Combining authors features and tweet's features for refine search's results in microblogs

Soumaya Cherichi [1], Rim Faiz [1]
[1]*LARODEC, IHEC Carthage University of Carthage Carthage Presidency, Tunisia*
Soumayacherichi@gmail.com, Rim.Faiz@ihec.rnu.tn

Keywords: Social Information Retrieval, Microblogs, Knowledge Discovery,Relevant Tweets, Measuring Impact

Abstract: The rapid growth of online data due to the internet and the widespread use of large databases have resulted in increased demand for knowledge discovery methodologies. However, if the development of information technology has made it possible to store larger data volumes at lower cost, quantity and quality of information provided to users has in turn, changed little. Companies more than ever need to transform data into valuable knowledge directly. The colossal size of data and user demand poses a challenge to the scientific community to be able to offer effective tools for Information Retrieval and Knowledge Discovery. Several works have proposed criteria for tweets search, but, this area is still not well exploited, consequently, search results are irrelevant. In this paper, we propose new features such as audience and RetweetRank. We investigate the impact of these criteria on the search's results for relevant information. Finally, we propose a new metric to improve the results of the searches in microblogs. More accurately, we propose a research model that combines content relevance, tweet relevance and author relevance. Each type of relevance is characterized by a set of criteria such as audience to assess the relevance of the author,Out Of Vobulary to measure the relevance of content and others. To evaluate our model, we built a knowledge management system. We used a corpus of subjective tweets talking about Tunisian actualities in 2012.

1 INTRODUCTION

Currently the World-Wide Web has a huge amount of data and is obviously a reliable source of information for many topics. Businesses, organizations, individuals and the media, each product today and uses masses phenomenal information of all kinds. The amount of information stored in the world doubles every twenty months, this information is a considerable wealth. The question that arises is how to find this information and transform data collections into new knowledge, understandable, useful and interesting in the context where it is located. Information retrieval systems solve one of the biggest problems of Knowledge discovery: quickly finding useful information within massive data stores and ranking the results by relevance. Recent years have revealed the accession of interactive media, which gave birth to a huge volume of data in blogs and micro-blogs more precisely. These micro-blogs attract more and more users due to the ease and the speed of information sharing especially in real time. Indeed a micro-blog is a stream of text that is written by an author. It is composed by regular and short updates that are presented to readers in reverse chronological order called time-line. Today, the service called Twitter is the most popular micro-blogging platform. While micro-blogging services are becoming more famous, the methods for organizing and providing access to data are also improving. Micro-bloggers as well as sending tweets are looking for the last updates according to their interests. Finding the most relevant tweets to a topic depends on the criteria of micro-blogs. Unlike other micro-blogging service, Twitter is positioned by the social relationship of subscription. Once, association led, it allows users to express their interest in the items of another micro-bloggers. The social network of Twitter is not limited to bloggers and subscription relationships; it also includes all the actors and data that interact in both contexts of use and publication of articles. We have analyzed the micro-blogging service Twitter and we have identified the main criteria of Twitter. But the question arises what is the impact of each feature on the quality of results? Our work consists in searching a new metric of features impact on the search results quality. Several criteria have been proposed in the literature (Ben Jabeur et al., 2012),(Damak et al., 2011), but there are still other criteria that have not

been exploited as audience which could be the size of the potential audience for a message: What is the maximum number of people who could have been exposed to a message? We have gathered the features on three groups: those related to content, those related to tweet and those related to the author. We have used the coefficient of correlation with human judgment to define our final score(Akermi and Faiz, 2012). Our experimental result uses a corpus of thousand subjective tweets which are neither answers nor retweets, and we also collected a corpus of human judgments to find the correlation coefficient. The remainder of this paper is organized as follows. In section 2, we give an overview of difference between Knowledge Discovery and Information Retrieval. In section 3, we describe the task Twitter Information Retrieval. In Section 4, we present all the features that we have used to calculate our score. In section 5, we propose our new metric measure, then in section 6 we discuss experiments and obtained results. Finally, section 7 concludes this paper and outlines future work.

