4. La question qui se pose maintenant est : si le modèle arrive à restituer tous les microblogs pertinents, est-ce que l'emploi de facteurs de pertinence supplémentaires permet de promouvoir les microblogs pertinents parmi l'ensemble des résultats ?

Dans ce chapitre, nous évaluons l'impact réel des facteurs de pertinence souvent utilisés dans les approches de l'état de l'art sur la qualité des microblogs restitués vis-à-vis d'un besoin en information.

2 Description des facteurs de pertinence

Nous décrivons dans cette section les 14 facteurs de pertinence que nous considérons, classés par groupe. Nous considérons cinq groupes de facteurs de pertinence : celui lié au contenu des *microblogs*, celui lié à leur hypertextualité, celui qui se base sur les hashtags, celui lié aux auteurs des *microblogs* et enfin un groupe de facteurs relatifs à la qualité des *microblogs*. Nous cherchons à évaluer l'impact de ces facteurs de pertinence comme précédemment, c'est-à-dire sur l'évaluation de la pertinence d'un tweet par rapport à une requête.

Nous utiliserons les notations suivantes dans la suite :

- q est la requête (composée de mots-clés 'topic' et caractérisée par une date),
- C_q est le corpus des tweets publiés avant la date de la requête,
- T_q est l'ensemble des tweets restitués par un moteur de recherche donné calculant la pertinence par rapport à q uniquement sur le contenu des tweets ($T_q \subseteq C_q$),
- t est un tweet $\in T_q$ sur lequel on applique le facteur de pertinence.

2.1 Facteurs de pertinence basés sur le contenu des tweets

Nous avons considéré quatre facteurs de pertinence relatifs à certaines spécificités de contenu des microblogs : la popularité d'un tweet (5.1), la longueur faible des tweets (5.2), la correspondance des termes entre les tweets et la requête (5.3) et la qualité du langage d'écriture du tweet (5.4).

- Popularité du tweet (Duan et al., 2010) : ce facteur de pertinence estime la popularité d'un tweet dans T_q. On suppose qu'un tweet est populaire si on trouve plusieurs autres tweets ayant un contenu similaire. La similarité $sim(t_i, t_j)$ entre chaque paire de tweets est calculée avec un modele vectoriel qui prend également en compte la fréquence des termes de la requête dans le tweet (Cohen et al., 2007). On note le vecteur contenant les termes du tweet courant par t_i. Ce facteur de pertinence est calculé de la manière suivante :

$$f_1(t_i, q) = \frac{\sum_{t_j \in T_q, i \neq j} sim(t_i, t_j)}{|T_q| - 1} \tag{5.1}$$

- Longueur du tweet (Duan et al., 2010) : intuitivement, plus une phrase est longue, plus elle contient de l'information. Nous avons calculé ce facteur de pertinence en comptant le nombre de termes dans un tweet. On note $l(t_i)$ le nombre de termes dans un tweet t_i dans T_q. Ce facteur de pertinence est calculé de la manière suivante :

$$f_2(t_i) = \frac{l(t_i)}{max_{t_j \in T_q} l(t_j)} \tag{5.2}$$

– Correspondance exacte des termes : ce facteur favorise les tweets qui contiennent les termes de la requête q. La valeur $nb(t_i, q)$ correspond au nombre de termes en commun entre t_i et q :

$$f_3(t_i, q) = \frac{nb(t_i, q)}{max_{t_j \in T_q} nb(t_j, q)} \tag{5.3}$$

– Qualité du langage (Duan et al., 2010) : ce facteur de pertinence représente la proportion des termes qui existent dans un dictionnaire [1] par rapport à tous les termes du tweet t_i. La valeur $dic(term)$ est binaire : 1 si le terme existe dans le dictionnaire, 0 sinon :

$$f_{14}(t_i) = \frac{\sum_{term \in t_i} dic(term)}{l(t_i)} \tag{5.4}$$

2.2 Facteurs de pertinence basés sur l'hypertextualité

Nous considérons trois facteurs de pertinence additionnels qui peuvent indiquer la qualité de l'information publiée dans les tweets :

– Présence d'une URL dans le tweet (Nagmoti et al., 2010 ; Zhao et al., 2011) : partager des URLs est une manière de confirmer l'information publiée dans un tweet. Ceci permet également d'attirer l'attention sur un contenu présent sur le web. Ainsi, on fait l'hypothèse que la présence d'une URL indique que le tweet a un caractère informatif renforcé. Ce facteur de pertinence est binaire :

$$f_4(t_i) = \begin{cases} 1 & \text{si } t_i \text{ contient une URL} \\ 0 & \text{sinon} \end{cases} \tag{5.5}$$

– Fréquence des URLs (Zhao et al., 2011) : compte le nombre d'URLs publiées dans un tweet t_i :

$$f_5(t_i, q) = \frac{|\{w \in t_i / isURL(w)\}|}{max_{t_j \in T_q} |\{w \in t_j / isURL(w)\}|} \tag{5.6}$$

– Fréquence de l'URL dans le corpus : ce facteur de pertinence permet de calculer la popularité des URLs publiées dans un tweet t_i dans le corpus C_q. On note par $freq(url)$ le nombre de fois ou une URL apparaît dans le corpus C_q :

$$f_6(t_i, q) = \frac{\sum_{url \in t_i} freq(url)}{max_{t_j \in T_q} \sum_{url \in t_j} freq(url)} \tag{5.7}$$

2.3 Facteurs de pertinence basés sur les hashtags

– Présence de hashtag (Metzler et Cai, 2011).

$$f_7(t_i) = \begin{cases} 1 & \text{si } t_i \text{ contient un hashtag} \\ 0 & \text{sinon} \end{cases} \tag{5.8}$$

1. `http://code.google.com/p/language-detection/`

– Fréquence de hashtags du tweet (Duan et al., 2010). On note la fréquence d'un hashtag dans le corpus C_q par $freq(h)$:

$$f_8(t_i) = \sum_{h \in t_i} freq(h) \tag{5.9}$$

– Hashtags de la requête dans le tweet : calcule le nombre de termes d'une requête q qui apparaissent sous forme d'un hashtag dans un tweet t_i.

$$f_9(t_i, q) = \frac{|\{w \in q / \#w \in t_i\}|}{max_{t_j \in T_q} |\{w \in q / \#w' \in t_j\}|} \tag{5.10}$$

2.4 Facteurs de pertinence basés sur la popularité des auteurs

Afin de tenir compte de la popularité des auteurs, nous avons considéré deux facteurs de pertinence spécifiques aux auteurs de microblogs.

– Nombre de tweets de l'auteur (Nagmoti et al., 2010) : l'objectif de ce facteur de pertinence est de valoriser les tweets publiés par des auteurs actifs par rapport aux tweets publiés par des auteurs moins actifs. On note par $a(t_i)$ l'auteur du tweet t_i et $N(a(t_i))$ le nombre de tweets publiés par l'auteur du tweet t_i dans le corpus C_q.

$$f_{10}(t_i) = N(a(t_i)) \tag{5.11}$$

– Nombre de citations de l'auteur (Zhao et al., 2011) : plus un auteur est mentionné, plus il est populaire. $M(a(t_i))$ indique combien de fois un auteur du tweet t_i a été mentionné dans le corpus C_q :

$$f_{11}(t_i) = M(a(t_i)) \tag{5.12}$$

2.5 Facteurs de pertinence relatifs à la qualité des tweets

Nous avons également analysé deux autres critères particularisant les tweets :

– Retweet (Metzler et Cai, 2011). Si un utilisateur aime un tweet publié par un de ses amis, il va probablement le commenter et le partager de nouveau. Dans ce cas, le nouveau message va être précédé par *RT* (ou marqué en tant que retweet).

$$f_{12}(t_i) = \begin{cases} 1 & \text{si } t_i \text{ contient RT} \\ 0 & \text{sinon} \end{cases} \tag{5.13}$$

– Fraîcheur (Magnani et al., 2012). C'est la différence entre la date de la publication du tweet t_i et la date de la soumission de la requête q, mesurée en secondes. $tmp(t_i)$ est le timestamp en seconde d'un tweet t_i (c'est-à-dire sa date de publication).

$$f_{13}(t_i, q) = \frac{tmp(q) - tmp(t_i)}{max_{t_j \in T_q} tmp(q) - tmp(t_j)} \tag{5.14}$$

3 Méthodologie

Notre analyse est réalisée en trois phases : tout d'abord, nous avons évalué les facteurs de pertinence en nous basant sur les distributions de leurs scores, ensuite, en combinant linéairement leurs scores avec le score du modèle de restitution, et finalement en utilisant les techniques de sélection d'attributs pour des algorithmes d'apprentissage.

3.1 Étude de la distribution des scores

L'intuition derrière cette étude est que les facteurs de pertinence reflétant la pertinence distinguent les tweets pertinents des non pertinents. Ces facteurs de pertinence n'aurons pas le même comportement avec les tweets pertinents et les tweets non pertinents. Pour évaluer un facteur, nous avons observé la distribution de ses scores dans les tweets. Si la distribution des scores d'un facteur de pertinence est la même pour les tweets pertinents et non pertinents, ce facteur ne permettra pas ainsi de différencier les deux classes de tweets, et ne sera pas considéré comme facteur utile à cette tâche. Dans le cas contraire, lorsque la distribution des scores d'un facteur de pertinence est différente entre les tweets pertinents et non pertinents, ce facteur permettra dans ce cas de différencier les deux classes de tweets, et il sera par conséquent considéré comme facteur utile.

3.2 Étude par la combinaison linéaire des scores

Dans un deuxième temps, nous avons évalué l'impact direct de chacun des facteurs de pertinence sur la qualité des résultats. Nous avons ainsi combiné linéairement le score de chaque facteur de pertinence avec le score du modèle de restitution textuel employé. L'intuition derrière cette étude est que les facteurs de pertinence utiles vont promouvoir les tweets pertinents dans l'ensemble des résultats et ceci, de manière générale, va améliorer la qualité des résultats.

Ensuite, nous avons testé la complémentarité des facteurs de pertinence : peuvent-ils se compléter afin d'améliorer les résultats ? L'idée est de voir si les facteurs de pertinence ont des comportements différents lorsqu'ils sont combinés avec d'autres facteurs de pertinence, par rapport à leur comportement lorsqu'ils sont employés seuls. Afin d'étudier ce dernier point, il faudrait effectuer toutes les combinaisons possibles des facteurs de pertinence entre eux, ceci impliquant un nombre très important de runs ($C_{14}^2 + C_{14}^3 + C_{14}^4 + C_{14}^5 + C_{14}^6 + C_{14}^7 + C_{14}^8 + C_{14}^9 + C_{14}^{10} + C_{14}^{11} + C_{14}^{12} + C_{14}^{13} = 16\,368$ runs).

Afin de réduire le nombre de cas à prendre en compte, nous avons décidé d'observer le comportement des différents groupes qu'ils forment :

- Groupe G1 lié au contenu des tweets, composé des facteurs de pertinence f_1, f_2 et f_{14},
- Groupe G2 lié à l'hypertextualité, composé des facteurs de pertinence f_4, f_5 et f_6,
- Groupe G3 lié aux hashtags publiés dans tweets, composé des facteurs de pertinence f_7, f_8, et f_9.
- Groupe G4 lié aux auteurs des tweets, composé des facteurs de pertinence f_{10}, et f_{11},
- Groupe G5 lié aux critères qualitatifs des tweets, composé des facteurs de pertinence f_{12} et f_{13}.

3.3 Étude avec les techniques de sélection d'attributs

Le problème de l'étude précédente réside dans la sélection des groupes des facteurs de pertinence. Tant que nous n'avons pas essayé toutes les combinaisons des facteurs de pertinence possible, il est impossible de juger convenablement la complémentarité des facteurs de pertinence, et d'identifier les meilleures combinaisons. En outre, l'étude précédente se base simplement sur des combinaisons de scores.

Dans cette troisième étude plus approfondie, nous nous sommes ainsi appuyés sur des techniques de sélection d'attributs pour déterminer les meilleurs facteurs de pertinence à considérer dans une tâche de recherche de microblogs.

Les techniques de sélection d'attributs visent à identifier et enlever le maximum d'information redondante et non pertinente en amont d'un processus à base d'apprentissage (Hall et Holmes, 2003). Elles permettent également de sélectionner de manière automatique les sous-ensembles de facteurs de pertinence permettant d'avoir les meilleurs résultats.

Cette phase a fait ressortir plusieurs ensembles de facteurs. Ensuite, nous avons évalué l'efficacité de ces ensembles en les appliquant sur des techniques d'apprentissage dans un contexte de recherche de microblogs.

4 Expérimentations

4.1 Étude par la distribution des scores

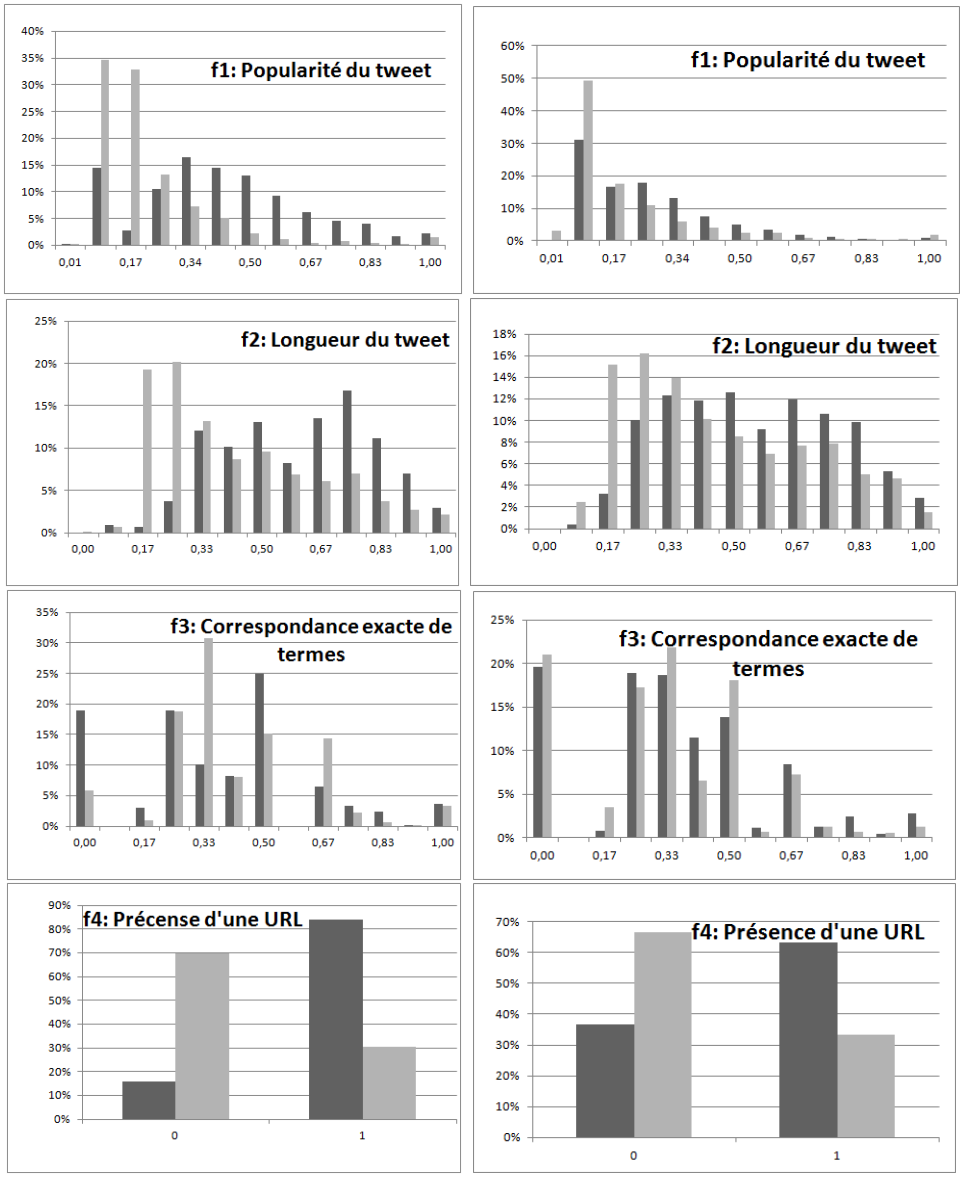
4.1.1 Cadre expérimental

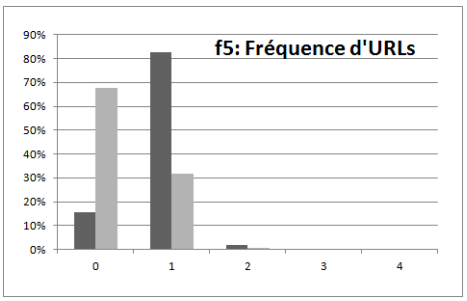
Nous nous sommes basés sur le modèle vectoriel comme modèle de restitution. Les scores des facteurs de pertinence sont ensuite calculés pour chaque tweet résultat.

