Recherche d'information dans les microblogs : que manque-t-il aux approches classiques ?

Firas Damak[1]

Université de Toulouse – IRIT UMR 5505 CNRS
118 route de Narbonne, F-31062 Toulouse cedex 9
firas.damak@irit.fr

RÉSUMÉ. *Nous nous intéressons dans cet article à la recherche d'information dans les microblogs. Les modèles de RI classiques, conçus pour des textes plus longs que les 140 caractères d'un microblog, ne sont pas forcément adaptés pour ces derniers. Une analyse de leurs résultats nous a permis d'identifier la différence de vocabulaire entre les microblogs et la requête comme étant la raison principale de leur manque de performance. Pour améliorer la qualité de la recherche, nous proposons d'étendre les microblogs grâce au texte des URL qu'ils contiennent, et également d'étendre les requêtes avec WordNet ou en utilisant des articles de presse. Les résultats montrent l'intérêt de l'extension des tweets, celui de l'extension des requêtes restant à prouver.*

ABSTRACT. *This paper deals with information retrieval in microblogs. Classical IR models were originally designed for texts longer than 140 characters (i.e., the maximum microblog length). They fail to perform well with microblog corpora. The failure analysis we conducted shows that the vocabulary mismatch is the main problem we have to deal with. We thus propose to extend tweets with the text of the URL they contain, and to extend queries in two ways (WordNet and news articles). Results show the interest of tweet extension, whereas interest of query expansion is still to be proved.*

MOTS-CLÉS : *Microblog, Twitter, moteur de recherche, analyse des défaillances*

KEYWORDS: *Microblog, Twitter, search engine, failure analysis*

1. Directeur de thèse : Mohand BOUGHANEM

1. Introduction

Depuis quelques années, la quantité d'information publiée sur les plateformes de microblogging augmente exponentiellement. Prenons par exemple le cas de Twitter[1], la plateforme de microblogging la plus populaire sur le web, 200 millions de microblogs ou tweets sont publiés chaque jour[2]. 1,6 milliards de requêtes sont également émises chaque jour[3].

Plusieurs travaux de recherche se sont focalisés sur la problématique de la recherche d'information dans les microblogs (Ounis *et al.*, 2011). La majorité des approches présentes dans la littérature ont traduit les spécificités des plateformes de microblogging *via* diverses caractéristiques (c.-à-d. *features*) telles que la popularité de l'auteur du microblog, la qualité du langage utilisé, la fraîcheur du microblog, etc. en supplément à la pertinence du contenu du microblog. L'emploi de ces caractéristiques 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 (Damak *et al.*, 2013). Cependant ces améliorations demeurent dépendantes du modèle de RI : les caractéristiques utilisées n'auront aucun effet si ce dernier ne capture pas le maximum de microblogs pertinents vis-à-vis d'un besoin en information.

Dans cet article nous réalisons une analyse pour déterminer les problèmes rencontrés par les modèles de RI classiques, qui affectent la qualité de leurs résultats sur des corpus de microblogs. Dans un deuxième temps, nous proposons et testons quelques hypothèses dans l'objectif d'améliorer les performances des moteurs de recherche de microblogs. La suite de cet article est organisée de la façon suivante : la section 2 dresse un état de l'art des approches actuelles pour la recherche d'information (RI) dans les microblogs. Dans la section 3, nous discutons les différents problèmes affectant l'efficacité des modèles de RI, observés à l'issue de notre analyse. La section 4 détaille quelques propositions permettant d'améliorer la qualité des moteurs. Finalement, nous discutons les résultats obtenus dans la section 5 et synthétisons les pistes prometteuses.

2. État de l'art sur la recherche dans les microblogs

De nombreuses approches de sélection de microblogs pertinents répondant à un besoin en information ont été proposées récemment dans la littérature. Parmi ces approches, celle de (Metzler *et al.*, 2005; Metzler *et al.*, 2011) utilise une technique d'apprentissage pour classer les microblogs selon des caractéristiques telles que le score de pertinence du contenu, l'intervalle de temps entre la diffusion du microblog et la soumission de la requête, la présence d'un hashtag, la présence d'une URL, la longueur du microblog et le pourcentage de termes qui ne font pas partie du lexique anglais. Duan *et al.* (2010) ont employé le modèle Okapi-BM25 pour mesurer la pertinence du contenu, certaines caractéristiques spécifiques à Twitter (p. ex., la popularité des

1. http://twitter.com
2. http://blog.twitter.com/2011/06/200-million-tweets-per-day.html
3. http://engineering.twitter.com/2011/05/engineering-behind-twitters\
-new-search.html

tweets, la fréquence des retweets, la fréquence des hashtags, la longueur des tweets, la présence d'URL) ainsi que des caractéristiques reflétant l'importance des auteurs (p. ex., nombre d'abonnés, nombre de mentions). SVM-RANK (Joachims, 2005) a également été utilisé pour construire le modèle de recherche.

