Relevant Information Management in Microblogs

Soumaya Cherichi and Rim Faiz

LARODEC, IHEC Carthage

University of Carthage

Carthage Presidency, Tunisia

soumayacherichi@gmail.com

Rim.Faiz@ihec.rnu.tn

Soumaya Cherichi obtained her Master in Computer science from the University of Carthage, in Tunisia. She is currently a Ph.D student at the Higher Institute of Management (ISG), University of Tunis, in Tunisia. Her research interests Natural Language Processing, Information Retrieval, and Semantic Web. She has published papers in different conference like KMIKS, WIMS, and KDIR.

Rim Faiz obtained her Ph.D. in Computer Science from the University of Paris- Dauphine, in France. She is currently a Professor in Computer Science at the Institute of High Business Study (IHEC), University of Carthage, in Tunisia. Her research interests include Artificial Intelligence, Machine Learning, Natural Language Processing, Information Retrieval, Text Mining, and Semantic Web. She has published several papers and has served as PC member and reviewer for several international conferences and journals. Dr. Faiz is also responsible of the Master "ECommerce" and Master "Business Intelligence" at IHEC, University of Carthage.

Chapter 6

Relevant Information Management in Microblogs

Modern information retrieval (IR) has come to terms with numerous new media in efforts to help people find information in increasingly diverse settings. Among these new media are so-called microblogs. Twitter, the most popular microblogging service till recently, has known an exponential growth. 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 management (KM). Several works have proposed tools for tweets search, but, this area is still not well exploited. This gave us the idea to offer new tool more developed that uses new features such as audience and RetweetRank for ranking relevant tweets. 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, OOV (Out Of Vobulary) to measure the relevance of content and others. To evaluate our model, we built a knowledge management system. We used a collection of subjective tweets talking about Tunisian actualities in 2012.

Keywords: Twitter, Micro-blog, information retrieval,

knowledge management, relevant tweets, measuring impact,

social relation, features set,

1. Introduction:

Business organizations and the media product and use today masses of phenomenal information of all kinds. The problem that we face 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 management (KM): 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.

Peoples' motivations for microblogging are diverse (Jansen, M. Zhang, Sobel, & Chowdury, 2009; Java, Song, Finin, & Tseng, 2007; Krishnamurthy, Gill, & Arlitt, 2008; Zhao & Rosson, 2009), ranging from offering autobiographical updates to reporting on current news events (Diakopoulos & Shamma, 2010; Shamma, Kennedy, & Churchill, 2009) or crisis situations (Longueville, R. S. Smith, & Luraschi, 2009; Sakaki, Okazaki, & Matsuo, 2010; Vieweg, Hughes, Starbird, & Palen, 2010). Peoples' use of Twitter to communicate during political unrest (e.g. the disputed 2009 election in Iran) brought increased attention to microblogs as a communication medium (Evgeny, 2009; Grossman, 2009). Twitter has played a role in important events, but the service also allows people to communicate among a relatively small social circle, and a sizeable part of Twitter's success is because of this function.

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. And since the association is 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 contributors 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 11] and [CHA 10], 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 gathered the features on three groups: those related to content, those related to tweet and those related to the author. We used the coefficient of correlation with human judgment to define our score.

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 describe the task Twitter Information Retrieval. In section 3 we present all the features that we have used to calculate our score. In Section 4, we discuss experiments and obtained results. Finally, section 5 concludes this paper and outlines future work.

2. Twitter Information Retrieval:

Recent years have witnessed the advent of interactive media and especially Web 2.0. This led to a huge volume of data from blogs, discussion forums and commercial sites. Indeed, blogs, being the figurehead of Web 2.0, are characterized by their evaluative usage, in the sense that users are using them to express themselves freely and share their opinions on their interests.

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 microblogs 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 11]. 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 an URL.

Given the specificity of micro-blogs, looking for tweets is facing several challenges such as indexing the flow of items [SAN 09], spam detections, diversification of results and evaluating the quality of tweets [NAG 10]. 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. 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.