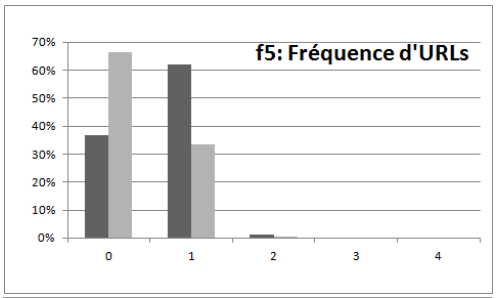
Nous avons utilisé les requêtes des deux éditions 2011 et 2012 de la tâche microblog de TREC. Dans les expérimentations de cette étude, nous considérons les tweets moyennement pertinents et hautement pertinents (*qrels all-rel*) L'ensemble

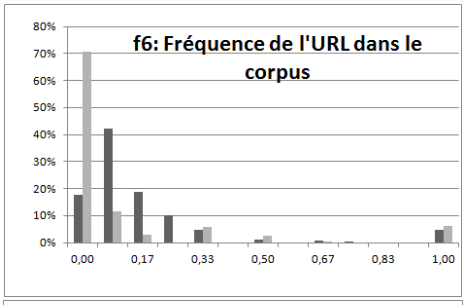
des tweets à analyser est construit de la manière suivante : d'abord, nous avons sélectionné uniquement les requêtes ayant au moins 100 tweets pertinents (ce qui représente 14 requêtes de 2011 et 13 requêtes de 2012). Nous avons fait ce choix pour avoir un nombre suffisant de tweets à étudier. Pour chacune d'entre elles, nous avons gardé tous les tweets pertinents en nous référant aux jugements de pertinence. Pour chaque requête, nous avons ajouté le même nombre de tweets non pertinents que de tweets pertinents. Les tweets non pertinents sont sélectionnés en fonction de leurs scores du modèle vectoriel. Nous avons gardé ceux ayant les scores les plus importants. Les tweets de toutes les requêtes ont été fusionnés pour tracer la distribution globale dans les figures qui suivent.

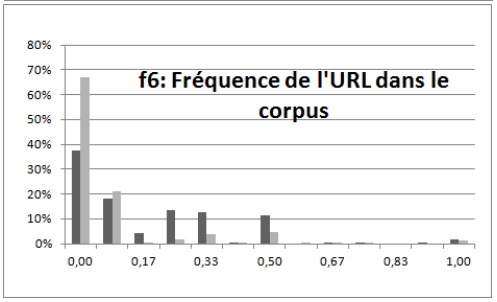
4.1.2 Résultats

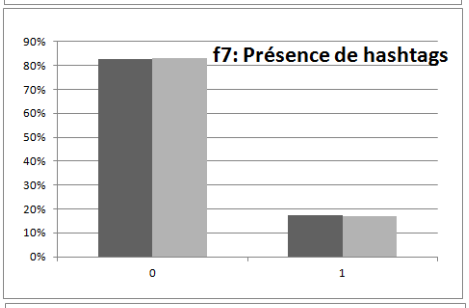


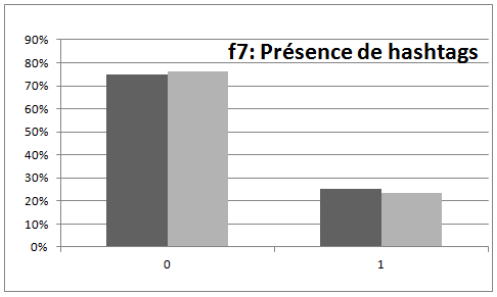

f5: Fréquence d'URLs
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0
1
2
3
4

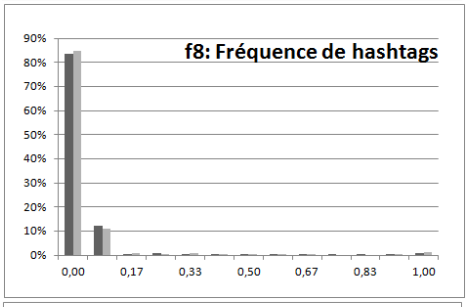

f5: Fréquence d'URLs
70%
60%
50%
40%
30%
20%
10%
0%
0
1
2
3
4

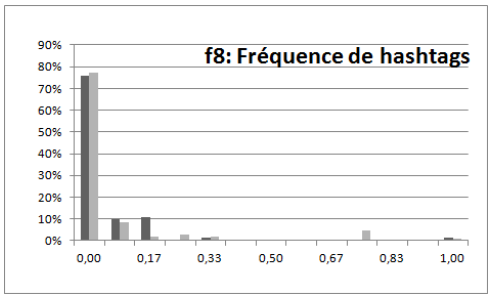

f6: Fréquence de l'URL dans le corpus
80%
70%
60%
50%
40%
30%
20%
10%
0%
0,00
0,17
0,33
0,50
0,67
0,83
1,00

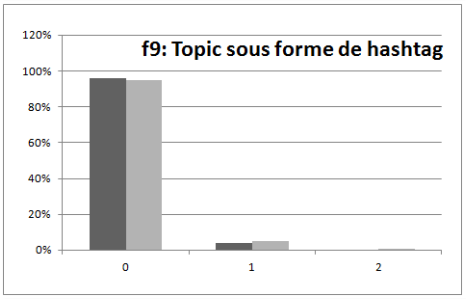

f6: Fréquence de l'URL dans le corpus
80%
70%
60%
50%
40%
30%
20%
10%
0%
0,00
0,17
0,33
0,50
0,67
0,83
1,00

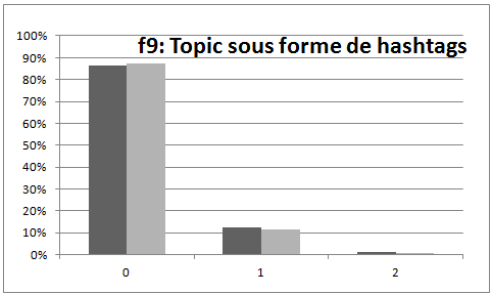

f7: Présence de hashtags
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0
1

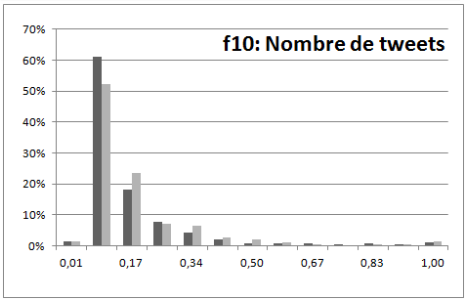

f7: Présence de hashtags
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0
1

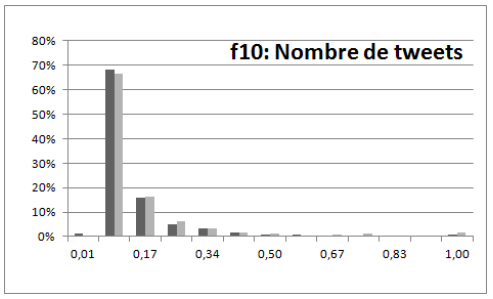

f8: Fréquence de hashtags
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0,00
0,17
0,33
0,50
0,67
0,83
1,00


f8: Fréquence de hashtags
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0,00
0,17
0,33
0,50
0,67
0,83
1,00


f9: Topic sous forme de hashtag
120%
100%
80%
60%
40%
20%
0%
0
1
2


f9: Topic sous forme de hashtags
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0
1
2


f10: Nombre de tweets
70%
60%
50%
40%
30%
20%
10%
0%
0,01
0,17
0,34
0,50
0,67
0,83
1,00


f10: Nombre de tweets
80%
70%
60%
50%
40%
30%
20%
10%
0%
0,01
0,17
0,34
0,50
0,67
0,83
1,00

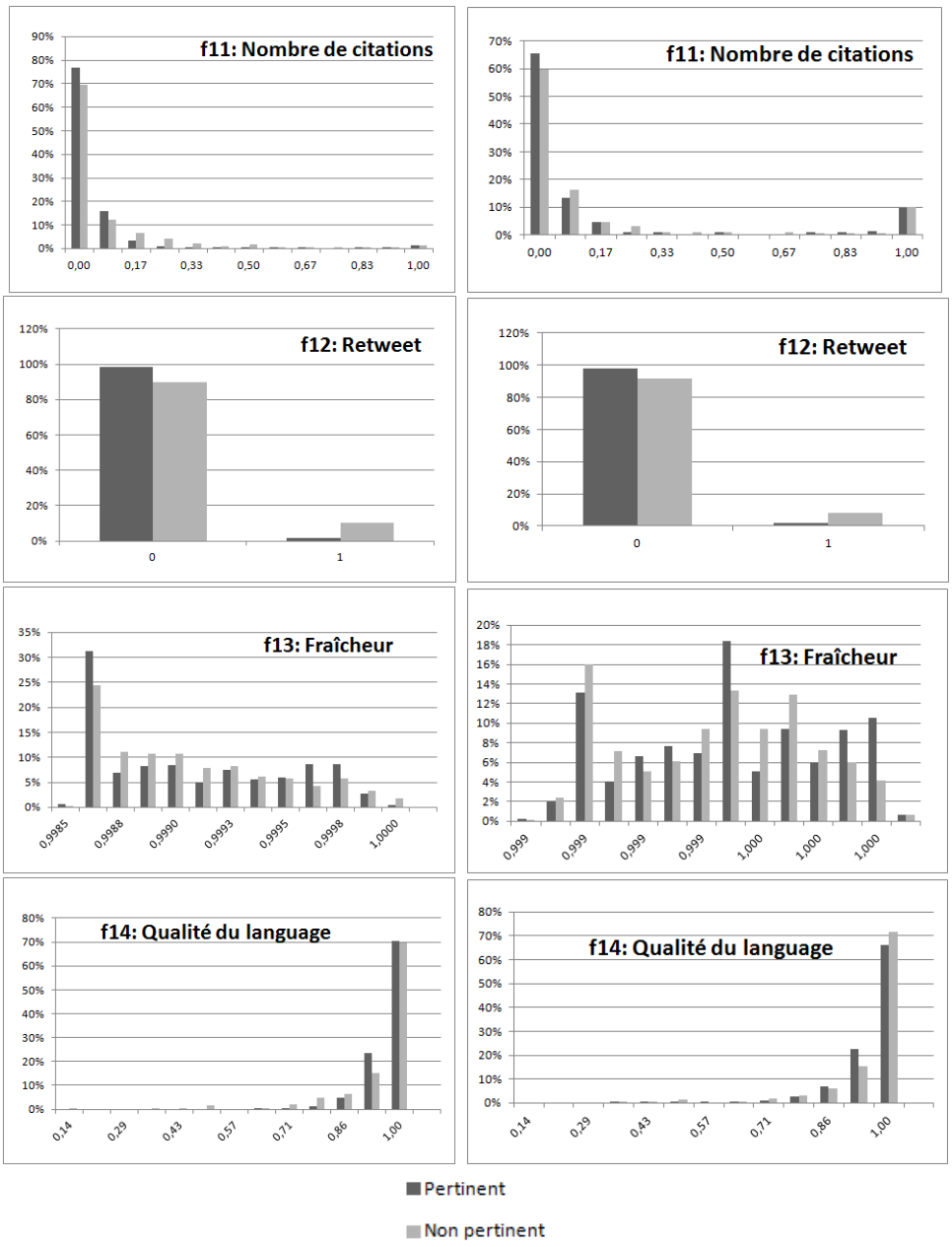


FIGURE 5.1 – Distribution des scores des tweets pertinents et des tweets non pertinents (requêtes de 2011 à gauche et celles de 2012 à droite).

La figure 5.1 montre la distribution des scores de tweets pertinents et des tweets non pertinents pour cette première étude. Les intervalles ont été calculés avec la loi de Sturges (1926). À part le facteur fraîcheur (f_{13}), aucune différence dans les distributions entre les requêtes de 2011 et de 2012 n'est constatée.

Nous pouvons observer que les facteurs de pertinence popularité du tweet (f_1), longueur du tweet (f_2), correspondance exacte des termes (f_3), présence d'URL (f_2), fréquence d'URLs (f_5), importance d'URLs (f_6) et fraîcheur (f_{13}) ne présentent pas la même distribution des scores entre les tweets pertinents et les tweets non per-

tinents. Ces critères obtiennent leurs meilleurs scores avec les tweets pertinents et reflètent probablement ainsi la pertinence. La différence entre les deux populations de scores (des tweets pertinents et des tweets non pertinents) est statiquement significative selon le test t pairé et bilatéral avec $p < 0,05$.

4.2 Étude par la combinaison linéaire des scores

Dans cette section, nous comparons tout d'abord l'apport des différentes facteurs de pertinence pour raffiner le processus de RI sur les *microblogs*. Puis, nous positionnons nos résultats par rapport aux résultats officiels de la tâche Microblog à TREC 2011 et 2012. Enfin, nous généralisons nos conclusions en faisant abstraction du moteur de recherche Lucene utilisé jusqu'alors (Damak et al., 2012).

4.2.1 Cadre expérimental

Le score final d'un tweet (équation 5.16) est calculé en combinant le score du modèle vectoriel et les scores des facteurs de pertinence (équation 5.15). Le score facteurs de pertinence est calculé par une combinaison linéaire. On réalise différentes normalisations de sorte que $f_n(t_i, q) \in [0, 1]$ et $f_{sources}(t_i, q) \in [0, 1]$. Nous avons exclu le facteur de pertinence f_3 (*correspondance exacte de termes*) des sources à évaluer afin de l'utiliser par la suite pour la généralisation des résultats. Cette source est nommée dans les expérimentations « *Base* ». Nous avons utilisé les requêtes des deux éditions 2011 et 2012 de la tâche Microblog.

$$f_{sources}(t_i, q) = f(f_1(t_i, q), f_2(t_i, q), f_4(t_i, q), f_5(t_i), ..., f_{12}(t_i), f_{13}(t_i), f_{14}(t_i)) \quad (5.15)$$

$$score(t_i, q) = \alpha * VSM(t_i, q) + (1 - \alpha) * f_{sources}(t_i, q) \qquad \alpha \in [0, 1] \qquad (5.16)$$

Dans nos expérimentations, nous considérons les tweets moyennement pertinents et hautement pertinents (*qrels all-rel*). Les résultats présentés par la suite sont évalués en fonction d'un classement sur le score de pertinence, contrairement à la tâche Microblog de l'édition de 2011 qui évalue en réordonnant les résultats sur la date des tweets au préalable, ce qui ne rend pas compte de la qualité des facteurs de pertinence. Pour nos analyses, nous utilisons les 5000 premiers résultats renvoyés par Lucene.

4.2.2 Résultats

4.2.2.1 Combinaison source par source. Le tableau 5.1 montre les résultats obtenus en considérant les facteurs de pertinence décrits dans la section 2 un par un. Un astérisque indique que la différence est statistiquement significative selon le test t de Student (1908) pairé et bilatéral avec $p < 0,05$.

Système	édition 2011		édition 2012	
	P@30	MAP	P@30	MAP
VSM	0,3544	0,3141	0,2842	0,1871
VSM + f_1	0,3027*	0,2280*	0,1966*	0,1371*
VSM + f_2	0,2701*	0,2241*	0,2729*	0,1616*
VSM + f_4	**0,3986***	**0,3348***	**0,3463***	**0,2202***
VSM + f_5	0,3517	0,3062*	0,3260*	0,2062*
VSM + f_6	0,3238	0,2570*	0,2469*	0,1645*
VSM + f_7	0,1619*	0,1067*	0,1542*	0,0905*
VSM + f_8	0,2823*	0,2078*	0,2379*	0,1517*
VSM + f_9	0,2275*	0,1903*	0,2246*	0,1478*
VSM + f_{10}	0,1850*	0,1724*	0,2040*	0,1311*
VSM + f_{11}	0,3245*	0,2475*	0,2627*	0,1625*
VSM + f_{12}	0,0299*	0,0867*	0,0362*	0,0845*
VSM + f_{13}	0,3456*	0,3134*	0,2808	0,1860*
VSM + f_{14}	0,3517*	0,3067*	0,2842	0,1813*

Tableau 5.1 – Apport de chaque facteur de pertinence par rapport au modèle vectoriel (baseline VSM).