2 KNOWLEDGE DISCOVERY VS INFORMATION RETRIEVAL

Knowledge discovery and information retrieval pursue the same aim, which is to find information in a data set), but differ in their responses and the means implemented. Their fundamental difference is the nature of the information they return. Information Retrieval models independently seek information and data from a collection of documents. Then selects the information that address a given topic (subject expressed by a query). Such a system is open (requests are not fixed a priori). Knowledge discovery conduct an analysis of raw documents in order to extract only specific information that would interest the user, this information is specified a priori (Montaner et al., 2003). Despite their differences, these techniques reveal complementary. The combination of knowledge discovery and information retrieval has indeed a great potential in the creation or improvement of applications knowledge extraction from data. There are several ways to combine these two systems:

- Using Information Retrieval in pretreatment of Knowledge discovery: facing a very large volume of data, it can provide to a Knowledge discovery system one sub collection that do involving only the most relevant documents.;
- UsingKnowledge discovery to refine the results of an Information Retrieval system by improving the modeling phase documents: the information discovered from each document via a form through a Knowledge discovery process can be used to create an index that models the document.;
- Techniques specific to Knowledge discovery can also be used to complement traditional approaches to information retrieval for categorizing, filtering and ordering documents according to their relevance.;

3 MICROBLOG INFORMATION RETRIEVAL

A micro-blogging service is at once a communication mean and a collaboration system that allows sharing and disseminating text messages. In comparison with other social networks on the Web (for example Facebook , Myspace, linkedIn, FourSquare), the microblog's articles are particularly short and submitted in real time to report a recent event. At the time of this writing, several micro-blogging services exist. In this paper, we will focus on the micro-blogging service Twitter which is the most popular and widely used. Twitter is characterized from similar sites by certain features and functionalities. An important characteristic is the presence of social relationships subscription. This directional relationship allows users to express their interest on the publications of a particular blogger. Twitter is distinguished from similar websites by some key features (Ben Jabeur et al., 2012). The main one consists on the following social relationship. This directed association enables users to express their interest in other micro-bloggers posts, called tweets, which doesn't exceed 140 characters. Moreover, Twitter is marked by the retweet feature which gives users the ability to forward an interesting tweet to their followers. A blogger, also called twitterer, can annotate his tweets using ♯ hashtags or send it to a specific user through the user @ mentions. Finally, a tweet can also share a Web resource referenced by a URL.

Twitter employs a social-networking model called *"following"*, in which each *"twitterer"* is allowed to choose who he wants to follow without seeking any permission. Conversely, he may also be followed by others without granting permission. In one instance of "following" relationship, the twitterer whose updates are being followed is called the *"friend"*, while the one who is following is called the *follower* (Weng et al., 2010).

Given the specificity of micro-blogs, looking for tweets is facing several challenges such as indexing the flow of items (Sankaranarayanan et al., 2009),

spam detections (Yardi et al., 2010), diversification of results (Choudhury et al., 2011), and evaluating the quality of tweets (Nagmoti et al., 2010; Pal and Counts, 2011). We find that most approaches for information retrieval in micro-blogs don't take into account all the features to narrow the search. In fact, each feature has a unique impact on the other ones. Based on this observation and to improve the results of research, we will try to overcome these limitations by measuring the impact of these criteria. We will propose a measurement Metric Impact Criteria for Improving Outcomes Research

The search for tweets is an information retrieval task ad-hoc whose objective is to select the items relevant micro-blogs in response to a query(Santos et al., 2010). The definition of relevance in the search for tweets is not limited to textual similarity but also takes account of social interactions in the network. In this context, the relevance of the items depends also on the tweets' technical specificities and the importance of the author. Compared to Web search, the search for tweets provides brief, concise and accurate information on a current topic (Java et al., 2007) (for example Obama Visits Aurora Shooting Victims, Families ♯AuroraShooting, in this example the current topic is Aurora Shooting). It can also receive real-time information about an event that just happened a few seconds ago. Finally, the search for tweets provides access to news with a diversity of views of bloggers (Ben Jabeur et al., 2011).In this work, we address first the issues of integrating criteria of relevance and importance of measuring tweets' features and those of the author.