Regarding the relevance of content, several studies have used Okapi BM25 algorithm [ROB 98], other studies like work of Duan et al., [DUA 10], 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, 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”. Although there is a 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 10]. 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., [BAL 08] propose to assess its expertise about the application and that based on a language model. Language modeling IR relies on generative models to rank documents against queries. Under this approach, we assume that the terms in a document di were generated by a particular probability distribution Mi.

To avoid numeric underflow, we typically use log-probabilities, replacing the product in Eq. 2 with a summation of logarithms. In either case, in the language modeling approach, documents whose

corresponding language models have a high probability of generating q are ranked higher than those with low probability.

Other approaches [ZHA 07] and [NOL 09.] consider that the domain experts are connected by social relations and propose to explore the topology of the social network to identify them.

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.

3. 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

3.1. Features Set:

3.1.1. Content Relevance Features

The criterion "Content" refers to the thematic relevance traditionally calculated by IR systems standards. The thematic relevance is generally measured by one of several IR models. One of the models reference. Information Retrieval IR is the probabilistic model [Jon 00] with the weighting scheme BM25 as matching request document function. For this reason, we have adopted this model for the calculation of the thematic relevance. Of course, it is made possible to calculate using any other IR model. BM25 is a search function based "bag of words", it allows us to organize all documents based on the occurrences of the query terms given in the documents. (cf section 2)

We used four content relevance features:

1) Relevance(T,Q): we used OKAPI BM25 score which measures the content relevance between the query *Q* and tweet *T*.

$$\begin{aligned} TF-IDF_{(}w,Ti) &= TF_{(}w,Ti).IDF_{(}w,Ti) \\ &= TF_{w,Ti}((\log_2 * \frac{N}{DF_w}) + 1) \end{aligned} \qquad [1.1]$$

Knowing that: *w* is a term in the query *Q* and *Ti* is the tweet *i*.

2) Popularity(Ti,Tj,Q): with i and j in n and i≠j : 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 topic. We used cosine similarity, according to a study done by Sarwar et Al [SAR 01] cosine similarity is the most efficient similarity measure ,in addition, it is not sensitive to the size of each tweet:

$$Cosine(Ti,Tj) = \frac{\sum_{w \in (Ti \bigcap Tj)} TFIDF_{w,Ti} * TFIDF_{w,Tj}}{\sqrt{\sum_{w \in Ti} (TFIDF_{w,Ti})^2 * \sum_{w \in Tj} (TFIDF_{w,Tj})^2}} \qquad [1.2]$$

Knowing that *w* is a term in the query *Q*, *Ti* is tweet *i*, *Tj* is tweet j, *i and j in n and i≠j.*

3) 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.

$$Lg(Ti, Q) = \frac{Lg(Ti) - MinLg(T)}{MaxLg(T)} \quad [1.3]$$

4)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:

$$Quality(T) = 1 - \frac{NumberofOOV(Ti)}{Lg(Ti)} \quad [1.4]$$

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.

3.1.2 Tweet Relevance Features

We note that the thematic relevance depends solely on the item and query. Each tweet has many technical features, and each feature form selection criteria that we have exploited.

1) 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 forwarding of corresponding original tweets, sometimes with comments of retweeters. According to Duan et al. [DUA 10], they are supposed to contain no more information than the original tweets.

Figure 1: *Tweet retweeted*

$$Retweet(Ti, Q) = \frac{Retweet(Ti) - MinRetweet(T)}{MaxRetweet(T)} \quad [1.5]$$

2)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. Ultimately tweets that have received the most response are more relevant.