Comme nous pouvons le constater, et de façon assez surprenante, l'utilisation de tous les critères, sauf f_4 (2011) et f_4, f_5 (pour 2012), conduit à une dégradation des résultats. Concernant f_4 (présence d'une URL dans le tweet), on observe une hausse sensible des résultats (+12,4 % sur la P@30 sur les requêtes de 2011 et +21,85 % sur les requêtes de 2012).

4.2.2.2 Combinaison de plusieurs facteurs de pertinence. Les résultats précédents ne nous permettent pas de voir les interactions entre les facteurs de pertinence. Nous évaluons maintenant les groupes des facteurs de pertinence.

Les résultats sont décrits dans les tableau 5.2. Les combinaisons qui améliorent la P@30 sont les combinaisons qui mobilisent G2, qui contient le facteur de pertinence f_4. Ceci tend à confirmer que seul le facteur de pertinence f_4 a un intérêt dans notre système. Nous pouvons remarquer également que le groupe G1 n'améliore pas les résultats lorsqu'il est utilisé seul, mais améliore les résultats lorsqu'il est combiné avec l'un des autres groupes, en particulier G2. Le contraire de cette observation se manifeste pour le groupe G3. Ce groupe dégrade les résultats en le combinant avec n'importe quel autre groupe.

Système	édition 2011		édition 2012	
	P@30	MAP	P@30	MAP
VSM	0,3544	0,3141	0,2842	0,1871
VSM + G1	0,3449*	0,2996*	0,2938*	0,1816*
VSM + G2	0,3694	0,3233	0,3249	0,2091
VSM + G3	0,1833*	0,1332*	0,1643*	0,0973*
VSM + G4	0,2197*	0,1832*	0,1876*	0,1254*
VSM + G5	0,1578*	0,1797*	0,1390*	0,1134*
VSM + G1 + G2	**0,4014***	**0,3431***	**0,3441***	**0,2235***
VSM + G1 + G3	0,2920*	0,2374*	0,2298*	0,1421*
VSM + G1 + G4	0,3374*	0,2965*	0,2864*	0,1770*
VSM + G1 + G5	0,2769*	0,2520*	0,2288*	0,1551*
VSM + G2 + G3	0,2848*	0,2382*	0,2678*	0,1659*
VSM + G2 + G4	0,3306*	0,2947*	0,3085*	0,1968*
VSM + G2 + G5	0,2973*	0,2763*	0,2740*	0,1841*
VSM + G3 + G4	0,1906*	0,1395*	0,1608*	0,0976*
VSM + G3 + G5	0,2159*	0,1793*	0,1596*	0,1111*
VSM + G4 + G5	0,2170*	0,2064*	0,1644*	0,1182*
VSM + G1 + G2 + G3	0,3623*	0,3005*	0,3029*	0,1909*
VSM + G1 + G2 + G4	0,3946*	0,3354*	0,3390*	0,2178*
VSM + G1 + G2 + G5	0,3544*	0,3113*	0,2853*	0,1994*
VSM + G1 + G3 + G4	0,2906*	0,2388*	0,2205*	0,1409*
VSM + G1 + G3 + G5	0,2804*	0,2398*	0,2094*	0,1416*
VSM + G1 + G4 + G5	0,2864*	0,2538*	0,2282*	0,1552*
VSM + G2 + G3 + G4	0,3043*	0,2545*	0,2573*	0,1623*
VSM + G2 + G3 + G5	0,3087*	0,2608*	0,2520*	0,1669*
VSM + G2 + G4 + G5	0,3252*	0,2839*	0,2644*	0,1819*
VSM + G3 + G4 + G5	0,2159*	0,1802*	0,1684*	0,1116*
VSM + G1 + G2 + G3 + G4	0,3638	0,2991	0,2959	0,1859
VSM + G1 + G2 + G3 + G5	0,3478	0,2929	0,2731	0,1822
VSM + G1 + G2 + G4 + G5	0,3517	0,3108	0,2853	0,1977
VSM + G1 + G3 + G4 + G5	0,2906	0,2409	0,2041	0,1403
VSM + G2 + G3 + G4 + G5	0,3109	0,2602	0,2491	0,1629
VSM + G1 + G2 + G3 + G4 + G5	0,3464	0,2909	0,2690	0,1790

Tableau 5.2 – Apport de chaque groupe de facteurs de pertinence et de leurs combinaisons par rapport modèle vectoriel (baseline VSM).

4.2.3 Comparaison avec les résultats officiels de TREC

Nous avons comparé notre meilleur run résultat (VSM + G1 + G2) avec les résultats officiels de la tâche Microblog de TREC 2011. Les résultats sont présentés dans le tableau 4.1. Afin que la comparaison soit équitable, seuls sont présents dans le tableau les runs officiels automatiques n'utilisant pas de source externe et fonctionnant en temps réel, c'est à dire n'utilisant pas d'information future. Nous rappelons que lors de l'évaluation officielle, les tweets doivent être ordonnés par ordre chronologique inverse. Notre run est coupé à 30 résultats afin d'éviter le biais introduit par le tri chronologique, assimilable à l'introduction d'un critère indépendant de la pertinence qui introduit un paramètre aléatoire non souhaitable. Par conséquent, les résultats du tableau 5.3 diffèrent sur la MAP par rapport au tableau 5.1. À titre informatif, les résultats sans coupe de notre run sont également présentés dans le tableau. On note l'effet négatif sur les résultats du tri chronologique des tweets, et ce sur les deux mesures.

Groupe	Run	P@30	MAP
isi	isiFDL	0,4551	0,1923
FUB	DFReeKLIM30	0,4401	0,2348
CLARITY_DCU	clarity1	0,4211	0,2139
IRIT	VSM + $G1 + G2$ coupé à 30	0,4014	0,1857
Purdue_IR	myrun2	0,3993	0,2003
IRIT	VSM + f_4 coupé à 1000	0,1272	0,1549
IRIT	Run officiel (Damak et al., 2011)	0,2565	0,1940

Tableau 5.3 – Comparaison avec les résultats officiels de TREC 2011

Le run « VSM + $G1 + G2$ coupé à 30 » nous aurait permis de nous placer à la 4^{e} position des participants selon la P@30. Ces résultats améliorent notre participation officielle dont les détails sont donnés dans (Damak et al., 2011). Sur les requêtes de 2012, le run (VSM + f_4) nous aurait permis de nous placer à la 36^{e} position des participants selon la P@30. Ceci s'explique par le fait que les participants de la tâche de 2012 ont employé d'autres moyens, en particulier l'exploitation des contenus des URLs et l'expansion de requêtes dans leurs systèmes, dont nous avons montré leurs intérêts dans le chapitre précédent, ce qui a mis la barre de la pertinence très haute.

4.2.4 Généralisation des résultats

Les résultats que nous avons obtenus et présentés dans les sections précédentes sont liés à la performance du modèle vectoriel : ce sont sur les tweets renvoyés par ce modèle que nous appliquons les facteurs de pertinence. On pourrait donc penser que le score final d'un tweet dépend fortement du score du modèle vectoriel. Nous avons donc cherché à généraliser nos résultats précédents, en mettant en place une méthodologie d'évaluation indépendante du modèle vectoriel. Pour ce faire, nous avons sélectionné 5000 tweets avec Lucene, desquels nous avons enlevé le score associé. Ensuite, nous avons ajouté à cet ensemble les tweets pertinents manquants obtenus à partir des jugements de pertinence officiels (*qrels*). Comme il semble obligatoire d'avoir au moins un facteur de pertinence basé sur le contenu de la requête, la contribution du modèle vectoriel a été remplacée par un score très simple : le pourcentage de termes de la requête présents dans le tweet (Base). Ce score correspond au critère f_3. Le score final de chaque tweet est ensuite calculé selon la formule 5.16 dans laquelle le score du modèle vectoriel est remplacé par Base. Les résultats généralisés sur l'apport des facteurs de pertinence un à un sont présentés dans le tableau 5.4.

Système	édition 2011		édition 2012	
	P@30	MAP	P@30	MAP
Base	0,2184	0,1785	0,1793	0,1001
Base + f_1	0,2034	0,1629*	0,1339	0,0762*
Base + f_2	0,1531*	0,1155*	0,1741*	0,0909*
Base + f_4	**0,2449***	**0,2019***	**0,2316***	**0,1298***
Base + f_5	0,2565*	0,1876*	0,2126	0,1216*
Base + f_6	0,2095	0,1610*	0,1816	0,1065*
Base + f_7	0,1150*	0,0687*	0,1103	0,0638*
Base + f_8	0,1755*	0,1214*	0,1586*	0,0871*
Base + f_9	0,1884*	0,1424*	0,1591	0,0957*
Base + f_{10}	0,1190*	0,0980*	0,1339*	0,0819*
Base + f_{11}	0,2054	0,1481*	0,1638	0,0866*
Base + f_{12}	0,0245*	0,0634*	0,0241	0,0537*
Base + f_{13}	0,2068*	0,1536*	0,1839	0,1030*
Base + f_{14}	0,2367*	0,1790*	0,1764	0,0945*

Tableau 5.4 – Apport des facteurs de pertinence pour le cas général.

Nous constatons une nouvelle fois que seuls les facteurs de pertinence f_4 et f_5 semblent avoir un intérêt car les autres dégradent les résultats.

Système	édition 2011		édition 2012	
	P@30	MAP	P@30	MAP
Base	0,2184	0,1785	0,1793	0,1001
Base + G1	0,2150*	0,1578*	0,1776*	0,0965*
Base + G2	**0,2646***	**0,2002***	**0,2213***	0,1209*
Base + G3	0,1370*	0,0886*	0,1386*	0,0679*
Base + G4	0,1544*	0,1081*	0,1322*	0,0787*
Base + G5	0,0558*	0,0736*	0,0667*	0,0609*
Base + G1 + G2	0,2558*	0,1930	0,2195*	**0,1286**
Base + G1 + G3	0,2007*	0,1229*	0,1643*	0,0870*
Base + G1 + G4	0,2170*	0,1549*	0,1684*	0,0952*
Base + G1 + G5	0,1646*	0,1245*	0,1402*	0,0822*
Base + G2 + G3	0,2413*	0,1635*	0,1936*	0,1099*
Base + G2 + G4	0,2395*	0,1791*	0,1966*	0,1162*
Base + G2 + G5	0,2027*	0,1537*	0,1741*	0,1076*
Base + G3 + G4	0,1471*	0,0912*	0,1421*	0,0693*
Base + G3 + G5	0,1210*	0,0925*	0,1088*	0,0665*
Base + G4 + G5	0,1136*	0,0978*	0,0908*	0,0639*
Base + G1 + G2 + G3	0,2565*	0,1746*	0,2012*	0,1187*
Base + G1 + G2 + G4	0,2544*	0,1888*	0,2063*	0,1249*
Base + G1 + G2 + G5	0,2306*	0,1741*	0,1885*	0,1157*
Base + G1 + G3 + G4	0,1971*	0,1270*	0,1643*	0,0872*
Base + G1 + G3 + G5	0,1732*	0,1194*	0,1427*	0,0823*
Base + G1 + G4 + G5	0,1782*	0,1294*	0,1414*	0,0829*
Base + G2 + G3 + G4	0,2283*	0,1599*	0,1871*	0,1081*
Base + G2 + G3 + G5	0,2355*	0,1487*	0,1649*	0,1032*
Base + G2 + G4 + G5	0,2061*	0,1603*	0,1816*	0,1076*
Base + G3 + G4 + G5	0,1355*	0,0959*	0,1140*	0,0684*
Base + G1 + G2 + G3 + G4	0,2486	0,1737	0,1994	0,1161
Base + G1 + G2 + G3 + G5	0,2449	0,1637	0,1842	0,1094
Base + G1 + G2 + G4 + G5	0,2374	0,1725	0,1845	0,1140
Base + G1 + G3 + G4 + G5	0,1717	0,1200	0,1392	0,0828
Base + G2 + G3 + G4 + G5	0,2167	0,1483	0,1690	0,1026
Base + G1 + G2 + G3 + G4 + G5	0,2391	0,1629	0,1842	0,1085

Tableau 5.5 – Apport des groupes de facteur de pertinence et de leurs combinaisons pour le cas général.

Si l'on prend maintenant en compte les différents groupes de facteurs de perti-

nence (tableau 5.5), le meilleur groupe est G2, contenant le facteur f_4. Ces résultats correspondent aux résultats obtenus dans le paragraphe précédent.

4.2.5 Discussion

La conclusion principale de ces expérimentations est que la présence de liens hypertextes dans les tweets semble être un indicateur de pertinence, en complément à la pertinence textuelle. Ceci est cohérent avec les résultats du chapitre 4, dans lequel nous avons montré l'apport important de la prise en compte des URLs dans la restitution.

De manière générale, les mêmes observations ont été remarquées avec les requêtes de 2011 et de 2012. En outre, ces observations persistent, que ce soit avec le modèle vectoriel ou avec Base, ce qui montre qu'elles ne dépendent pas forcément du modèle vectoriel et qu'elles sont généralisables.

Concernant maintenant le protocole expérimental utilisé, les résultats que nous avons présentés dans cet article sont basés sur un ensemble de $N = 5000$ tweets renvoyés par Lucene. Nous avons fait ce choix dans le but de maximiser le rappel des tweets pertinents (environ 80 %). Nous avons également mené d'autres expérimentations avec une valeur plus petite pour N (1500), sans que nos conclusions ne changent.

D'autre part, nous avons constaté qu'il n'y a pas au moins 30 tweets pertinents par topic. Par exemple, le système idéal pour les requêtes de 2011, atteindrait une P@30 de 0,7619. Par ailleurs, la P@30 étant une mesure ensembliste, elle ne tient pas compte du classement des résultats. Pour ces deux raisons, la MAP, qui est une mesure sensible au rang, nous semblerait plus appropriée afin de classer les participations officielles.

Dans la section suivante, nous présentons une étude plus approfondie sur l'apport des facteurs de pertinence. Nous allons en effet nous baser sur les techniques de sélection d'attributs afin de détecter les groupes de facteurs de pertinence qui reflètent la pertinence et qui sont susceptibles d'être utiles dans la recherche de microblogs avec les techniques d'apprentissage.

4.3 Étude avec les techniques de sélection d'attributs

4.3.1 Cadre expérimental

Nous avons utilisé Weka [2] pour ces expérimentations. Weka est un outil open-source écrit entièrement en Java et qui rassemble la plupart des techniques d'apprentissage et des techniques de sélection d'attributs.

2. `http://www.cs.waikato.ac.nz/ml`

Nous avons procédé ainsi : les premiers 1500 tweets pour chaque topic ont été restitués avec le modèle vectoriel. Ensuite, les scores de tous les facteurs de pertinence ont été calculés pour chaque tweet. Nous avons identifié les tweets pertinents et les tweets non pertinents. L'ensemble obtenu contient 72 614 tweets, répartis en 2 129 tweets pertinents et 70 485 tweets non pertinents. On peut remarquer que les classes de cet ensemble 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 (Yen et Lee, 2006). Pour cette raison, nous avons appliqué une approche de sous-échantillonnage pour générer une collection équilibrée composé de 2 129 tweets pertinents et 2,129 tweets non pertinents. Les tweets non pertinents pour cette étude ont été sélectionnés de manière aléatoire. Finalement, nous avons appliqué les techniques de sélection d'attributs sur l'ensemble obtenu.

Cette phase a fait ressortir plusieurs ensembles de critères. Ensuite, nous avons évalué l'efficacité de ces ensembles en les appliquant sur des techniques d'apprentissage dans un contexte de recherche de microblogs. Nous avons utilisé dans cette deuxième phase les requêtes de 2011 pour l'apprentissage et les requêtes de 2012 pour l'évaluation.