Regarding the relevance of content, several studies have used Okapi BM25 algorithm (Robertson et al., 1999), other studies like work of *Duan et Al* (Duan et al., 2010), have added new features such as tweets' quality ie the tweet that contains the least amount of Out of vocabulary (OOV) is considered as the most informative one. Also *Duan et Al* (Duan et al., 2010), consider that the longer the tweet, the better amount of information it contains... The Merriam-Webster dictionary defines influence "as the power or capacity of causing an effect in indirect or intangible ways". Despite the large number of theories of influence in sociology, there is no tangible way to measure such a force nor is there a concrete definition of what influence means, for instance, in the spread of news(Cha et al., 2010). With the aim to measure the importance or influence of a blogger, many studies have suggested to study the quality of bloggers as a first step to estimate the relevance of their articles. We note that an article published by a major blogger is more relevant in this context than another article written by an "unknown". With the aim to measure the importance of a blogger, *Balog et Al*(Balog et al., 2008) propose to assess its expertise about the application and that based on a language model. Other approaches (Zhang et al., 2007) and (Noll et al., 2009)consider that the domain experts are connected by social relations and propose to explore the topology of the social network to identify them.

Beyond the expertise, *Weng et Al*(Weng et al., 2010) and *Nagmoti et Al* (Nagmoti et al., 2010) propose to measure the influence of bloggers by applying the algorithm-PageRank *TwitterRank* sensitive topic on the subscription network. For the integration of relevant factors, some modular approaches propose to calculate multiple factors of relevance based on the network structure and combine them later. These approaches are studying the social significance by applying the measures of social network analysis and estimate a global relevance by combining the thematic relevance and social importance. Other integrated approaches, model all factors of relevance to the network and view the global relevance by a transition probability (Yahia et al., ; Yang et al., 2010).

We introduce in this paper our approach for tweet search that integrates different criteria namely the social authority of micro-bloggers, the content relevance , the tweeting features as well as the hashtag's presence. We present in the next section the main features of our criteria.

4 FEATURES FOR TWEETS RANKING

Among the most important tasks for a ranking system tweet is the selection of features set. We offer three types of features to rank tweets:

1) Content features refer to those features which describe the content relevance between queries and tweets.

2) Tweet features refer to those features which represent the particular characteristics of tweets, as OOV and hashtags in tweet.

3) Author features refer to those features which represent the authority of authors of the tweets in Twitter

4.1 Features Set

4.1.1 Content Relevance Features

We used four content relevance features:

-Relavance(T,Q): we used OKAPI BM25 score measures the content relevance between query Q and

tweet T.

$$TF-IDF_{(}w,Ti) = TF_{(}w,Ti).IDF_{(}w,Ti)$$
$$= TF_{w,Ti}((\log_2 * \frac{N}{DF_w}) + 1)$$

knowing that: w is a term in the query Q an Ti is the tweet i.

-Popularity(Ti,Tj,Q) with $i \to n$ and $j \neq i$: it used to calculate the popularity of a tweet from the corpus. It measures the similarity between the tweets in the context of the tweet's theme. We have used cosine similarity, according to a study done by *Sarwar et Al* (Sarwar et al., 2001) cosine similarity is the most efficient similarity measure in addition it is not sensitive to the size of each tweet:

$$Cosine(Ti,Tj) = \frac{\sum_{w\in(Ti\bigcap Tj)} TFIDF_{w,Ti} * TFIDF_{w,Tj}}{\sqrt{\sum_{w\in Ti}(TFIDF_{w,Ti})^2 * \sum_{w\in Tj}(TFIDF_{w,Tj})^2}}$$

knowing that:

w : w is a term in the query Q, Ti : tweet i, Tj : tweet j and $i \to n$ and $j \neq i$

-Length of tweet (Lg(Ti,Q)):Length is measured by the number of characters that a tweet contains. It is said that more the tweet is long, more it contains information(Duan et al., 2010).

$$Lg(Ti,Q) = \frac{Lg(Ti) - MinLg(T)}{MaxLg(T)}$$

-Out of Vocabulary (OOV(Ti)):This feature is used to roughly approximate the language quality of tweets. Words out of vocabulary in Twitter include spelling errors and named entities.This feature aims to measure the quality language of tweet as follows

$$Quality(T) = 1 - \frac{Numberof OOV(Ti)}{Lg(Ti)}$$

with Number of OOV(Ti)is calculated as follows

```
String tweet[] = tweet.split(" ");
int count = 0;
for (int i = 1; i < tweet.length; i++)
if (checker.isNotCorrect(tweet[i]))
{
Number of oov ++;
}
```

The more number of out of vocabulary is small the more quality of tweet is better.