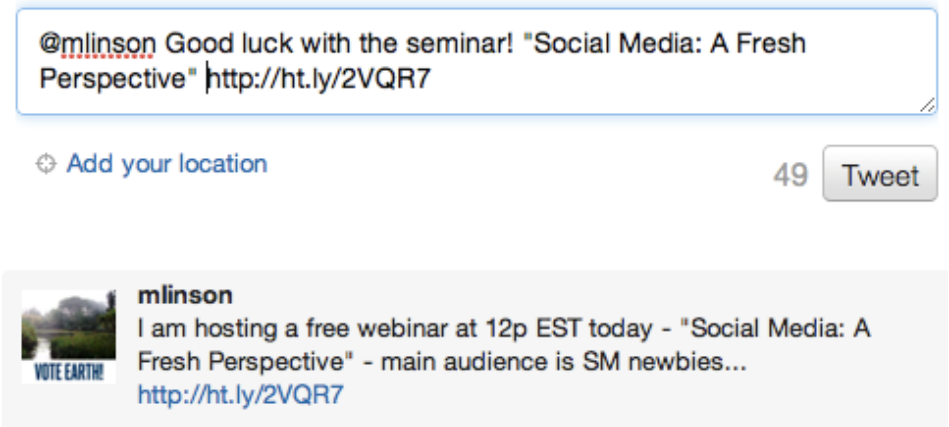

Figure 2: Tweet as reply

$$Reply(Ti, Q) = \frac{Reply(Ti) - MinReply(T)}{MaxReply(T)} \quad [1.6]$$

3) 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.

$$Favor(Ti, Q) = \frac{Favor(Ti) - MinFavor(T)}{MaxFavor(T)} \quad [1.7]$$

*4) 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

Figure 4:. *Tweet contains hasthag*

$$HashtagCount(Ti) = \sum \text{ of occurrences of hashtag} \qquad [1.8]$$

*5) Url count(Ti):*Twitter allows users to include URL 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 [DAM 12], tweets containing URLs are more informative

Figure 5: *Tweet contains URL*

$$URLCount(Ti) = \sum \text{ of occurrences of URL} \quad [1.9]$$

3.1.3. Author Relevance Features

Each blogger has specific characteristics such as number of follower and 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.

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

(2) *Mention Count (author):* 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.

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

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

(5) *Expertise(a):* this feature was found by conducting a survey that asks people to rate the expertise of the blogger from 0 to 10.

(6) *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* (ReTweets) *(e.g. RT @username*).This feature is an indicator of how a blogger is influential on twitter.

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

(8) *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?

4. 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.

As shown in Figure 6, we present the architecture of our approach to search for relevant tweets that consists of three steps:

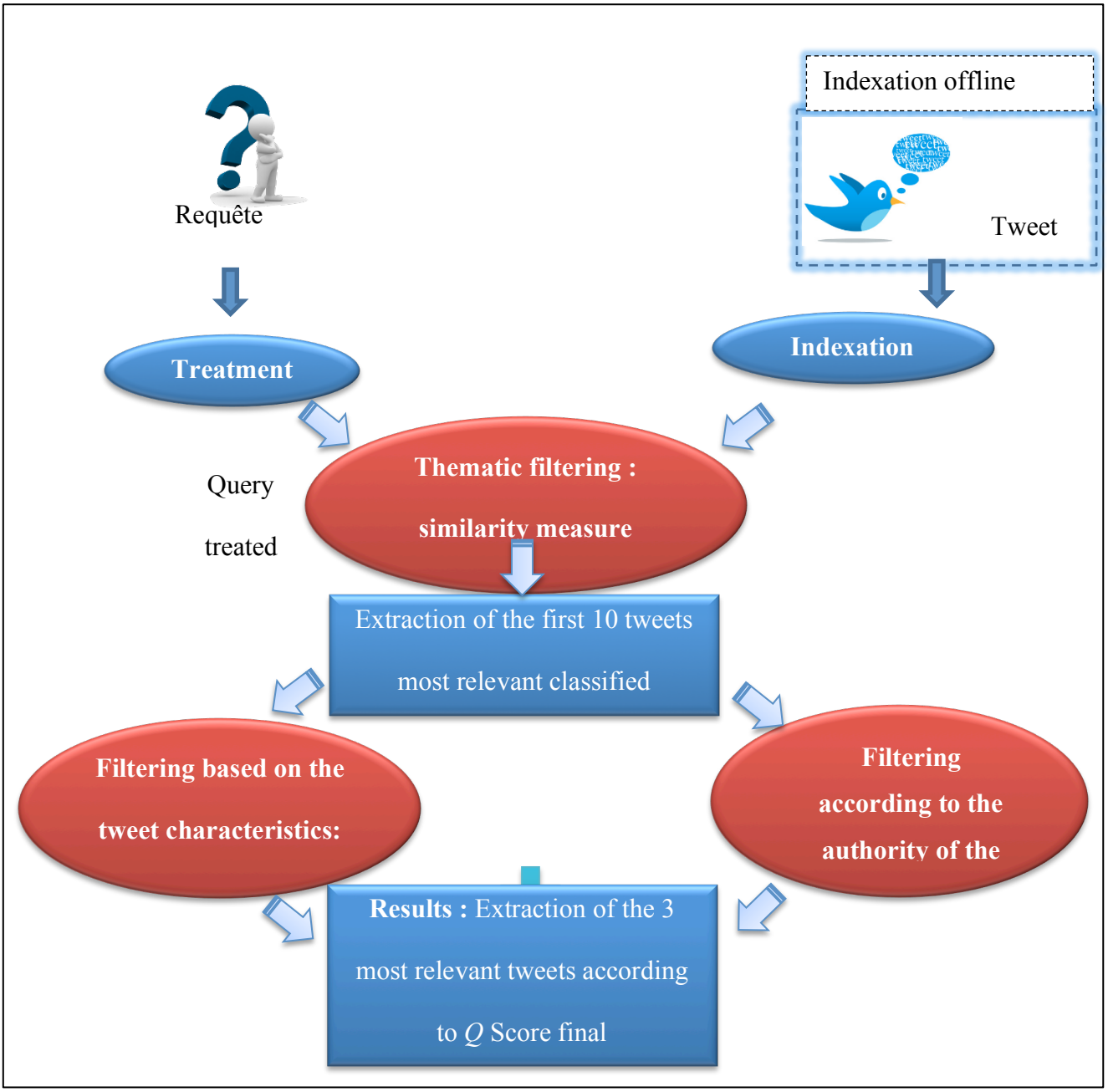

Figure 6: *General Architecture of our process*

A general algorithm called "extraction of the most relevant tweets" which describes the steps of our proposed approach:

Algorithm 3.1 : Extraction of the most relevant tweets

Input:

Q: User Query

C : Tweets' Collection

Begin

Pretreatment (*C ; Q*) : 1

Indexation(*C*) : *T* ← Indexed Tweet

Treatment(*Q*) : *Q* ← Query treatment

Thematic filtering: 2

1. Calculate the standard score of the content relevance

   For Each Indexed Tweet *Ti* in *C* :

   Calculate **Scorecontent (*Ti, Q*)**

   End

2. Sort Tweets by Scorecontent

   Vector C_{10} ← Extraction of the first 10 tweets most relevant classified for *Q*

3

Filtering based on the tweet characteristics:

3. Calculate the standard score of the specificity of tweet *Ti*

   For Each Tweet *T* in C_{10} :

   Calculate **Score Tweet (*Ti,Q*)**

   End

4. Sort Tweets by ScoreTweet

Filtering according to the authority of the author: 4

5. Calculate the standard score of the importance of the author who published *A* tweet *Ti*

   For each Tweet *T* in C_{10} :

   Calculate **Score Author (*A,Ti*)**

   End

6. Sort Tweets by **Score Author**

Filtring by final score : 5

7. Calculate final score :

   Score=Scorecontent + **Score Tweet** + **Score Author**

8. Sort tweets by their final scores

Vecteur C_3 ← Extraction of the first 3 tweets most relevant classified for Q

6

End Extraction of the 3 most relevant tweets according to Q Score final

Output : 3 most relevant tweets for Q

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:

$$\begin{aligned} Score(Ti,Q) = {} & scoreContent(Ti,Q) \\ & +\beta * scoreTweet(Ti,Q) \\ & +\gamma * scoreAuthor(Ti,Q) \end{aligned} \qquad [1.10]$$

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

Where Ti and Q represent respectively, tweet and request. β and γ on [0,1] are a weighting parameter[AKE 12]. 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:

(1) -Scorecontent(Ti,Q)=Relevance(T,Q) + Lg(Ti) + Popularity(Ti,Tj,Q) + Quality(Ti);

(2) -ScoreTweet(Ti,Q)= Url count(Ti)+ Hashtag Count(Ti) + Retweet(Ti) + Reply(Ti) + Favor(Ti);

(3) -ScoreAuthor(a,Q)= TwitterPageRank(a) +Audience(a)+ Tweet Count(a) + Mention Count(a) + Expertise(a) + RetweetRank(a) + Follower(a) + Following(a);

5. 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.