4.3.2 Résultats de l'étude

Le tableau 5.6 montre les facteurs de pertinence sélectionnés par les techniques de sélection d'attributs. Les facteurs de pertinence mis en avant par l'étude de la distribution des scores (section 3.3) correspondent à ceux ressortissant de cette étude (f_1, f_2, f_3, f_4, f_5, f_6, f_{13}). Ceci confirme l'importance de cet ensemble par rapport au reste des facteurs. Nous avons également effectué cette étude sans échantillonnage du corpus. Nous n'avons remarqué aucune différence sur les résultats de l'étude avec les techniques de sélection d'attributs.

Nous avons trouvé que les mêmes facteurs de pertinence sont mis en avant par l'étude de la distribution des scores et l'étude avec les techniques de sélection d'attributs. Ces facteurs de pertinence sont : facteurs de pertinence de contenu (*popularité du tweet, longueur du tweet, correspondance exacte des termes*), facteurs de pertinence d'hypertextualité (*présence d'URL, importance d'URLs, fréquence d'URLs*) et facteur de pertinence temporelle (*fraîcheur*).

D'autres facteurs de pertinence ont été sélectionnés pas les techniques de sélection d'attributs, quoique moins fréquemment : facteurs de pertinence de l'auteur (*nombre de tweets, nombre de citations*) et la qualité du langage. Finalement, les facteurs de pertinence des hashtags (*popularité du hashtag, présence de hashtags*) n'ont jamais été sélectionnés et semblent complètement non pertinents.

Algorithme	VSM	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14
Cfssubseteval	+	+	+	+	+	+	+						+	+	
ChisquaredAtt.Eval	+	+	+	+	+	+	+			+	+	+	+	+	+
FilteredAtt.Eval	+	+	+	+	+	+	+			+	+	+	+	+	+
FilteredSubsetEval	+	+	+	+	+	+	+							+	
Gain ration att eval	+	+	+	+	+	+	+			+	+	+	+	+	+
Info gain att eval	+	+	+	+	+	+	+			+	+	+	+	+	+
One att eval	+	+	+	+	+	+	+			+	+	+	+	+	+
ReliefFAttribute Eval	+	+	+	+	+	+	+			+	+	+	+	+	+
SVM Attribute Eval	+	+	+	+	+	+	+			+		+	+	+	+
SymetricalUncertEval	+	+	+	+	+	+	+			+	+	+	+	+	+
Consistency subset Eval	+	+	+	+	+	+	+			+	+	+	+	+	+
Wrapper subset Eval	+			+	+	+	+								
LatentSymanticAnalysis	+	+	+	+											
Total	13	12	12	13	12	12	12	0	0	9	8	9	10	11	9

Tableau 5.6 – Critères sélectionnés avec les techniques de sélection d'attributs

4.3.3 Application des résultats de l'étude dans la recherche

Dans cette section, nous évaluons certaines techniques d'apprentissage avec l'ensemble de facteurs identifiés lors des études précédentes. L'objectif est double : d'une part, nous cherchons à valider si la sélection d'attributs améliore effectivement les résultats d'une tâche de recherche de microblogs. D'autre part, nous cherchons à mesurer la performance de certaines techniques d'apprentissage dans ce type de classification.

Pour évaluer les techniques d'apprentissage, nous utilisons les résultats des requêtes de l'édition de 2011 comme collection d'apprentissage et les résultats de l'édition de 2012 pour les tests. Les modèles appris prédisent la classe de pertinence pour chaque tweet (pertinent ou non pertinent). Les tweets classés comme non pertinents sont ainsi supprimés. Les tweets classés comme pertinents sont triés selon les scores d'efficacité de la classification produit pas la technique d'apprentissage. Pour évaluer nos runs obtenus, nous nous basons sur la P@30 (la mesure officielle des tâches de 2011 et 2012).

Nous avons choisi de tester trois techniques d'apprentissage. Ce choix s'explique par le fait qu'elles sont les plus utilisées pour classer des documents de faible longueur. Par ailleurs, elles ont souvent montré leur efficacité dans la recherche de microblogs : SVM (Duan et al., 2010 ; Vosecky et al., 2012), J48 (Yuan et al., 2012) et Naive Bayes (Yuan et al., 2012).

Hall et Holmes (2003) 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 :

- Naive Bayes et Wrapper Subset Evaluation (WRP) qui utilise l'algorithme d'apprentissage ciblé afin d'estimer les meilleurs attributs. Ainsi, les facteurs sélectionnés dans ce cas sont le score de Lucene, f_3, f_4, f_5 et f_6.
- Naive Bayes et Correlation-based feature Selection (CFS) (*Lucene*, f_1, f_2, f_3, f_4, f_5, f_6, f_{12}, f_{13}).
- J48 et ReliefFAttribute Eval (RLF) (*Lucene*, f_1, f_2, f_3, f_5, f_6, f_9, f_{10}, f_{11}, f_{12}, f_{13}, f_{14}).
- SVM et SVM Attribute Eval qui évaluent les attributs en utilisant le classifieur SVM (*Lucene*, f_1, f_2, f_3, f_4, f_5, f_6, f_9, f_{11}, f_{12}, f_{13}, f_{14}).

Lucene	0,2842		
	Distribution de scores	Techniques de sélection	Tous les critères
J48	0,1627	0,0983 (RLF)	0,1000
Naive Bayes	**0,3305**	**0,3311** (WRP) **0,3356** (CFS)	0,2372
SVM	0,1689	0,1746 (SVM)	0,1729

Tableau 5.7 – Résultats (P@30), les scores en gras indiquent des améliorations significatives par rapport à la baseline

Le tableau 5.7 montre les résultats des trois techniques d'apprentissage appris avec les facteurs issus de l'étude de la distribution des scores, les facteurs ressortis de l'étude avec les techniques de sélection d'attributs et avec tous les facteurs. Les résultats ont été comparés avec le run nommé Lucene dans lequel seulement les scores de Lucene ont été utilisés pour trier les tweets et qui représente notre baseline.

4.3.4 Discussion et limites

L'objectif principal de cette étude était d'identifier la meilleure combinaison de facteurs de pertinence. Les facteurs de pertinence mis en évidence sont les mêmes que celles de l'étude par la distribution des scores.

Cette étude nous a permis également de vérifier si la sélection des attributs améliore l'efficacité des techniques d'apprentissage. Les critères identifiés par SVM attribute Eval, WRP, CFS, et par l'étude de la distribution des scores confirment l'hypothèse. À part J48 appris avec les critères sélectionnés avec RLF, tous les résultats ont été améliorés par rapport aux runs créés avec tous les critères. Nous notons

également que les techniques d'apprentissage, à part J48, ont été plus efficaces avec les techniques de sélection d'attributs qu'avec les critères ressortis de la distribution des scores.

Nous avons pu identifier également la meilleure technique d'apprentissage pour une tâche de recherche de microblogs. Nous pouvons remarquer que seul Naive Bayes dépasse Lucene (+18 % avec les critères sélectionnés en utilisant CFS et +16 % avec les critères ressortis de la distribution des scores). Les autres techniques d'apprentissage n'ont pas réussi à améliorer les résultats.

Nous avons comparé le run obtenu en utilisant Naive Bayes appris avec les critères obtenus de CFS avec les autres participants de la tâche Microblog de 2011. Nous avons fait apprendre Naive Bayes avec les critères de CFS et nous avons réalisé une validation croisée avec les requêtes de 2011. Nous avons obtenu une P@30 moyenne de 0,3707, ce qui nous aurait classé à la 5^{e} place parmi tous les participants qui n'ont pas utilisé des informations futures et qui ont soumis des runs automatiques. Cette précision est réduite de 10 % en utilisant le même modèle sur les requêtes de l'édition de 2012. En outre, les techniques d'apprentissage telles que J48 et SVM ont obtenu un gain de 80 % d'efficacité lorsqu'elles sont testées et croisées sur les requêtes de l'édition de 2011. Cependant, elles n'ont pas fonctionné comme prévu sur les requêtes de 2012. Toutes ces observations soulèvent la question suivante : les requêtes et les jugements de pertinence des tâches des deux années ont-ils été construits de la même manière ?

Dans le but de contrôler ce biais potentiel de la collection, nous avons fusionné les requêtes de 2011 et 2012 et nous avons répété les mêmes étapes. Nous avons obtenus une P@30 moyenne de 0,3435. Ce bon résultat confirme que Naive Bayes appris avec les critères obtenus avec CFS est le plus adapté à la recherche de microblogs.

5 Conclusion

Nous avons évalué dans ce chapitre les facteurs de pertinence souvent utilisés pour évaluer la pertinence des microblogs vis-à-vis d'un besoin en information. Nous avons montré expérimentalement ceux qui reflètent la pertinence. Nous avons calculé les scores des facteurs de pertinence. Ces scores ont été employés dans des combinaisons linéaires ou avec des techniques d'apprentissages, ou bien pour étudier leurs distributions dans les tweets pertinents et dans les tweets non pertinents. Les trois analyses ont montré l'importance des facteurs de pertinence liés aux URLs des tweets, ce qui complète encore une fois nos conclusions du chapitre précédent. Les facteurs liés aux hashtags ou à l'importance des auteurs n'ont cependant pas montré leur intérêt.

L'emploi de certains facteurs de pertinence permet d'améliorer les résultats d'une

tâche de recherche de microblogs lorsqu'ils sont utilisés afin de réordonner les résultats fournis par un modèle de RI classique. Cependant ces améliorations demeurent dépendantes du modèle de RI : les facteurs interviennent pour le classement des tweets candidats (identifiés au préalable).

Nous notons que le meilleur résultat de toutes les expérimentations de ce chapitre est obtenu par la combinaison linéaire du score du modèle vectoriel avec les scores des facteurs de pertinence des groupes G1 et G2, et non pas avec l'apprentissage.

Notre travail présente cependant quelques limites. D'abord, nous n'avons pas calculé des poids quantifiant les importances des critères de pertinence (même avec les techniques de sélection d'attributs). Intuitivement, il semblerait que certains soient plus pertinents que d'autre pour la restitution de microblogs. De même, nous n'avons aucune idée de la manière dont les critères sont combinés dans les techniques d'apprentissage (boite noire). Ensuite, nous n'avons pas pu évaluer d'autres facteurs utilisés dans certaines approches de recherche de microblogs, tels que la fréquence de retweet, le nombre d'abonnés d'un auteur. Ces facteurs nécessitent des informations supplémentaires que nous ne possédons pas dans le corpus utilisé pour nos expérimentations. Un accès ouvert à Twitter semble nécessaire pour obtenir ces informations et les évaluer, ce qui n'est pas possible.

La recherche d'information dans les microblogs implique la prise en compte automatique de la fraîcheur dans la pertinence. Ce facteur à été sélectionné 11 fois par les 13 techniques de sélection d'attributs que nous avons employé dans la section 4.3 (tableau 5.6). Nous traitons plus finement ce facteur dans le chapitre suivant.

Chapitre 6

Prise en compte du temps dans la recherche de microblogs

1 Introduction

Ounis et al. (2011) ont défini la recherche de microblogs de la façon suivante : *en cherchant dans les microblogs, l'utilisateur cherche à avoir l'information la plus récente, et pertinente, par rapport à un besoin d'information.* Teevan et al. (2011), quand à eux, ont également montré que l'une des principales motivations des utilisateurs qui utilisent un moteur de recherche de microblogs concerne l'information récente. Nous avons, de notre part, montré dans le chapitre précédent que le facteur temps est souvent sélectionné comme facteur pertinent pour la recherche de microblogs.

Ces trois constats suggèrent que la fraîcheur est un facteur de pertinence crucial pour la restitution de microblogs. Dans un premier temps, nous avons intégré la fraîcheur de deux manières différentes dans le calcul de la pertinence des tweets. Nous avons (i) renforcé les scores de pertinence des tweets récents par rapport à la date de soumission de la requête. Ensuite, nous avons (ii) favorisé les termes qui apparaissent fréquemment au moment de soumission de la requête. Dans un deuxième temps (iii) nous avons exploité les distributions temporelles des termes des tweets potentiellement pertinents dans le calcul de la pertinence. L'idée ainsi est de promouvoir un tweet restitué contenant des termes fréquemment utilisés le jour de sa publication.

2 Emploi de la fraîcheur dans la restitution des microblogs

Nous prenons en compte à ce niveau le facteur fraîcheur par rapport à la date de soumission de la requête dans la mesure de la pertinence. Ce facteur peut être pris en compte de différentes manières. Dans un premier temps, nous proposons d'amplifier les scores de pertinence du contenu d'un tweet en fonction de sa proximité temporelle avec la date de la requête. Dans un deuxième temps, nous proposons de favoriser les termes fréquemment utilisés au moment de la soumission de la requête. Nous avons choisi d'utiliser la méthode de *Kernel Laplace* utilisée dans (Lv et Zhai, 2009) pour amplifier les scores du modèle de restitution [1] en fonction de la fraîcheur du tweet. La formule de Kernel est :

$$k(i,j) = \frac{1}{2b}\ exp\left(\frac{-|i-j|}{b}\right) avec\ \sigma^2 = 2b^2 \tag{6.1}$$

Dans notre cas, i et j représentent respectivement les dates en jour de la soumission de la requête et la date de publication du tweet. le facteur σ est le facteur qui permet de modifier le degré d'amplification des scores.

2.1 Favoriser des tweets récents

Une façon simple de prendre en compte la fraîcheur d'un tweet est d'amplifier son score de pertinence de contenu en fonction de sa date de proximité temporelle avec la requête. L'intuition ici est que certains tweets, même ayant un score de pertinence de contenu faible, sont pertinents du fait de leur fraîcheur par rapport à la date de soumission de la requête. En contrepartie, d'autres tweets, même ayant des scores de pertinence de contenu élevés, ne sont pas pertinents du fait de leur distance temporelle importante par rapport à la date de la soumission de la requête.

Le score de chaque tweet devient ainsi :

$$RSVT_1(q,d,\sigma) = RSV(q,d) * k_\sigma(t_q, t_d) \tag{6.2}$$

avec $k_\sigma(t_q, t_d)$ est le score du facteur de Kernel. Nous avons fait varier la valeur de σ pour observer l'impact de l'amplification sur les résultats. Le tableau 6.1 montre l'ensemble des résultats.

1. Les méthodes que nous proposons sont basées sur le modèle BM25 avec $K_1 = 0,1$ et $b = 0$. Nous avons choisi cette configuration car elle a obtenu de meilleurs résultats que le modèle vectoriel et que le modèle BM25 de base (chapitre 4). Nous n'avons pas employé d'autres facteurs comme l'expansion de requêtes ou de documents (l'objectif ici est d'étudier uniquement le facteur temporel). Nous nous sommes basés sur les 60 requêtes de l'édition 2012 de la tâche Microblog de TREC.

σ	Rappel	P@30	MAP
$RSV(q,d)$	**0,6643**	**0,3186**	**0,2170**
$RSVT_1(q,d,2)$	0,2388	0,0432	0,1175
$RSVT_1(q,d,14)$	0,4849	0,2305	0,1178
$RSVT_1(q,d,18)$	0,5427	0,2379	0,1362
$RSVT_1(q,d,28)$	0,5950	0,2729	0,1695
$RSVT_1(q,d,32)$	0,6082	0,2797	0,1782
$RSVT_1(q,d,50)$	0,6295	0,2910	0,1938
$RSVT_1(q,d,90)$	0,6520	0,2960	0,2024
$RSVT_1(q,d,230)$	0,6597	0,3119	0,2111
$RSVT_1(q,d,350)$	0,6633	0,3153	0,2155

Tableau 6.1 – Amplification des scores de pertinence de contenu en fonction de leur fraîcheur

Comme les résultats le montrent, l'amplification des scores du modèle de restitution n'a pas amélioré les résultats. En faisant augmenter σ, l'effet de l'amplification diminue, et les résultats se rapprochent des résultats du modèle de recherche de base.