Cependant, même si ces caractéristiques peuvent aider à la sélection de microblogs pertinents, leur pouvoir discriminant est limité par l'extraction préalable des microblogs « candidats », qui se base principalement sur la similarité de contenu entre un microblog et un besoin en information exprimé sous forme de requête. Le modèle de RI employé peut être défaillant quant à la sélection des microblogs pertinents (rappel faible). Dans ce cas, la prise en compte des caractéristiques comme celles employées dans les approches déjà citées ne permettra pas, pour autant, de restituer des microblogs qui n'auraient pas été sélectionnés. La problématique que nous considérons dans cet article est donc : *dans quelle mesure est-il possible d'augmenter le rappel, étape préalable à l'emploi des caractéristiques pour réordonner la liste des microblogs candidats ?*

Dans cet article, nous réalisons une analyse pour déterminer quels sont les problèmes réduisant l'efficacité des modèles de RI dans le cas de la recherche de microblogs. Nous proposons ensuite des pistes pour améliorer le rappel.

3. Analyse des défaillances des modèles de RI classiques dans le cadre de la recherche de microblogs

3.1. *Méthodologie*

Notre analyse a pour but de déterminer quels sont les facteurs pénalisant le rappel des modèles de RI. 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 (Ounis *et al.*, 2011) et sur les *topics* fournis aux participants de la tâche de 2012. Dans cette tâche, il s'agit, pour un moteur de recherche, de restituer les tweets dont le contenu satisfait un besoin en information sous forme de mots-clés. Nos expérimentations ont porté sur le moteur de recherche *open source* Lucene[4], qui implémente une version modifiée du modèle vectoriel. Le modèle considère également la fréquence des termes de la requête dans le texte (Cohen *et al.*, 2007). La version de Lucene utilisée intègre le lemmatiseur Porter et utilise une liste de mots vides.

3.2. *Observations préliminaires*

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) Tout d'abord, nous avons souvent constaté une absence totale des termes des requêtes dans les tweets pertinents non retrouvés. Nous avons observé qu'un nombre

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

important de tweets traite du sujet de la requête sans avoir, pour autant, aucun terme en commun avec cette dernière. Nous pouvons par exemple citer le topic `British Government Cuts` 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 dans 37 topics sur 60 (62 %), à hauteur de 50 % des tweets pertinents non restitués (ce problème apparaît pour au moins 1 500 tweets pertinents non retrouvés parmi les 3 696 tweets pertinents non retrouvés sur toutes les requêtes).

2) Un autre problème concerne les noms propres et les entités nommées, qui sont orthographiés de différentes manières. Par exemple, pour le topic `Glen Beck`, dans certains tweets pertinents les auteurs utilisaient `Glenn` plutôt que `Glen`. Ce phénomène est présent dans 5 topics sur 60 (8 %), à hauteur de 50 % des tweets pertinents non restitués (dans au moins 200 tweets non retrouvés sur les 3 696).

3) Des termes différents ne sont pas appariés, alors qu'ils relèvent d'une même racine. Par exemple, pour le topic `somalian piracy` étaient présents dans les tweets jugés pertinents les termes `pirates` ou `pirate`. Ce phénomène est présent clairement dans 2 topics sur 60, à hauteur de 50 % des tweets pertinents non restitués. Cela représente au moins 100 tweets pertinents non retrouvés sur les 3 696.

4) Enfin les acronymes sont écrits de différentes manières. C'est le cas du topic `FDA approval of drugs` pour lequel les tweets pertinents contenaient également l'acronyme `USFDA`. Ce phénomène est présent clairement dans 2 topics sur 60, à hauteur de 50% des tweets pertinents non restitués. Cela représente au moins 50 tweets pertinents non retrouvés sur les 3696.

Suite à ces observations, nous avons voulu savoir pour combien de tweets la considération des contenus des URL qu'ils contiennent permettrait de gérer ce problème de vocabulaire. Nous avons donc analysé les contenus des URL des tweets pertinents non retrouvés et nous avons constaté que leur prise en compte aiderait à retrouver des tweets pertinents dans 15 topics sur 60 (25 %). Cela représente au moins 500 tweets pertinents non retrouvés.

4. Contributions et expérimentations

À l'issue de l'analyse des URL publiées dans les tweets pertinents, nous avons remarqué que la considération des pages web pointées par les URL pourrait améliorer la restitution des tweets pertinents. Une première proposition consiste alors en l'indexation d'un tweet selon 1) son contenu ainsi que 2) le contenu des documents pointés par les URL présents dans le tweet (2 646 611 tweets contiennent une URL dans la collection). Nous avons commencé par considérer les deux champs dans la recherche avec les requêtes originales (ReqO). Le Tab. 1 montre que le run ReqOTweetURL qui considère le contenu des URL atteint un rappel de 0,7171 (13,10 % d'amélioration par rapport à la baseline). Une deuxième solution consiste en l'extension de la requête. Nous avons considéré deux sources : Wordnet (en étendant chaque terme de la requête avec le

Run	Topic	Champ utilisé	Rappel
Baseline	ReqO	Tweets	0,6340
ReqOTweetURL	ReqO	Tweets+URL	0,7171*