5.1. 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. Our concern is that the database size is small. We describe in the following collection of articles and the approach for collecting relevance judgments.

5.1.1. Search Engine TWEETRIM

We built 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.

5.1.2. Tweet 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 contained a total of 3000 tweets 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.

5.1.3. Queries and relevance judgments

To perform queries and to collect the human judgement of relevance followed the following steps:

(1) We collected 500 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.

5.2. Results

5.2.1. Comparing the relevancy factors

In this experiment, we evaluate the factors relevant to know the specifics of tweets and blogger's authorities then we compare their performance independently. Figure 7 shows the values of correlation coefficients obtained by the different configurations of our metric measure.

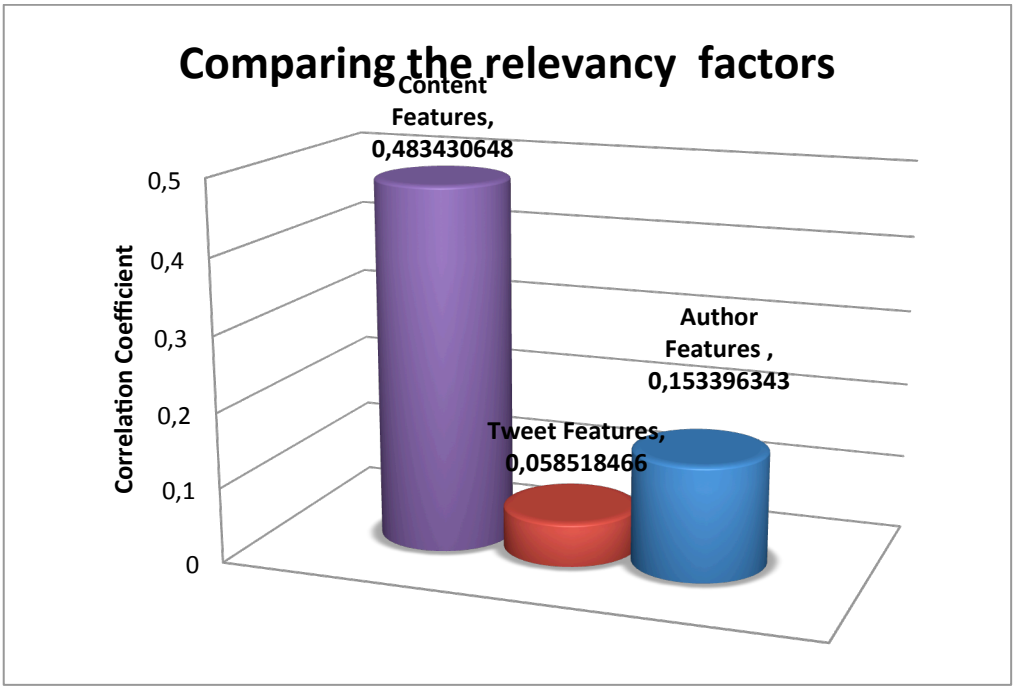

Figure 7: *Comparing the relevancy factors*

We emphasize that the content relevance already exists and we will improve our search result by varying our two other scores ScoreTweet and ScoreAuthor.

5.2.2. Estimation of weights

We make a comparison within the values the values of correlation coefficients and from these results, we observe that the best correlation coefficient between βScoreTweet+γScoreAuthor with human judgment score = 0,161456763 when β = 0,4 and thus γ= 0,6.