2.2 Favoriser les termes récents

L'intuition ici est de considérer que les termes les plus représentatifs pour exprimer un besoin en information dans les microblogs sont des termes fréquemment utilisés au moment de la soumission de la requête : un document, même ancien par rapport à la date de soumission de la requête, contenant des termes fréquemment utilisés au moment de la requête est plus pertinent qu'un document récent, contenant des termes fréquemment utilisés dans des périodes lointaines par rapport à la requête. Pour prendre en compte cette intuition, nous avons modifié le facteur IDF du modèle de restitution ($RSVT_2(q,d,\sigma)$) :

$$IDF = log\left(\frac{N-(R_i)_{temps}}{(R_i)_{temps}}\right) \tag{6.3}$$

$$(R_i)_{temps} = \sum_t (|R_i|_t * k_\sigma(t_q,t)) \tag{6.4}$$

avec t correspond à une fenêtre temporelle exprimée en jours et $|R_i|_t$ correspond au nombre de documents dans cette fenêtre temporelle. Le tableau 6.2 présente les résultats. Nous avons fait varier σ :

σ	Rappel	P@30	MAP
$RSV(q,d)$	0,6643	**0,3186**	**0,2170**
$RSVT_2(q,d,2)$	0,6640	0,3130	0,2156
$RSVT_2(q,d,10)$	0,6647	0,3130	0,2159
$RSVT_2(q,d,20)$	0,6657	0,3136	0,2160
$RSVT_2(q,d,30)$	0,6657	0,3136	0,2160
$RSVT_2(q,d,40)$	**0,6659**	0,3119	0,2157
$RSVT_2(q,d,50)$	**0,6659**	0,3085	0,2128

Tableau 6.2 – Amplification des scores des termes en fonction de leur fréquence d'apparition dans le temps

Encore une fois, la prise en compte de la fraîcheur n'a pas montré une amélioration significative (à part une légère amélioration de 0,28 % au niveau du rappel).

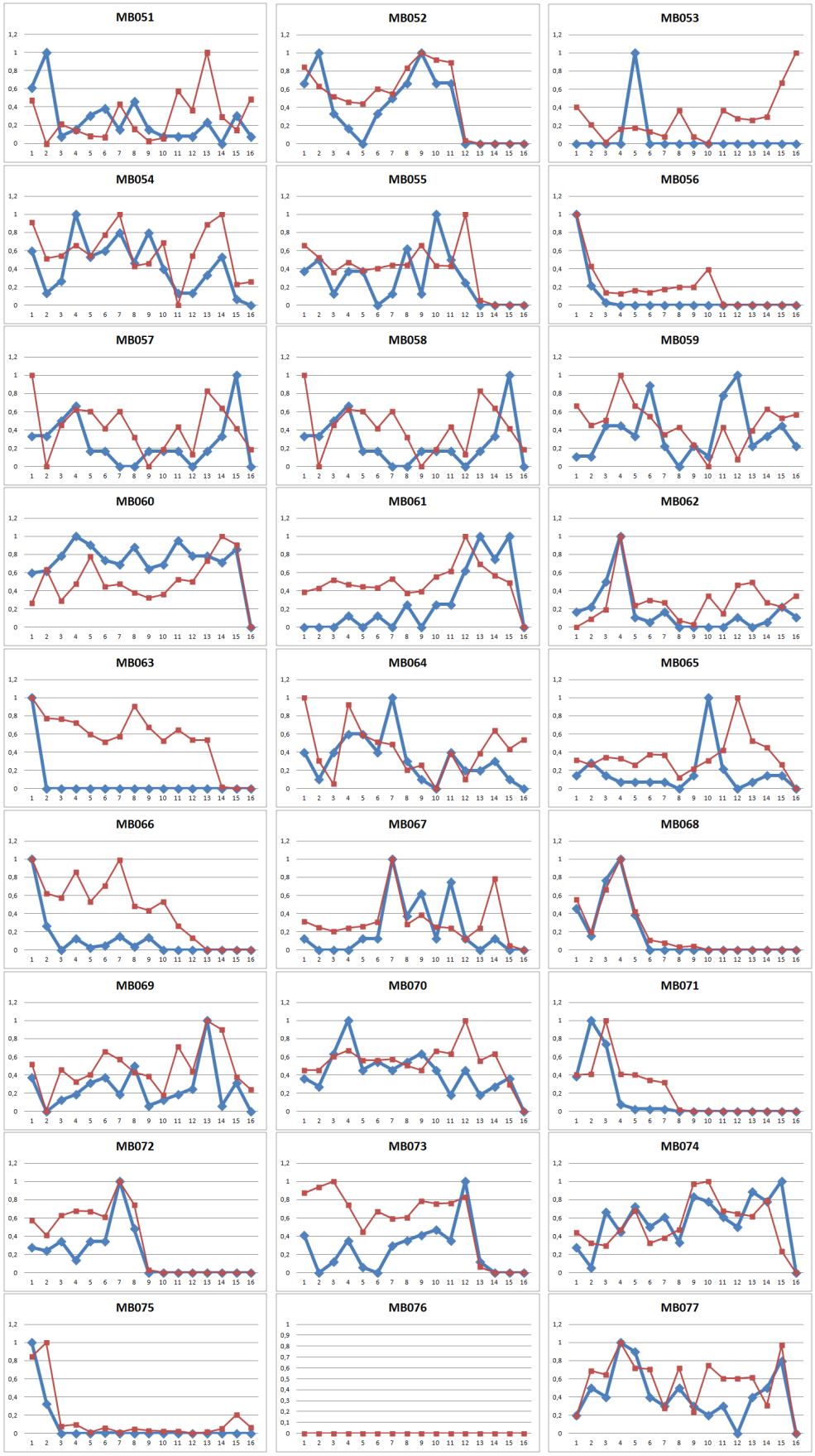
2.3 Observations

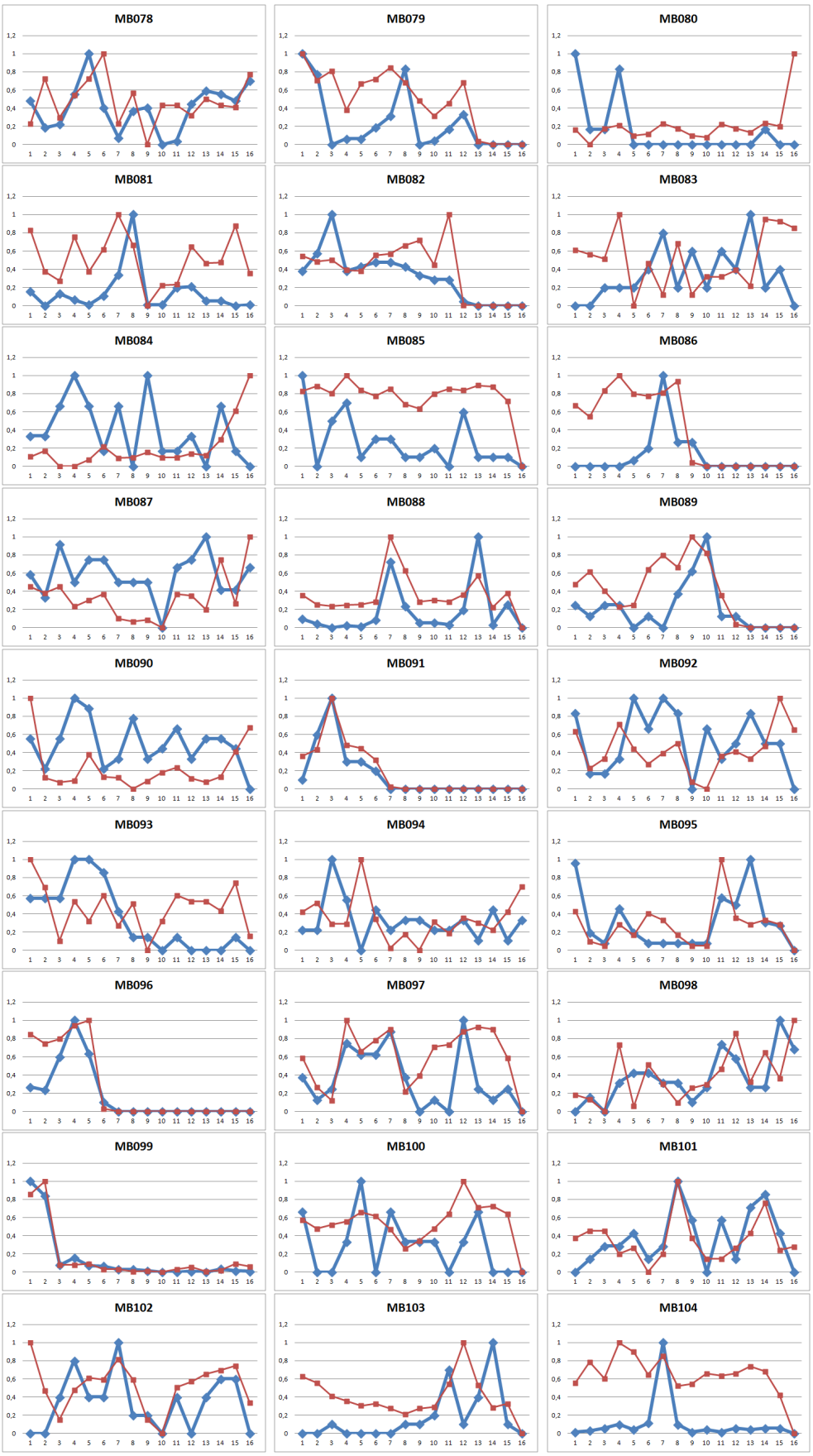
De manière générale, l'emploi de la fraîcheur dans les deux méthodes proposées n'apporte pas d'amélioration. Afin de vérifier si la fraîcheur a un impact sur les résultats, nous avons regardé la distribution temporelle des tweets pertinents et des tweets non pertinents pour l'ensemble des requêtes.

La figure 6.1 montre les ratios des distributions temporelles des tweets pertinents et non pertinents de chaque requête, ainsi que la distribution générale des tweets sur l'ensemble des requêtes (nommée somme). Nous nous sommes basés sur les jugements de pertinence (qrels) pour sélectionner ces tweets. Les courbes présentent le ratio des quantités de tweets pertinents (bleu/carrés inclinés) et des non pertinents (rouge) par jour. Nous pouvons remarquer que les distributions diffèrent d'une requête à une autre. Les tweets pertinents ne sont pas toujours récents par rapport à la date de la soumission des requêtes. En analysant chaque requête séparément, nous pouvons affirmer que la prise en compte de la fraîcheur pénalise les résultats de plusieurs requêtes dont les dates de la plupart des tweets pertinents sont relativement éloignés de sept jours de la date de soumission des requêtes (ex. MB088, MB089, MB095. . .).

Par ailleurs, nous remarquons que les tweets pertinents arrivent par rafales. Les positions de ces rafales sont différentes d'une requête à une autre. Nous avons ainsi testé une troisième méthode qui, pour le calcul du score d'un tweet, tient compte de la fréquence des tweets publiés le jour de sa publication. L'objectif est de promouvoir un tweet s'il est publié dans une période qui correspond à une rafale de tweets. Par exemple, pour la requête MB065, la plupart des tweets pertinents sont apparus dix jours avant la date de soumission de la requête. Ainsi, l'idée est de favoriser les

tweets publiés dans cette fenêtre temporelle.




MB078
MB079
MB080
MB081
MB082
MB083
MB084
MB085
MB086
MB087
MB088
MB089
MB090
MB091
MB092
MB093
MB094
MB095
MB096
MB097
MB098
MB099
MB100
MB101
MB102
MB103
MB104

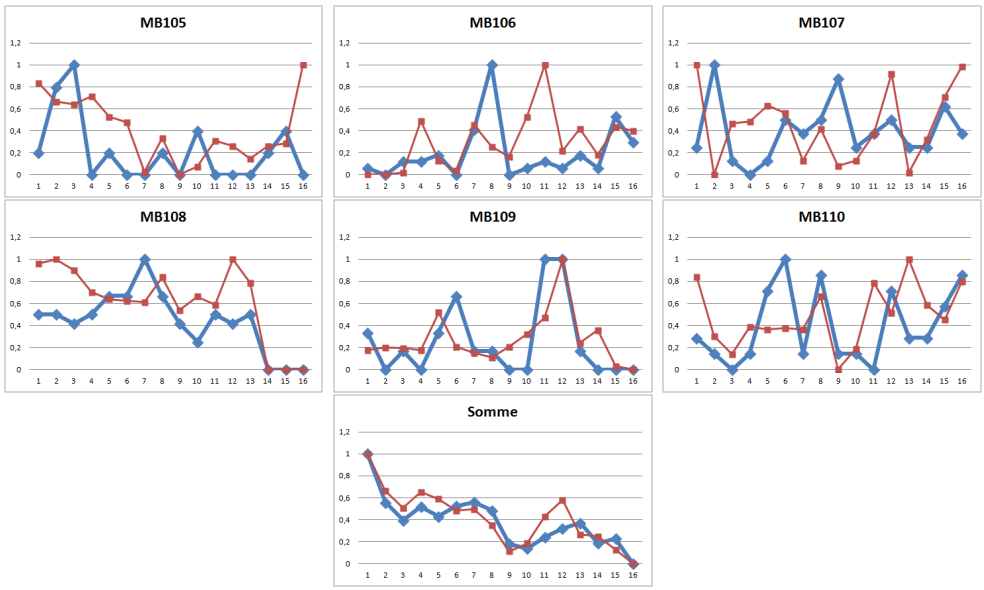


FIGURE 6.1 – Distribution temporelle des tweets pertinents et non pertinents pour les requêtes de TREC Microblog 2012. Les rectangles représentent les tweets pertinents tandis que les losanges représentent les tweets non pertinents.

3 Prise en compte de la fréquence temporelle

Nous prenons en compte à ce niveau les distributions temporelles des résultats. Nous essayons ainsi de favoriser les résultats qui apparaissent dans les périodes de rafales. Cette troisième méthode amplifie le score d'un terme dans un tweet publié à un instant t en fonction de la fréquence d'emploi de ce terme dans cette période t. Un même terme aura des scores différents en fonction de la date de soumission du document auquel il appartient. Ce score sera plus important si le terme appartient à un document publié dans une période de rafale de ce terme, que dans le cas où il appartient à un document publié dans une période où le terme n'est pas fréquemment utilisé. De cette manière, nous favorisons les résultats publiés dans des périodes de rafales.

Pour mettre en application cette intuition, nous avons employé un nouveau facteur : IDF_{new}.

$$IDF_{new} = IDF * 1/IDF_{local} \tag{6.5}$$

et

$$IDF_{local} = log\left(\frac{N - (R_i)_t}{(R_i)_t}\right) \tag{6.6}$$

avec $(R_i)_t$ est le nombre de tweets contenant le terme i le jour de la publication du tweet. IDF_{local} est le IDF d'un terme, mais sur une fenêtre temporelle d'un jour (est non pas sur toute la collection). Ainsi, un terme va avoir un IDF_{local} différent

pour chaque jour. Ce facteur est plus important dans un jour où le terme n'est pas fréquemment utilisé, que dans un jour où il est fréquemment utilisé (supposé correspondant à un jour de rafale). Pour cette raison, nous employons l'inverse de ce facteur : $1/IDF_{local}$. Le modèle qui prend en compte IDF_{new} dans le modèle de restitution est nommé $RSVT_3(q, d)$.

	Rappel	P@30	MAP
$RSV(q, d)$	**0,6643**	0,3186	**0,2170**
$RSVT_3(q, d)$	0.6469	**0.3198**	0.2087

Tableau 6.3 – Prise en compte de la fréquence temporelle.

Le tableau 6.3 montre que cette méthode n'a pas donné d'amélioration significative sur l'ensemble des requêtes.

Afin de mieux comprendre l'impact de nos méthodes, nous avons analysé les résultats requête par requête afin de voir si l'effet négatif de nos méthodes apparaît sur toutes les requêtes ou uniquement sur certaines. L'objectif est de voir si la prise en compte de la fraîcheur ou du temps dans la restitution permet d'améliorer certaines requêtes en particulier, et d'identifier des spécificités de ces requêtes.

4 Analyse requête par requête

Dans cette section, nous analysons l'impact de nos trois méthodes, requête par requête. Nous considérons la mesure MAP, car elle tient compte à la fois du rappel et de la précision.

Le tableau 6.4 montre les requêtes pour lesquelles nous avons eu des améliorations avec chacune des trois méthodes. De manière générale, la modification de σ dans Kernel ne change pas les différentes observations.

La première remarque est que la deuxième approche améliore 51 requêtes parmi les 60. Cependant, nous n'avons identifié aucune particularité commune, que ce soit au niveau des distributions temporelles des résultats ou bien de la sémantique des requêtes, pour les requêtes n'ayant pas obtenu d'amélioration. Nous n'avons également pas trouvé de spécificités communes pour les requêtes améliorées avec la première et la dernière méthode. Par exemple, la première méthode a amélioré les requêtes MB059 « Glen Beck » et MB085 « Best Buy improve sales ». Ces deux requêtes n'ont aucune sensibilité temporelle. De même, pour la troisième approche, nous notons des améliorations pour les requêtes MB060 « fishing guidebook », MB064 « red light cameras », MB102 « school lunches »... Ces requêtes n'ont également aucune sensibilité temporelle.