4.1.2 Tweet Relevance Features

Each tweet has many technical features, and each feature form a selection criteria that we have exploited.

-Retweet(Ti,Q):A retweeted tweet usually includes an RT tag. Generally, sentences before RT are comments of the retweeter and sentences after RT are the original content, perhaps with some modifications. Here we only consider tweets including RT with the original content unmodified. Retweet (Ti,Q) is defined as the number of times a tweet is retweeted. In a rational manner, the most retweeted tweets are most relevant. Retweets are forwardings of corresponding original tweets, sometimes with comments of retweeters. They are supposed to contain no more information than the original tweets.(Duan et al., 2010)

$$Retweet(Ti,Q) = \frac{Retweet(Ti) - MinRetweet(T)}{MaxRetweet(T)}$$

-Reply(Ti):An @reply is any update posted by clicking the "Reply" button on a Tweet, it will always begin with @username. This feature aims to calculate the number of reply to a tweet.Likewise, tweets that have received the most response are more relevant

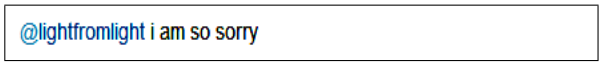

Figure 1: Tweet contains reply

$$Reply(Ti,Q) = \frac{Reply(Ti) - MinReply(T)}{MaxReply(T)}$$

-Favor(Ti):this feature aims to calculate the number of times a tweet is classified as a favorite. If a message is considered by many followers as a favorite, it means that it is relevant.

$$Favor(Ti,Q) = \frac{Favor(Ti) - MinFavor(T)}{MaxFavor(T)}$$

-Hashtag Count(Ti):The ♯ symbol, called a hashtag, is used to mark keywords or topics in a Tweet. It was created organically by Twitter users as a way to categorize messages. This feature aims to calculate the number of hashtags in tweet

$HashtagCount(Ti) = \sum$ of occurrences of hashtag

-Url count(Ti):Twitter allows users to include URL

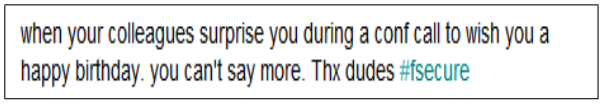

Figure 2: Tweet contains Hashtag ♯

as a supplement in their tweets.This feature aims to estimates the number of times that the URL appears in the tweet corpus. According to (Damak et al., 2012),tweets containing urls are more informative

$URLCount(Ti) = \sum$ of occurrences of URL

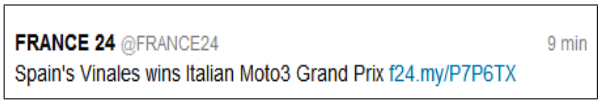

Figure 3: Tweet contains URL

4.1.3 Author Relevance Features

Each blogger has specific characteristics such as number of follower number of mention We said that users who have more followers and have been mentioned in more tweets, listed in more lists and retweeted by more important users are thought to be more authoritative. Apart these features we have added others such as hearing, TwitterPageRank, Expertise ...

-Tweet Count(a):this feature represents the number of tweet posted by the author

-Mention Count(a): A mention is any Twitter update that contains "@username" anywhere in the body of the Tweet , this means that @replies are also considered mentions. This feature aims to calculate the number of times an author is mentioned. **-**

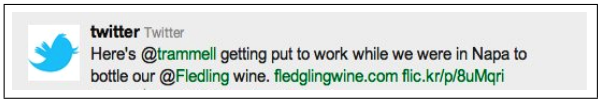

Figure 4: Tweet contains Mention @

Follower(a):this feature represents the number of follower to the author

-Following(a): this feature represents the number of subscriptions of the author (a) to other authors

-Expertise(a): this feature was found by conducting a survey that asks people to note the expertise of the blogger from 0 to 10

-RetweetRank(a):Retweet Rank looks up all recent retweets, number of followers, friends and lists of a user. It then compares these numbers with those of other users' and assigns a rank. Retweet Rank tracks both RTs posted using the Retweet button and other RTs (e.g. RT @username).This feature is an indicator of how a blogger is influential on twitter.

-TwitterPageRank(a): this feature represents the rank of author of the total twitter users using PageRank Algorithm

-Audience(a): is the size of the potential audience for a message. What is the maximum number of people who could have been exposed to a message?