Tableau 1. *Comparaison de l'utilisation des champs Tweets et Tweets+URL avec la requête originale. * indique une différence significative entre le rappel de la baseline et celui du run considéré, selon le test t de Student pairé et bilatéral avec $p < 0,05$.*

Run	Requête	Champ utilisé	Rappel†	Rappel‡
ReqWN(Pond)Tweet	ReqO+WN	Tweets	0.6305	0.6362
ReqWN(Pond)TweetURL	ReqO+WN	Tweets+URL	0.7179	0.7201
Req3Act(Pond)Tweet	ReqO+3Act	Tweets	0.5691	0.5923
Req3Act(Pond)TweetURL	ReqO+3Act	Tweets+URL	0.6310	0.6970
Req7Act(Pond)Tweet	ReqO+7Act	Tweets	0.5985	0.6156
Req7Act(Pond)TweetURL	ReqO+7Act	Tweets+URL	0,5561	0,7032

Tableau 2. *Récapitulatif des différents runs testés sans (†) et avec (‡) pondération des termes ajoutés aux requêtes (1 500 résultats par requête).*

premier synset retrouvé) et les actualités correspondantes aux topics (avec 3 « 3Act » ou 7 « 7Act » mots-clés des articles de journaux renvoyés à travers l'API Alchemy[5]).

Dans un premier temps, nous ne pondérons pas les termes ajoutés dans la requête. Les résultats sont présentés dans le Tab. 2(†). Ils sont comparés aux résultats du run ReqOTweetURL présenté précédemment : aucune amélioration significative n'est constatée. Dans un second temps, nous avons pondéré les termes ajoutés aux requêtes avec un poids de 0,8 (choix arbitraire pour ces premières expérimentations). Les résultats sont présentés dans le Tab. 2(‡). Nous constatons que la pondération améliore les résultats par rapport à la non pondération (Recall‡ par rapport au Recall†). Cependant, aucune amélioration significative n'est à remarquer par rapport au run.

5. Discussion

Au niveau des analyses des facteurs limitant l'efficacité d'un moteur de recherche sur les microblogs, nous avons montré 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, sémantiquement similaires. Nous avons identifié d'autres facteurs pour lesquels des solutions doivent être trouvées, comme les noms écrits de différentes manières, ou bien les termes qui peuvent être concaténés.

5. http://www.alchemyapi.com/

Concernant nos expérimentations, la conclusion principale est que la prise en compte des contenus des URL paraît indispensable pour la recherche des microblogs (une amélioration de 13,10 % par rapport à la baseline). L'extension de requête, que ce soit sur le champ Tweet ou sur les champs Tweet+URL, n'améliore généralement pas le rappel du run ReqOTweetURL (Tab. 1) utilisant les deux champs Tweets et URL. Ceci est probablement causé par le rajout de termes « parasites » à la requête initiale, et des expérimentations complémentaires sont nécessaires.

6. Conclusion

Dans cet article nous avons réalisé une analyse de défaillance afin de déterminer quels sont les facteurs limitant l'efficacité d'un modèle de recherche classique de RI dans le cas de recherche de microblogs. Nous avons trouvé que la concision des microblogs (140 caractères au plus) en est le facteur principal. Nous avons proposé quelques solutions pour enrichir les microblogs (contenu des URL) et les requêtes (WordNet et les mots clés des articles de journaux). Les expérimentations ont montré que le fait d'utiliser les contenus des URL améliore les résultats de façon significative. Les méthodes d'extension de requêtes n'ont cependant pour le moment pas montré leur intérêt. En termes de perspectives, nous envisageons de creuser la piste de l'extension de requête, en faisant varier dans un premier temps le nombre de termes ajoutés par requête et leur pondération. Nous étudierons également d'autres sources pour l'extension des tweets et des requêtes, telles que Wikipedia.

7. Bibliographie

Cohen D., Amitay E., Carmel D., « Lucene and Juru at TREC 2007 : 1-Million Queries Track. », *TREC'07*, p. -1–1, 2007.

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

Duan Y., Jiang L., Qin T., Zhou M., Shum H.-Y., « An empirical study on learning to rank of tweets », *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, p. 295–303, 2010.

Furnas G. W., Deerwester S., Dumais S. T., Landauer T. K., Harshman R. A., Streeter L. A., Lochbaum K. E., « Information retrieval using a singular value decomposition model of latent semantic structure », SIGIR '88, ACM, New York, NY, USA, p. 465–480, 1988.

Joachims T., « A support vector method for multivariate performance measures », ICML '05, ACM, New York, NY, USA, p. 377–384, 2005.

Metzler D., Cai C., « USC/ISI at TREC 2011 : Microblog Track (Notebook Version) », *TREC'11 : 20th Text Retrieval Conference*, NIST, November, 2011.

Metzler D., Croft W. B., « A Markov random field model for term dependencies », SIGIR '05, ACM, New York, NY, USA, p. 472–479, 2005.

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