$RSVT_1(q, d, \sigma)$	MB058, MB059, MB063, MB066, MB067, MB071, MB075, MB079, MB080, MB085, MB091, MB093, MB107
$RSVT_2(q, d, \sigma)$	MB051, MB052, MB053, MB054, MB056, MB057, MB059, MB060, MB062, MB063, MB064, MB065, MB067, MB069, MB070, MB071, MB072, MB073, MB074, MB075, MB076, MB077, MB078, MB079, MB080, MB081, MB083, MB084, MB085, MB086, MB087, MB088, MB089, MB090, MB091, MB092, MB093, MB094, MB095, MB098, MB099, MB100, MB101, MB102, MB103, MB105, MB106, MB107, MB108, MB109, MB110
$RSVT_3(q, d)$	MB051, MB054, MB057, MB059, MB066, MB069, MB070, MB075, MB077, MB079, MB080, MB081, MB085, MB086, MB088, MB089, MB092, MB093, MB094, MB095, MB096, MB098, MB100, MB101, MB102, MB107, MB108, MB109

Tableau 6.4 – Requêtes améliorées sur la mesure MAP pour les 3 méthodes

Ensuite, nous avons identifié manuellement les requêtes sensibles au temps. L'objectif est de voir si, pour ces requêtes et avec la prise en compte du temps, les résultats ont été améliorés. Dans le cas contraire, nous essayons de comprendre les raisons. Ces requêtes correspondent principalement à des événements (par exemple : « Hu Jintao visit to the United States », « Autralian Open Djokovic vs. Murray », « fashion week in NYC »...). Nous avons sélectionné 13 requêtes qui parlent explicitement d'événements et qui sont clairement sensibles au temps (MB051, MB057, MB061, MB065, MB067, MB071, MB075, MB079, MB086, MB093, MB096, MB098, MB106).

L'impact de la prise en compte du temps pour ces requêtes diffère d'une approche à une autre :

- 8/13 (MB051, MB057, MB061, MB065, MB086, MB096, MB098, MB106) n'ont pas été améliorées avec la première approche.
- 11/13 (MB051, MB057, MB065, MB067, MB071, MB075, MB079, MB086, MB093, MB098, MB106) n'ont pas été améliorées avec la deuxième approche.
- 5/13(MB061, MB065, MB067, MB071, MB106) n'ont pas été améliorées avec la troisième approche.

La première remarque est que la troisième approche est celle qui arrive à amélio-

rer le nombre le plus important de requêtes sensibles au temps (8/13). Concernant les deux premières approches (se basant sur la fraîcheur par rapport à la date de la soumission de la requête), la cause principale pour laquelle il n'y avait pas d'amélioration consiste en la concentration des tweets pertinents dans des dates lointaines par rapport à la date de la requête. C'est la cas des requêtes MB057, MB061, MB065, MB067, MB079, MB086, MB093, MB098 et MB106. Ainsi, la prise en compte de la fraîcheur n'a pas montré d'intérêt. Concernant les requêtes MB071, MB075 et MB096, nous pouvons remarquer à partir des courbes de distributions temporelles que les tweets pertinents sont proches de la date de la soumission de la requête. Cependant, nous notons que, pour ces requêtes, les courbes des tweets pertinents sont très similaires aux courbes des tweets non pertinents. Ainsi, le fait de favoriser les tweets récents va impliquer les tweets pertinents et les tweets non pertinents, ce qui explique la dégradation des résultats pour elles. Finalement, la requête MB051 se caractérise par l'apparition de la grande partie des tweets pertinents à une date récente par rapport à la date de la requête, et les distributions des tweets pertinents et des tweets non pertinents ne sont pas similaires. Nous avons ainsi regardé les résultats restitués pour cette requête et nous avons remarqué que le modèle de RI employé n'a pas restitué une grande partie des tweets pertinents apparus récemment par rapport à la date de la requête. Ceci est du aux problèmes de vocabulaires étudiés dans le chapitre 3. Par conséquent, la prise en compte de la fraîcheur n'a pas montré son effet.

Concernant la troisième approche (qui prend en compte des distributions temporelles des résultats), nous avons étudié les résultats des requêtes pour lesquelles il n'y avait pas eu d'amélioration. Nous avons trouvé que le modèle de restitution de base (sans l'intégration de la fraîcheur) a restitué tous les tweets pertinents apparus dans les périodes de rafales. Par conséquent, la prise en compte du temps a favorisé uniquement la restitution des tweets non pertinents, pour ces périodes, ce qui a engendré une dégradation des résultats.

5 Conclusion

Nous avons étudié l'impact de la prise en compte du temps dans la recherche de microblogs. Nous avons proposé trois méthodes qui prennent en compte le temps de façons différentes. De manière générale, nous avons trouvé que la fraîcheur n'est pas un facteur de pertinence. Ce constat vient à l'encontre la définition de la tâche de recherche de microblogs dans TREC et aussi de l'état de l'art. Dans la collection utilisée pour nos expérimentations, la date de la soumission des requêtes correspond à la date de publication du tweet pertinent le plus récent. Cependant, nous avons trouvé, que pour plusieurs requêtes, la majorité des tweets pertinents sont publiés sept jours

avant la date de soumission de la requête. Nous avons également proposé une méthode qui se focalise sur les fenêtres de concentration temporelle des termes des requêtes dans la restitution. Cette approche n'a également pas montré d'amélioration significative. Toutefois, c'est la seule approche qui a obtenu une P@30 meilleure que celle du modèle BM25. Des études plus approfondies sur ce point doivent être réalisées. De plus, nous avons regardé les résultats de chaque requête avec chacune des trois approches. Nous avons trouvé que chaque approche améliore les résultats de certaines requêtes et dégrade les résultats d'autres. Cependant, nous n'avons pas trouvé de spécificités communes pour les requêtes ayant obtenu des améliorations, ni pour celles qui ont subi des dégradations. Finalement, nous avons identifié manuellement les requêtes sensibles au temps. Nous avons trouvé que c'est la troisième approche qui a amélioré la plus grande partie de ces requêtes. Ces résultats nous encouragent à prendre en compte le temps dans la restitution, en particulier avec les requêtes sensibles au temps. Il reste maintenant à savoir comment les identifier.

Chapitre 7

Conclusion générale

Synthèse

Nous nous sommes intéressés dans ces travaux à la RI adhoc dans les microblogs. L'objectif est de retrouver les microblogs répondant à un besoin d'information spécifié par un utilisateur. Pour réaliser nos expérimentations, nous nous sommes basés sur le corpus fourni par la campagne d'évaluation internationale TREC (Text Retrieval Conference) dans la tâche Microblog des éditions de 2011 et 2012. Nos différentes contributions ont également fait l'objet de participations aux trois tâches de Microblogs de TREC (2011, 2012 et 2013). Nos contributions se situent à plusieurs niveaux :

- Afin de déterminer exactement les facteurs limitant les performances des modèles classiques de RI dans un corpus de microblogs, nous avons mené une analyse de défaillance d'un modèle de recherche usuel. Nous avons sélectionné les microblogs pertinents mais non retrouvés par le modèle de recherche. Ensuite, nous avons identifié les facteurs empêchant leur restitution. Nous avons trouvé que **le problème principal vient de la concision des microblogs**. Cette concision engendre une correspondance limitée entre les termes des microblogs et les termes des requêtes, même s'ils sont sémantiquement semblables. Toutefois, ce facteur est apparu sous différentes formes : **absence totale des termes de certaines requêtes dans les documents pertinents, caractère non discriminant des termes de requêtes...** Nous avons également identifié des problèmes de lemmatisation : **termes non appariés quoique dérivant d'une même racine, ou des termes concaténés sous formes de hashtags ou de citations**. Outre le problème de vocabulaire, nous avons remarqué que, pour plusieurs requêtes, **certains termes n'ont pas un caractère discriminant**. Par conséquent, ces termes n'aident pas à sélectionner les résultats pertinents.
- Afin de compenser l'impact de la concision des microblogs, nous avons pro-

posé et testé plusieurs solutions. Nous avons proposé d'étendre les requêtes (i) en exploitant des ressources de type actualités, (ii) en utilisant la base lexicale WordNet, (iii) en appliquant des techniques de réinjection de pertinence de l'état de l'art. Ces techniques ont souvent prouvé leur efficacité : Rocchio pour identifier les termes susceptibles de ramener la pertinence ainsi que pour la pondération des termes de la nouvelle requête, et le mécanisme naturel d'extension de requêtes du modèle BM25. Dans Rocchio, nous avons testé différentes méthodes de calcul de poids de termes d'expansion. Nous avons enfin étendu les microblogs grâce aux liens (URLs) qu'ils contiennent. Nos expérimentations ont montré que **l'emploi des URLs et l'expansion de requêtes a partir du *feedback* sont primordiales pour la RI dans les microblogs**. L'expansion de requêtes avec les articles d'actualité améliore uniquement la précision. La plupart de ces expérimentations (expansion de requêtes et de microblogs) ont été réalisées en se basant sur le modèle vectoriel et sur le modèle probabiliste comme modèle de restitution. Ceci nous a permis de comparer les comportements des deux modèles sur les microblogs et avec les deux types d'expansion. De manière générale, nous avons trouvé que **le modèle vectoriel est plus performant que modèle probabiliste au niveau de la sélection des microblogs pertinents** (meilleur rappel). Cependant, **le modèle probabiliste met davantage en valeur les microblogs pertinents restitués par rapport à tous les microblogs restitués** (meilleure précision).

– Un deuxième volet de notre travail concerne l'étude des facteurs de pertinence utilisés pour identifier les microblogs pertinents. Nous avons repris les facteurs souvent utilisés dans l'état de l'art (facteurs liés au contenu, facteurs liés aux auteurs, facteurs liés aux URLs, facteurs liés aux hashtags et facteurs liés à la qualité des tweets) et nous les avons évalués. Nous avons réalisé cette analyse selon trois axes. Dans le premier axe, nous avons étudié le comportement des facteurs de pertinence dans les documents pertinents et les avons comparés à leur comportement dans les documents non pertinents. Dans le deuxième axe, nous avons analysé l'impact de la combinaison des scores des facteurs avec le score de pertinence du contenu, calculé avec un modèle de RI usuel. Dans le troisième axe, nous avons utilisé des techniques d'apprentissage ainsi que des algorithmes de sélection d'attributs qui peuvent être utiles en entrée de ces techniques d'apprentissages. De manière générale, nous avons montré que **les facteurs liés aux URLs publiées dans les tweets sont les plus discriminants**. **Les facteurs liés aux auteurs ou aux hashtags ne reflètent pas la pertinence**. Nous avons également comparé différentes techniques d'apprentissage souvent utilisées dans l'état de l'art pour la recherche de microblogs. Nous avons trouvé que **Naive Bayes est le plus adapté** pour ce

type de recherche et ceci en considérant les meilleurs critères de pertinence identifiés.

– Afin de prendre en compte l'aspect temporel dans la restitution des microblogs pertinents vis-à-vis d'un besoin en information, nous avons proposé trois méthodes qui intègrent le temps dans le calcul de la pertinence. Cette intégration du temps n'a cependant pas montré son intérêt dans nos méthodes. Une analyse plus poussée, requête par requête, nous a permis de voir que la fraîcheur ne représente en effet pas un facteur de pertinence pour la restitution de microblogs.

Limites et perspectives

Nous commençons par présenter nos perspectives à court terme pour arriver à celles à long terme :

– Dans un premier temps, nous aimerions compléter le traitement des différentes formes du problème de vocabulaire soulignées dans le chapitre 3. Nous avons trouvé que, dans plusieurs cas, les tweets pertinents contiennent les termes des requêtes concaténés sous forme de hashtags. Nous avons testé une méthode pour décomposer ces hashtags. Cette méthode se basait sur les lettres majuscules pour identifier le début de chaque terme composant. Cependant, elle ne nous a pas permis d'améliorer les résultats. Une solution à ce problème consiste à employer l'algorithme de segmentation proposé dans le livre « Beautiful Data » (Segaran et Hammerbacher, 2009), permettant de décomposer les termes concaténés. La même approche peut être employée également pour résoudre les problèmes reliés aux lemmatiseurs : termes non appariés dérivant d'une même racine.

– Dans le chapitre 5, nous n'avons pas pu évaluer certains facteurs de pertinence tels que le nombre de fois un tweet a été retweeté ou le nombre de fois il a été favori. Nous n'avons pas ces informations dans la collection d'évaluation utilisée. La solution ainsi consiste à créer une nouvelle collection contenant toutes les informations requises.

– Considérer la fraîcheur dans la restitution des microblogs n'a pas montré un intérêt. Toutefois, nous avons trouvé que, dans la plupart des cas, les tweets arrivent par rafales. L'idée ainsi est de trouver un moyen pour identifier les fenêtres temporelles correspondant aux rafales de tweets au préalable et les utiliser comme *feedback* ou comme source d'expansion de requêtes.

– La grande majorité des travaux réalisés sur les microblogs, et en particulier nos travaux, emploient Twitter comme cadre applicatif. Notre objectif est d'étudier ainsi si nos résultats et nos observations sont valables également sur les autres

plate-formes de microblogging telles que Blipper et Tumblr.

- La tâche de recherche de microblogs consiste à restituer des microblogs pertinents vis-à-vis d'un besoin en information. Nous avons trouvé, regardant les résultats des qrels de la tâche Microblog de TREC, que plusieurs tweets pertinents ont exactement le même contenu et ramènent les mêmes informations. Dans le cas idéal, un utilisateur devra ainsi consulter tous les tweets pertinents (parfois des centaines) pour s'assurer d'avoir vu tous les aspects d'une requêtes. Pour simplifier la tâche, créer un synthétiseur de résultats permettant d'une part d'éliminer les informations qui se répètent, et d'autre part de représenter les résultats d'une manière plus lisible.
- Une des principales caractéristiques des plate-formes de microblogging est leur aspect social. Les utilisateurs ne produisent pas uniquement du contenu informatif, mais ils peuvent s'impliquer dans des conversations avec d'autres utilisateurs, en commentant, aimant et partageant leurs publications. Ainsi, il est important dans ce cas de pouvoir restituer tout le contexte d'un tweet. Une méthode de présenter le contexte est d'extraire la conversation à laquelle un tweet appartient. L'identification des critères permettant d'extraire des conversations à partir des microblogs représente un vrai défi. Les microbloggeurs discutent entre eux sans utiliser forcément les moyens explicites de conversations donnés par les plate-formes (retweet, hashtag, citation, réponse...).
- Finalement, agréger des informations de différentes sources (Web, images, wiki, actualités...) pour répondre aux besoins en information, a montré son intérêt (Kopliku et al., 2011). Cette technique permet de présenter à l'utilisateur des résultats variés et complémentaires. Considérer les microblogs (information fraîche) en plus des sources employées dans (Kopliku et al., 2011) semble très utiles, étant donné l'importance des microblogs aujourd'hui, en particulier, en tant que source d'information. L'objectif ainsi est d'étudier l'apport de la prise en compte des microblogs en complément des autres sources d'information du Web, pour répondre aux besoins en informations.

Références

Aboulnaga, Y., et Clarke, C. L. (2012). Frequent Itemset Mining for Query Expansion in Microblog Ad-hoc Search. In *TREC'12 : 21th Text Retrieval Conference.* National Institute of Standards and Technology (NIST).

Attardi, G., et Simi, M. (2006). Blog mining through opinionated words. In E. M. Voorhees et L. P. Buckland (Eds.), *Trec* (Vol. Special Publication 500-272). National Institute of Standards and Technology (NIST).

Baccianella, A. E. S., et Sebastiani, F. (2010). Sentiwordnet 3.0 : An enhanced lexical

resource for sentiment analysis and opinion mining. In *Proceedings of the seventh conference on international language resources and evaluation (lrec'10).* Valletta, Malta : European Language Resources Association (ELRA).

Baeza-Yates, R. A., et Ribeiro-Neto, B. (1999). *Modern information retrieval.* Boston, MA, USA : Addison-Wesley Longman Publishing Co., Inc.