5 METRIC MEASURE OF THE IMPACT OF CRITERIA TO IMPROVE SEARCH RESULTS

We introduce a research model that combines tweets relevant content, the specificities of tweets and the authority of bloggers. This model considers the specificities of tweets and the authority of bloggers as important factors which contribute to the relevance of the results. The search for tweets is a task of information retrieval whose goal is to select the relevant sections in response to a user's request. To present an accurate list of articles, our model combines a score of content's relevance, a score of author's authority and a score of tweets' specificities. The objective of this combination is to provide a list of tweets that cover the subject of the request and are posted by major bloggers. After normalizing the feature scores, these three scores are combined linearly using the following formula:

$$\begin{aligned} Score(Ti,Q) = {} & scoreContent(Ti,Q) \\ & +\beta * scoreTweet(Ti,Q) \\ & +\gamma * scoreAuthor(Ti,Q) \end{aligned} \quad (1)$$

with score(Ti,Q) on [0, 2] and $\beta+\gamma$=1.

where Ti and q represent respectively, tweet and request. $\beta, and \gamma$ on[0,1] are a weighting parameter(Akermi and Faiz, 2012). Scorecontent (Ti, Q) is the normalized score of the relevance of content. Scoretweet is the normalized score of the specificity of the tweet Ti and ScoreAuthor (A, Ti) is the normalized score of the importance of the author A corresponds to the blogger who published the tweet Ti.

We note that:

$$\begin{aligned} Scorecontent(Ti,Q) = {} & Relevance(T,Q) + Lg(Ti) \\ & +Popularity(Ti,Tj,Q) \\ & +Quality(Ti); \end{aligned} \quad (2)$$

$$\begin{aligned} ScoreTweet(Ti,Q) = {} & Urlcount(Ti) + HashtagCount(Ti) \\ & +Retweet(Ti) + Reply(Ti) \\ & +Favor(Ti); \end{aligned} \quad (3)$$

$$\begin{aligned} ScoreAuthor(A,Q) = {} & TwitterPageRank(a) + Audience(a) \\ & +TweetCount(a) + MentionCount(a) \\ & +Expertise(a) + RetweetRank(a) \\ & +Follower(a) + Following(a); \end{aligned} \quad (4)$$

6 EXPERIMENTAL EVALUATION

We conducted a series of preliminary experiments on a collection of articles from Twitter, in order to evaluate the performance of our model.

6.1 Description of the collection

With the absence of a standard framework for evaluating information retrieval in micro-blogs, we collected a set of articles and queries. We describe in the following collection of articles and the approach for collecting relevance judgments.

6.1.1 Search Engine TWEETRIM

We construct a search engine that we have called "TWEETRIM", which allows to calculate all scores and display the most relevant tweets according to these score. It has as input a query composed of three keywords and as output a set of relevant tweets relative to the query.

6.1.2 Tweets set

We built a collection of articles, metadata about relationships subscription and reply. This corpus is collected manually ie a thousand blogs and thousands of tweets have been browsed. This collection includes a total of 1000 articles published by 50 active Tunisian bloggers who are interested on the Tunisian news, we chose the period of March 4, 2012 until June 4, 2012.

6.1.3 Queries and relevance judgments

To construct queries and the collection of human judgments of relevance we followed the following steps:

1) we collected 300 queries on recent actualities in Tunisia from users,

2) then, we used the system that we have built which allows us to view the 10 results are especially relevant according to the score of the content,

3) and then, we asked 300 users to judge the 10 first results of each query.

We suppose that the content relevance already exists and we will improve our search result by varying our two other scores ScoreTweet and ScoreAuthor. We calculate the correlation coefficient between our scores and the corpus, which allowed us to find our weighting coefficients β and γ.

6.2 Results

6.2.1 Comparing the relevancy factors

In this experiment, we evaluate the factors relevant to know the specifics of tweets and blogger's authority then we compare their performance independently. Figure 5 shows the values of correlation coefficients obtained by the different configurations of our metric measure. We emphasize that the content relevance already exists and we will improve our search result by varying our two other scores ScoreTweet and ScoreAuthor.