Bai, J., Nie, J.-Y., Cao, G., et Bouchard, H. (2007). Using query contexts in information retrieval. In *Proceedings of the 30th annual international acm sigir conference on research and development in information retrieval* (pp. 15–22). New York, NY, USA : ACM.

Bamman, D., Eisenstein, J., et Schnoebelen, T. (2012). Gender in twitter : Styles, stances, and social networks. *CoRR*, *abs/1210.4567*.

Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., et Su, Z. (2007). Optimizing web search using social annotations. In *Proceedings of the 16th international conference on world wide web* (pp. 501–510). New York, NY, USA : ACM.

Belkin, N. J., et Croft, W. B. (1992). Information filtering and information retrieval : Two sides of the same coin ? *Commun. ACM*, *35*(12), 29–38.

Ben Jabeur, L., Damak, F., Tamine, L., Cabanac, G., Pinel-Sauvagnat, K., et Boughanem, M. (2013). IRIT at TREC Microblog Track 2013. In E. M. Voorhees et (Eds.), *Text REtrieval Conference (TREC), Gaithersburg, USA,*. National Institute of Standards and Technology (NIST).

Ben Jabeur, L., Damak, F., Tamine, L., Pinel-Sauvagnat, K., Cabanac, G., et Boughanem, M. (2012). IRIT at TREC Microblog 2012 : Adhoc Task. In E. M. Voorhees et L. P. Buckland (Eds.), *Text REtrieval Conference (TREC), Gaithersburg, USA,*. National Institute of Standards and Technology (NIST).

Ben Jabeur, L., Tamine, L., et Boughanem, M. (2011). Un modèle de recherche d'information sociale dans les microblogs : cas de twitter. In *Conférence sur les modèles et l'analyse des réseaux : Approches mathématiques et informatique.*

Ben Jabeur, L., Tamine, L., et Boughanem, M. (2012). Active microbloggers : Identifying influencers, leaders and discussers in microblogging networks. In L. Calderón-Benavides, C. González-Caro, E. Chávez, et N. Ziviani (Eds.), *String processing and information retrieval* (Vol. 7608, p. 111-117). Springer Berlin Heidelberg.

Bernstein, M., Suh, B., Hong, L., Chen, J., Kairam, S., et Chi, E. (2010). Eddi : interactive topic-based browsing of social status streams. In *Acm symposium on user interface software and technology* (p. 303-312). New York, NY : ACM.

Blei, D. M., Ng, A. Y., et Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, *3*, 993–1022.

Bollen, J., Pepe, A., et Mao, H. (2009). Modeling public mood and emotion : Twitter sentiment and socio-economic phenomena. *CoRR*, *abs/0911.1583*.

Brin, S., et Page, L. (1998). The anatomy of a large-scale hypertextual web search

engine. *Comput. Netw. ISDN Syst.*, *30*, 107–117.

Buckley, C., et Voorhees, E. M. (2000). Evaluating evaluation measure stability. In *Proceedings of the 23rd annual international conference on research and development in information retrieval* (pp. 33–40). New York, NY, USA : ACM SIGIR.

Cai, Y., et Li, Q. (2010). Personalized search by tag-based user profile and resource profile in collaborative tagging systems. In *Proceedings of the 19th acm international conference on information and knowledge management* (pp. 969–978). New York, NY, USA : ACM.

Cappelletti, R., et Sastry, N. (2012). Iarank : Ranking users on twitter in near real-time, based on their information amplification potential. In *Proceedings of the 2012 international conference on social informatics* (pp. 70–77). Washington, DC, USA : IEEE Computer Society.

Carmel, D., Zwerdling, N., Guy, I., Ofek-Koifman, S., Har'el, N., Ronen, I., et al. (2009). Personalized social search based on the user's social network. In *Proceedings of the 18th acm conference on information and knowledge management* (pp. 1227–1236). New York, NY, USA : ACM.

Carpineto, C., Mori, R. de, Romano, G., et Bigi, B. (2001). An information-theoretic approach to automatic query expansion. *ACM Trans. Inf. Syst.*, *19*(1), 1–27.

Che Alhadi, A., Gottron, T., Kunegis, J., et Naveed, N. (2011). Livetweet : Microblog retrieval based on interestingness and an adaptation of the vector space model. In *Proc. text retrieval conference (TREC).*

Cheng, F., Zhang, X., He, B., Luo, T., et Wang, W. (2013). A survey of learning to rank for real-time twitter search. In *Proceedings of the 2012 international conference on pervasive computing and the networked world* (pp. 150–164). Berlin, Heidelberg : Springer-Verlag.

Choi, J., et Croft, W. B. (2012). Temporal models for microblogs. In *Proceedings of the 21st acm international conference on information and knowledge management* (pp. 2491–2494). New York, NY, USA : ACM.

Cleverdon, C. W., Mills, J., et Keen, M. (1966). Factors determining the performance of indexing systems.

Cohen, D., Amitay, E., et Carmel, D. (2007). Lucene and juru at trec 2007 : 1-million queries track. In *TREC'07 : 7th Text Retrieval Conference* (pp. -1–1).

Damak, F. (2013). Recherche d'information dans les microblogs : que manque-t-il aux approches classiques ? In *Rencontres Jeunes Chercheurs en Recherche d'Information (RJCRI), Neuchâtel, 03/04/2013-05/04/2013* (pp. 475–480). Association Francophone de Recherche d'Information et Applications (ARIA).

Damak, F., Jabeur, L. B., Cabanac, G., Pinel-Sauvagnat, K., Lechani, L., et Boughanem, M. (2011). IRIT at TREC Microblog 2011. In E. M. Voorhees et (Eds.), *Text REtrieval Conference (TREC), Gaithersburg, USA,*. National

Institute of Standards and Technology (NIST).

Damak, F., Pinel-Sauvagnat, K., et Cabanac, G. (2012). Recherche de microblogs : quels critères pour raffiner les résultats des moteurs usuels de RI ? In *Conférence francophone en Recherche d'Information et Applications (CORIA), Bordeaux, France, 21/03/2012-23/03/2012* (pp. 317–328). LABRI.

Damak, F., Pinel-Sauvagnat, K., Cabanac, G., et Boughanem, M. (2013). Effectiveness of State-of-the-art Features for Microblog Search. In *SAC'13 : ACM Symposium on Applied Computing.* ACM.

Diakopoulos, N. A., et Shamma, D. A. (2010). Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1195–1198). New York, NY, USA : ACM.

Dong, A., Chang, Y., Zheng, Z., Mishne, G., Bai, J., Zhang, R., et al. (2010). Towards recency ranking in web search. In *Proceedings of the third acm international conference on web search and data mining* (pp. 11–20). New York, NY, USA : ACM.

Dong, A., Zhang, R., Kolari, P., Bai, J., Diaz, F., Chang, Y., et al. (2010). Time is of the essence : improving recency ranking using twitter data. In *In www.*

Duan, Y., Jiang, L., Qin, T., Zhou, M., et Shum, H.-Y. (2010). An empirical study on learning to rank of tweets. In *Proceedings of the 23rd international conference on computational linguistics* (pp. 295–303).

Efron, M. (2010). Hashtag retrieval in a microblogging environment. In *Proceedings of the 33rd international acm sigir conference on research and development in information retrieval* (pp. 787–788). New York, NY, USA : ACM.

Efron, M. (2011a). Information search and retrieval in microblogs. In (Vol. 62, pp. 996–1008). New York, NY, USA : John Wiley & Sons, Inc.

Efron, M. (2011b). The university of illinois graduate school of library and information science at TREC 2011. In *TREC'11 : 20th Text Retrieval Conference.* National Institute of Standards and Technology (NIST).

Efron, M., et Golovchinsky, G. (2011). Estimation methods for ranking recent information. In *Proceedings of the 34th international acm sigir conference on research and development in information retrieval* (pp. 495–504). New York, NY, USA : ACM.

Efron, M., Organisciak, P., et Fenlon, K. (2012). Improving retrieval of short texts through document expansion. In *Proceedings of the 35th international acm sigir conference on research and development in information retrieval* (pp. 911–920). New York, NY, USA : ACM.

Endarnoto, S., Pradipta, S., Nugroho, A., et Purnama, J. (2011). Traffic condition information extraction amp ; visualization from social media twitter for android mobile application. In *Electrical engineering and informatics (iceei),*

2011 international conference on (p. 1-4).

Feng, W., et Wang, J. (2013). Retweet or not ? : Personalized tweet re-ranking. In *Proceedings of the sixth acm international conference on web search and data mining* (pp. 577–586). New York, NY, USA : ACM.

Ferguson, P., O'Hare, N., Lanagan, J., Phelan, O., et McCarthy, K. (2012). An investigation of term weighting approaches for microblog retrieval. In *Proceedings of the 34th european conference on advances in information retrieval* (pp. 552–555). Berlin, Heidelberg : Springer-Verlag.

Frank, J. R., Bauer, S. J., Kleiman-Weiner, M., Roberts, D. A., Tripuraneni, N., Zhang, C., et al. (2013). Evaluating stream filtering for entity profile updates for trec 2013. In *TREC'13 : 22th Text Retrieval Conference.*

Frank, J. R., Kleiman-Weiner, M., Roberts, D. A., Niu, F., Zhang, C., Re, C., et al. (2012). Building an Entity-Centric stream filtering test collection for TREC 2012. In *Proc. of trec.* National Institute of Standards and Technology (NIST).

Furnas, G. W., Deerwester, S., Dumais, S. T., Landauer, T. K., Harshman, R. A., Streeter, L. A., et al. (1988). Information retrieval using a singular value decomposition model of latent semantic structure. In *Proceedings of the 11th annual international acm sigir conference on research and development in information retrieval* (pp. 465–480). New York, NY, USA : ACM.

Furnas, G. W., Landauer, T. K., Gomez, L. M., et Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Commun. ACM*, *30*(11), 964–971.

Gudivada, V., Raghavan, V., Grosky, W. I., et Kasanagottu, R. (1997). Information retrieval on the world wide web. *Internet Computing, IEEE*, *1*(5), 58-68.

Hall, M. A., et Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *IEEE Trans. on Knowl. and Data Eng.*, *15*(6), 1437–1447.

Han, B., et Baldwin, T. (2011). Lexical normalisation of short text messages : Makn sens a #twitter. In *Proceedings of the 49th annual meeting of the association for computational linguistics : Human language technologies - volume 1* (pp. 368–378). Stroudsburg, PA, USA : Association for Computational Linguistics.

Hatzivassiloglou, V., et McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics* (pp. 174–181). Stroudsburg, PA, USA : Association for Computational Linguistics.

Jabeur, L., Tamine, L., et Boughanem, M. (2012). Featured tweet search : Modeling time and social influence for microblog retrieval. In *IEEE/WIC/ACM International Conference on Web Intelligence, Macau, China* (pp. 166–173). IEEE Computer Society - Conference Publishing Services.

Jansen, B. J., Zhang, M., Sobel, K., et Chowdury, A. (2009a). Micro-blogging as online word of mouth branding. In *Chi '09 extended abstracts on human factors in computing systems* (pp. 3859–3864). New York, NY, USA : ACM.

Jansen, B. J., Zhang, M., Sobel, K., et Chowdury, A. (2009b). Twitter power : Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.*, *60*(11), 2169–2188.

Java, A., Song, X., Finin, T., et Tseng, B. (2007). Why we twitter : understanding microblogging usage and communities. In *WebKDD'07 : Proceedings of the 9th webkdd and 1st sna-kdd 2007 workshop on web mining and social network analysis* (pp. 56–65).

Joachims, T. (2005). A support vector method for multivariate performance measures. In *Proceedings of the 22nd international conference on machine learning* (pp. 377–384). New York, NY, USA : ACM.

Jones, K. S., et Rijsbergen, C. van. (1976). *Information retrieval test collections* (Rapport technique).

Karamuftuoglu, M. (1998). Collaborative information retrieval : toward a social informatics view of ir interaction. *J. Am. Soc. Inf. Sci.*, *49*(12), 1070–1080.

Kazai, G., et Milic-Frayling, N. (2008). Trust, authority and popularity in social information retrieval. In *Proceedings of the 17th ACM conference on information and knowledge management* (pp. 1503–1504). New York, NY, USA : ACM.

Klas, C.-P., et Fuhr, N. (2000). A new effective approach for categorizing Web documents. In *Proceedings of the 22th bcs-irsg colloquium on ir research.*

Koolen, M., Kazai, G., et Craswell, N. (2009). Wikipedia pages as entry points for book search. In *In proceedings of the second acm international conference on web search and data mining (wsdm 2009).* ACM Press.

Kopliku, A., Damak, F., Pinel-Sauvagnat, K., et Boughanem, M. (2011). Interest and Evaluation of Aggregated Search. In *IEEE/WIC/ACM International Conference on Web Intelligence, Lyon.* ACM.

Korfiatis, N., Poulos, M., et Bokos, G. (2006). Evaluating authoritative sources using social networks : an insight from wikipedia. *Online Information Review*, *30*(3), 252-262.

Kumar, N., et Carterette, B. (2013). Time based feedback and query expansion for twitter search. In *Proceedings of the 35th european conference on advances in information retrieval* (pp. 734–737). Berlin, Heidelberg : Springer-Verlag.

Kwak, H., Lee, C., Park, H., et Moon, S. (2010). What is twitter, a social network or a news media ? In *Proceedings of the 19th international conference on world wide web* (pp. 591–600). New York, NY, USA : ACM.

Lampos, V., et Cristianini, N. (2010). Tracking the flu pandemic by monitoring the social web. In *Cognitive information processing (cip), 2010 2nd international*

workshop on (p. 411-416).

Lee, C., Kwak, H., Park, H., et Moon, S. (2010). Finding influentials based on the temporal order of information adoption in twitter. In *Www'10 : Proceedings of the 19th international conference on world wide web* (pp. 1137–1138). New York, NY, USA : ACM.

Li, R., Lei, K. H., Khadiwala, R., et Chang, K.-C. (2012). Tedas : A twitter-based event detection and analysis system. In *Data engineering (icde), 2012 ieee 28th international conference on* (p. 1273-1276).

Li, Y., Luk, W. P. R., Ho, K. S. E., et Chung, F. L. K. (2007). 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* (pp. 797–798). New York, NY, USA : ACM.

Lin, Y., Li, Y., Xu, W., et Guo, J. (2012). Microblog retrieval based on term similarity graph. In *Computer science and network technology (iccsnt), 2012 2nd international conference on* (p. 1322-1325).

Lv, Y., et Zhai, C. (2009). Positional language models for information retrieval. In *Proceedings of the 32nd international acm sigir conference on research and development in information retrieval* (pp. 299–306). New York, NY, USA : ACM.

Macdonald, C., et Ounis, I. (2006). Voting for candidates : Adapting data fusion techniques for an expert search task. In *Proceedings of the 15th acm international conference on information and knowledge management* (pp. 387–396). New York, NY, USA : ACM.

Magnani, M., Montesi, D., et Rossi, L. (2012). Conversation retrieval for microblogging sites. *Inf. Retr.*, *15*(3-4), 354-372.

Mandl, T. (2007). *Recent developments in the evaluation of information retrieval systems : Moving towards diversity and practical relevance.*

Manning, C. D., Raghavan, P., et Schütze, H. (2008). *Introduction to information retrieval*. New York, NY, USA : Cambridge University Press.

Massoudi, K., Tsagkias, E., Rijke, M. de, et Weerkamp, W. (2011). Incorporating query expansion and quality indicators in searching microblog posts. In *Ecir 2011 : 33rd european conference on information retrieval* (pp. 362–367). Dublin : Springer.

Mayfield, J., et McNamee, P. (2003). Single n-gram stemming. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval* (pp. 415–416). New York, NY, USA : ACM.

McCreadie, R., et Macdonald, C. (2013). Relevance in microblogs : Enhancing tweet retrieval using hyperlinked documents. In *Proceedings of the 10th conference on open research areas in information retrieval* (pp. 189–196). Paris, France, France : Le centre de hautes études internationales d'informatique documen-

taire.

Metzler, D., et Cai, C. (2011). USC/ISI at TREC 2011 : Microblog Track (Notebook Version). In *TREC'11 : 20th Text Retrieval Conference.* National Institute of Standards and Technology (NIST).

Missen, M. M. S., Boughanem, M., et Cabanac, G. (2009, juin). Challenges for Sentence Level Opinion Detection in Blogs (regular paper). In *International Conference on Computer and Information Science (ICIS), Shanghai, China, 01/06/2009-03/06/2009* (pp. 347–351). IEEE Computer Society.

Miyanishi, T., Seki, K., et Uehara, K. (2013). Combining recency and topic-dependent temporal variation for microblog search. In *Ecir* (p. 331-343).

Nagmoti, R., Teredesai, A., et De Cock, M. (2010). Ranking approaches for microblog search. In *Proceedings of the 2010 ieee/wic/acm international conference on web intelligence and intelligent agent technology* (pp. 153–157). Washington, USA : IEEE Computer Society.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., et Smith, N. A. (2010). From tweets to polls : Linking text sentiment to public opinion time series. In *Icwsm.*

Okazaki, M., et Matsuo, Y. (2010). Semantic twitter : analyzing tweets for real-time event notification. In *Proceedings of the 2008/2009 international conference on social software : recent trends and developments in social software* (pp. 63–74). Berlin, Heidelberg : Springer-Verlag.

Ounis, I., Lin, J., et Soboroff, I. (2011). Overview of the TREC-2011 Microblog Track. In *TREC'11 : 20th Text Retrieval Conference.*

Ounis, I., Lin, J., et Soboroff, I. (2012). Overview of the TREC-2012 Microblog Track. In *TREC'12 : 21th Text Retrieval Conference.*

Pang, B., et Lee, L. (2008). Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, *2*(1-2), 1–135.

Peters, I., Kipp, M. E. I., Heck, T., Gwizdka, J., Lu, K., Neal, D. R., et al. (2011). Social tagging & folksonomies : Indexing, retrievingâĂę and beyond ? *Proceedings of the American Society for Information Science and Technology*, *48*(1), 1–4.

Phelan, O., McCarthy, K., et Smyth, B. (2009). Using twitter to recommend real-time topical news. In *Recsys'09 : Proceedings of the third acm conference on recommender systems* (pp. 385–388). New York, NY, USA : ACM.

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

Ramage, D., Dumais, S. T., et Liebling, D. J. (2010). Characterizing microblogs with topic models. In *ICWSM'10* (pp. -1–1).

Ravikumar, S., Balakrishnan, R., et Kambhampati, S. (2012). Ranking tweets considering trust and relevance. In *Proceedings of the ninth international workshop on information integration on the web* (pp. 4 :1–4 :4). New York, NY, USA :

ACM.

Robertson, S. (2004). Understanding inverse document frequency : On theoretical arguments for idf. *Journal of Documentation*, *60*, 2004.

Robertson, S., et Sparck Jones, K. (1988). Document retrieval systems. In P. Willett (Ed.), (pp. 143–160). London, UK, UK : Taylor Graham Publishing.

Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., et Gatford, M. (1996). Okapi at trec-3. In (pp. 109–126).

Rocchio, J. J. (1971). Relevance feedback in information retrieval.

Sakaki, T., Okazaki, M., et Matsuo, Y. (2010). Earthquake shakes twitter users : real-time event detection by social sensors. In *Proceedings of the 19th international conference on world wide web* (pp. 851–860). New York, NY, USA : ACM.

Salton, G. (1968). *A comparison between manual and automatic indexing methods* (Rapport technique). Ithaca, NY, USA.

Salton, G., et Buckley, C. (1997). Readings in information retrieval. In K. Sparck Jones et P. Willett (Eds.), (pp. 355–364). San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

Salton, G., Wong, A., et Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, *18*(11), 613–620.

Sanderson, M. (2010). Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, *4*(4), 247-375.

Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D., et Sperling, J. (2009). Twitterstand : news in tweets. In *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems* (pp. 42–51). New York, NY, USA : ACM.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International conference on new methods in language processing* (p. 44-49). Manchester, UK.

Segaran, T., et Hammerbacher, J. (2009). *Beautiful Data : The Stories Behind Elegant Data Solutions* (Original éd.). O'Reilly Media. Paperback.

Shamma, D. A., Kennedy, L., et Churchill, E. F. (2009). Tweet the debates : Understanding community annotation of uncollected sources. In *Proceedings of the first sigmm workshop on social media* (pp. 3–10). New York, NY, USA : ACM.

Song, S., Li, Q., et Zheng, N. (2010). A spatio-temporal framework for related topic search in micro-blogging. In *Proceedings of the 6th international conference on active media technology* (pp. 63–73). Berlin, Heidelberg : Springer-Verlag.

Student. (1908). The probable error of a mean. *Biometrika*, *6*(1), 1–25.

Sturges, H. A. (1926). The Choice of a Class Interval. *Journal of the American Statistical Association*, *21*(153), 65–66.

Teevan, J., Ramage, D., et Morris, M. R. (2011). #twittersearch : a comparison of microblog search and web search. In *Wsdm'11 : Proceedings of the fourth acm international conference on web search and data mining* (pp. 35–44). New York, NY, USA : ACM.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., et Welpe, I. M. (2010). Predicting elections with twitter : What 140 characters reveal about political sentiment. In *Icwsm.*

Uysal, I., et Croft, W. B. (2011). User oriented tweet ranking : a filtering approach to microblogs. In C. Macdonald, I. Ounis, et I. Ruthven (Eds.), *Cikm* (p. 2261-2264). ACM.

Vechtomova, O., et Wang, Y. (2006). A study of the effect of term proximity on query expansion. *J. Information Science*, *32*(4), 324-333.

Voorhees, E. M. (2006). Overview of the trec 2006. In *TREC'06 : 6th Text Retrieval Conference.*

Vosecky, J., Leung, K. W.-T., et Ng, W. (2012). Searching for quality microblog posts : Filtering and ranking based on content analysis and implicit links. , 397-413.

Weng, J., Lim, E.-P., Jiang, J., et He, Q. (2010). Twitterrank : finding topic-sensitive influential twitterers. In *Wsdm'10 : Proceedings of the third acm international conference on web search and data mining* (pp. 261–270). New York, NY, USA : ACM.

Wilson, T., Wiebe, J., et Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (pp. 347–354). Stroudsburg, PA, USA : Association for Computational Linguistics.

Xu, S., Bao, S., Cao, Y., et Yu, Y. (2007). Using social annotations to improve language model for information retrieval. In *Proceedings of the sixteenth acm conference on conference on information and knowledge management* (pp. 1003–1006). New York, NY, USA : ACM.

Yamaguchi, Y., Takahashi, T., Amagasa, T., et Kitagawa, H. (2010). Turank : Twitter user ranking based on user-tweet graph analysis. In *Wise'10* (p. 240-253).

Yen, S.-J., et Lee, Y.-S. (2006). Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. In *Intelligent control and automation* (Vol. 344, p. 731-740). Springer Berlin / Heidelberg.

Yuan, Q., Cong, G., et Thalmann, N. M. (2012). Enhancing naive bayes with various smoothing methods for short text classification. In *Proceedings of the 21st international conference companion on world wide web* (pp. 645–646). New York, NY, USA : ACM.

Zhao, L., Zeng, Y., et Zhong, N. (2011). A weighted multi-factor algorithm for

microblog search. In *Proceedings of the 7th international conference on active media technology* (pp. 153–161). Berlin, Heidelberg : Springer-Verlag.

Zhongyuan, H., Xuwei, L., Muyun, Y., Hoaliang, Q., Sheng, L., et Tiejun, Z. (2012). HIT at Trec 2012 Microblog Track. In *TREC'12 : 21th Text Retrieval Conference.* National Institute of Standards and Technology (NIST).

Résumé

Notre travail se situe dans le contexte de recherche d'information (RI) sociale et s'intéresse plus particulièrement à la recherche de microblogs. Les microblogs sont des messages de faible longueur à travers lesquels les utilisateurs publient des informations sur différents sujets : des opinions, des événements, des statuts... Les microblogs occupent aujourd'hui une part considérable de l'information générée sur le web. Dans Twitter, la plate-forme de microblogging la plus populaire, le nombre de microblogs par jour peut atteindre 500 millions. Les microblogs ont une forme différente des traditionnels documents. Leur taille est réduite par rapport aux blogs et aux articles publiés sur le web (140 caractères pour Twitter). De plus, les microblogs peuvent contenir une syntaxe spécifique telle que les #hashtags, les @citations ou bien encore des URLs. Les plateformes de microblogging représentent également un modèle de réseau social différent des autres réseaux sociaux. Les relations entre les utilisateurs ne sont pas forcément réciproques et les abonnements sont sans restrictions entre microbloggeurs.

Les utilisateurs de plateformes de microblogging, outre la publication de microblogs, effectuent également des recherches. Les motivations de ces recherches sont diverses. Certaines sont similaires à la recherche sur le web (comme par exemple la recherche d'actualités), et d'autres sont spécifiques à la recherche de microblogs (comme par exemple la recherche temps réel ou d'informations sociales). Dans Twitter, 1,6 milliards de requêtes sont ainsi émises chaque jour.

Les modèles de RI doivent s'adapter aux spécificités des microblogs : fraîcheur, aspect social et spécificités syntaxiques doivent ainsi être pris en compte. C'est dans ce contexte de recherche d'information dans les microblogs que se situent plus particulièrement nos travaux. Nous nous plaçons plus précisément dans le cadre de la recherche adhoc. L'objectif est de retrouver les microblogs répondant à un besoin d'information spécifié par un utilisateur.

Nos travaux visent à améliorer la qualité des résultats de recherche d'information adhoc dans les microblogs. Nos contributions se situent à plusieurs niveaux :

-Afin de déterminer exactement les facteurs limitant les performances des modèles de recherche classiques dans un corpus de microblogs, nous avons mené à une analyse de défaillance d'un modèle de recherche usuel. Nous avons sélectionné les microblogs pertinents mais non retrouvés par le modèle de recherche. Ensuite, nous avons identifié les facteurs empêchant leur restitution. Nous avons trouvé que le problème principal vient de la concision des microblogs. Cette concision engendre une correspondance limitée entre les termes des microblogs et les termes des requêtes, même s'ils sont sémantiquement similaires.

-Afin de compenser l'impact de la concision des microblogs, nous avons proposé et testé plusieurs solutions. Nous avons proposé d'étendre les requêtes (i) en exploitant des ressources de type actualités, (ii) en utilisant la base lexicale Wordnet, (iii) en applicant des techniques de réinjection de pertinence de l'état de l'art qui ont souvent prouvé leur efficacité : Rocchio pour identifier les termes susceptibles de ramener la pertinence ainsi que pour la pondération des termes de la nouvelle requête, et le mécanisme naturel d'extension

de requêtes du modèle BM25. Dans Rocchio, nous avons testé différentes méthodes de calcul de poids de termes d'expansion. Nous avons enfin étendu les microblogs grâce aux liens (URLs) qu'ils contiennent. Nos expérimentations ont montré que l'emploi des URLs et l'expansion de requêtes sont primordiales pour la RI dans les microblogs. La plupart de ces expérimentations (expansion de requêtes et de microblogs) ont été réalisées en se basant sur le modèle vectoriel et sur le modèle probabiliste comme modèle de restitution. Ceci nous a permis de comparer les comportements des deux modèles sur les microblogs et avec les deux types d'expansion. De manière générale, nous avons trouvé que le modèle vectoriel est plus performant que modèle probabiliste au niveau de la sélection des microblogs pertinents (meilleur rappel). Cependant, le modèle probabiliste met plus en valeur les microblogs pertinents restitués par rapport à tous les microblogs restitués (meilleure précision).

-Un deuxième volet de notre travail concerne l'étude des critères utilisés pour identifier les microblogs pertinents. Nous avons repris les critères souvent utilisés dans l'état de l'art (critères de contenu, critères sur l'importance des auteurs, critères sur les URLs) et nous les avons évalués. Nous avons réalisé cette analyse selon 3 axes. Dans le premier axe, nous avons analysé l'impact de la combinaison des scores des critères avec le score de pertinence du contenu, calculé avec un modèle de RI usuel. Dans le deuxième axe, nous avons étudié le comportement des critères dans les documents pertinents et les avons comparés avec leurs comportements dans les documents non pertinents. Dans le troisième axe, nous avons utilisé des techniques d'apprentissage ainsi que des algorithmes de sélection de critères qui peuvent être utiles en entrée de ces techniques d'apprentissages. De manière générale, nous avons montré que les critères en relation avec les URLs publiées dans les tweets sont les plus discriminants. Les critères liés aux auteurs ne reflètent pas la pertinence.

-Afin de prendre en compte l'aspect temporel dans la restitution des microblogs pertinents vis-à-vis d'un besoin d'information, nous avons proposé trois méthodes qui intègrent le temps dans le calcul de la pertinence. Cette intégration du temps n'a cependant pas montré son intérêt dans nos méthodes.

Pour réaliser nos expérimentations, nous nous sommes basés sur le corpus fourni par la campagne d'évaluation internationale TREC (Text Retrieval Conference) dans la tâche Microblogs des années 2011 et 2012. Nos différentes contributions ont également fait l'objet de participations aux trois tâches de Microblogs de TREC (2011, 2012 et 2013).

Title

Étude des facteurs de pertinence dans la recherche de microblogs.

Abstract

This work deals with the context of social information retrieval (IR), more particularly the retrieval of microblogs. Microblogs are messages of short length. They contain information on various topics :opinions, events, articles... Microblogs represent a significant part of the information generated on the Web. In the case of Twitter, the most popular platform, the number of microblogs can reach 500 million per day. Microblogs have a different form from traditional documents. Their length is reduced compared to traditional blogs and

articles on the web (only 140 characters in the case of Twitter). Moreover, microblogs can have specific syntax such as #hashtags, @mentions or shortened URLs... Microblogging platforms are a social network model different from other social networks. Relationships between users are not necessarily reciprocal and subscriptions are unrestricted between microbloggers. Users of microblogging platforms do not only produce but they also search for information. The motivations of this research are diverse. Some are inspired from Web search (e.g. the search for news) and others are specific to the search for microblogs (e.g. real-time search or social information). In Twitter, 1.6 billion queries are issued every day. Though, the IR models must adapt to the specificities of microblogs : freshness, social aspect and syntactic characteristics must therefore be taken into account. The aim of our work is to improve the quality of the results of adhoc information retrieval in microblogs. Our contributions are at several levels :

- In order to accurately determine the factors limiting the performance of conventional models of search in a corpus of microblogs, we conducted an analysis of failure of a conventional model search. We selected relevant microblogs. However, they are not found by the search pattern. Then, we identified the factors preventing their return. We found that the main problem is the shortness of microblogs.

- To offset the impact of the shortness of microblogs, we proposed and tested several solutions : to extend the queries by (i) exploiting news articles, (ii) using the WordNet lexical database, (iii) applying techniques of relevance feedback of the state of art which often proved effective : Rocchio to identify terms likely to bring relevance and for weighting the terms of the new query, and the natural extension mechanism queries of the BM25 model. Using Rocchio, we tested different methods of calculating the weight of expansion terms. We finally extended microblogs thanks to the links (URLs) they contain. Our experiments have shown that the use of URLs and the expansion of the query are crucial for IR in microblogs. Most of these experiments (expansion of queries and microblogs) were performed on the basis of the vector model and the probabilistic model, as a model of restitution. This allowed us to compare the behavior of the two models on microblogs and with the two types of expansion. In general, we found that the Vector Space Model is more efficient than the probabilistic one in the selection of relevant microblogs (better recall). However, the probabilistic model puts more value on relevant microblogs returned over all returned microblogs (better precision).

- A second part of our work is concerned with the study of the features used to identify relevant microblogs. We selected the features often used in the state of art (content features, features on the importance of authors, URLs features and quality features).Then, we evaluated them. We conducted this analysis in 3 axes. In the first axis, (i) we studied the behavior of the features in the relevant documents and compared them with their behavior in non-relevant documents. In the second axis, (ii) we analyzed the impact of the combination of the features scores with the content's score, calculated with a model of conventional IR. In the third axis, (iii) we used learning techniques as well as algorithms of feature selection that may be useful as input to the learning techniques. In general, we have shown that the features related to URLs posted in tweets are the most discriminating.

The features related to the authors do not reflect the relevance.

- To take into account the temporal aspect when selecting relevant microblogs, we have proposed three methods that incorporate time in the calculation of relevance. However, this integration of time did not show any positive impact in our methods.

To perform our experiments, we used the corpus provided by TREC (Text Retrieval Conference) international survey in the task Microblogs for the years 2011 and 2012. Our various contributions have also been the subject of participations for the three tasks of Microblogs TREC (2011, 2012 and 2013).

ISBN : ␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