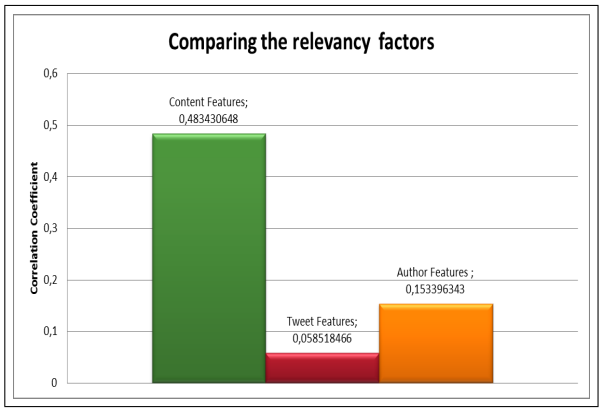

Figure 5: Comparing the relevancy factors

6.2.2 Estimation of weights

We compare, in Figure 6, the values of correlation coefficients and from these results, we observe that the best correlation coefficient between β ScoreTweet+γScoreAuthor with human jugement score = 0,161456763 when β = 0,4 and thus γ = 0,6.

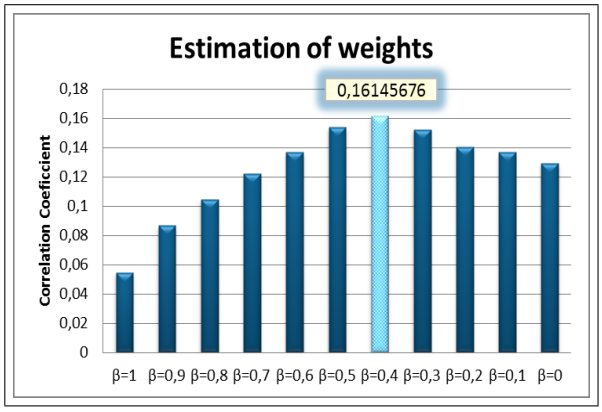

Figure 6: Estimation of weights

6.2.3 Evaluation of our model

We compare, in Figure 7, the values of correlation's coefficients obtained by Tweet Features and Author Features with the parameters β, γ values respectively (1.0) and (0.1) obtained by experiments and the third configuration with β=0.4 and γ=0.6. we notice that the performance of the last 2 configurations are very close with a slight advantage for the combination "Tweet Features & Author Features " on the model based only on the specificities of the tweet and the

importance of the author. We conclude that Author features has more impact on the search's results then Tweet features. However, we have not recorded in

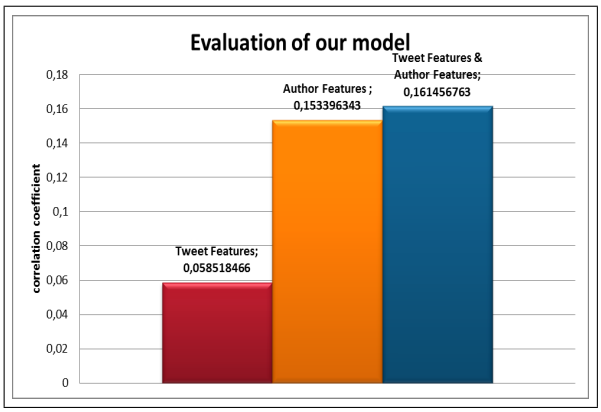

Figure 7: Comparing correlation coefficients

these experiments an improvement over the reference model. This amounts to the small size of the data collection and the small number of relevant documents. We also explain these results by the thematic nature of human judgments of relevance based on the documents ranked according to the first score of the content and not on social interactions.

7 CONCLUSIONS

Research conducted under the auspices of knowledge management varies greatly in direction and scope. There are several approaches that have been proposed which are based on the features. Therefor the choice of characteristics is important to obtain a satisfactory result and close to the human judgment. We have proposed in this paper a new metric for Social Research on twitter. This has to integrate relevance of content, the specificities of tweets and the author's importance where we incorporate new features such as the hearing. The preliminary experimental evaluation we conducted on a collection of articles of Twitter shows the measurement that we propose allows a better assessing the impact of bloggers and tweets' technical specificities.

Looking ahead, we plan to conduct experiments under the Micro-blog TREC evaluation framework that will include a collection of articles and queries for larger and whose relevance judgments are social. We also need to evaluate the influence of each feature independently. We plan to compare the performance of our model with other models for social searching for tweets.

REFERENCES

Akermi, I. and Faiz, R. (2012.). Hybrid method for computing word-pair similarity based on web content. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics, WIMS'12, New York, NY, USA*, ACM.

Balog, K., de Rijke, M., and Weerkamp, W. (2008). Bloggers as experts: feed distillation using expert retrieval models. In *SIGIR*, pages 753–754.

Ben Jabeur, L., Tamine, L., and Boughanem, M. (2011). Un modéle de recherche d'information sociale dans les microblogs : cas de Twitter (regular paper). In *Conférence sur les Modéles et l'Analyse des Réseaux : Approches Mathématiques et Informatique (MARAMI), Grenoble*.

Ben Jabeur, L., Tamine, L., and Boughanem, M. (2012). Uprising microblogs: A Bayesian network retrieval model for tweet search. In *ACM Symposium on Applied Computing (SAC), Riva del Garda (Trento), Italy*.

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P. (2010). Measuring user influence in twitter: The million follower fallacy. In *in ICWSM 10: Proceedings of international AAAI Conference on Weblogs and Social*.

Choudhury, M. D., Counts, S., and Czerwinski, M. (2011). Find me the right content! diversity-based sampling of social media spaces for topic-centric search. In Adamic, L. A., Baeza-Yates, R. A., and Counts, S., editors, *ICWSM*. The AAAI Press.

Damak, F., Ben Jabeur, L., Cabanac, G., Pinel-Sauvagnat, K., Tamine, L., and Boughanem, M. (2011). IRIT at TREC Microblog 2011 (regular paper). In Ellen M., V. and Lori P., B., editors, *Text REtrieval Conference (TREC), Gaithersburg, USA*.

Damak, F., Pinel-Sauvagnat, K., and Cabanac, G. (2012). Recherche de microblogs : quels critàres pour raffiner les résultats des moteurs usuels de RI ? In *Conférence francophone en Recherche d'Information et Applications (CORIA), Bordeaux, France*.

Duan, Y., Jiang, L., Qin, T., Zhou, M., and Shum, H.-Y. (2010). An empirical study on learning to rank of tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 295–303, Stroudsburg, PA, USA. Association for Computational Linguistics.

Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA. ACM.

Montaner, M., López, B., and De La Rosa, J. L. (2003). A taxonomy of recommender agents on theinternet. *Artif. Intell. Rev.*, 19(4):285–330.

Nagmoti, R., Teredesai, A., and Cock, M. D. (2010). Ranking approaches for microblog search. In Huang, J. X., King, I., Raghavan, V. V., and Rueger, S., editors, *Web Intelligence*, pages 153–157. IEEE.

Noll, M. G., Au Yeung, C.-m., Gibbins, N., Meinel, C., and Shadbolt, N. (2009). Telling experts from spammers: expertise ranking in folksonomies. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 612–619, New York, NY, USA. ACM.

Pal, A. and Counts, S. (2011). Identifying topical authorities in microblogs. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 45–54, New York, NY, USA. ACM.

Robertson, S., Walker, S., Beaulieu, M., and Willett, P. (1999). Okapi at trec-7: Automatic ad hoc, filtering, vlc and interactive track. *In*, 21:253–264.

Sankaranarayanan, J., Samet, H., Teitler, B., Lieberman, M., and Sperling, J. (2009). Twitterstand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 42–51. ACM.

Santos, R., McCreadie, R., Macdonald, C., and Ounis, I. (2010). University of glasgow at trec 2010: Experiments with terrier in blog and web tracks. In *Proceedings of TREC 2010*.

Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, WWW '01, pages 285–295, New York, NY, USA. ACM.

Weng, J., Lim, E.-P., Jiang, J., and He, Q. (2010). Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 261–270, New York, NY, USA. ACM.

Yahia, S. A., Benedikt, M., and Bohannon, P. Challenges in searching online communities. *IEEE Data Eng. Bull*, 30:2007.

Yang, Z., Hong, L., and Davison, B. D. (2010). Topic-driven multi-type citation network analysis. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*, RIAO '10, pages 24–31, Paris, France, France.

Yardi, S., Romero, D. M., Schoenebeck, G., and Boyd, D. (2010). Detecting spam in a twitter network. *First Monday*, pages –1–1.

Zhang, J., Tang, J., and Li, J.-Z. (2007). Expert finding in a social network. In *DASFAA*, pages 1066–1069.