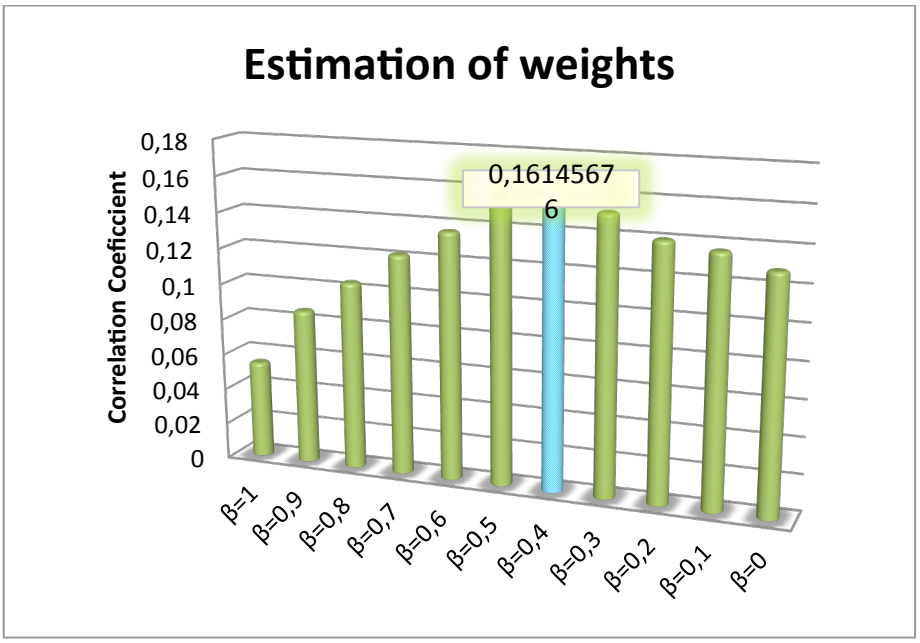

Figure 8: *Estimation of weights*

5.2.3

Evaluation of our model

We compare, in Figure 2, 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.

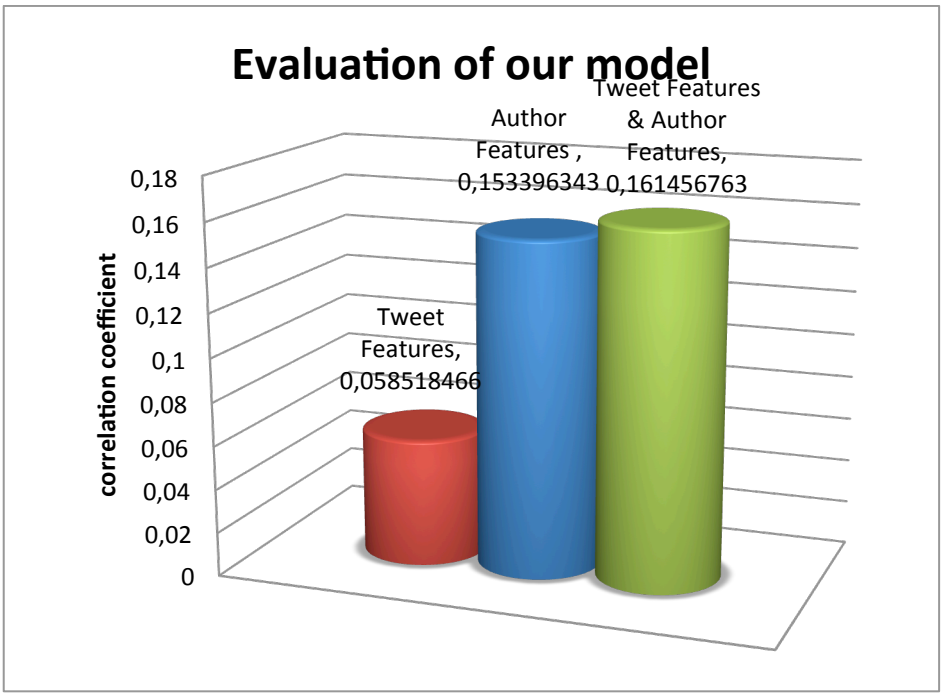

Figure.9: *Comparing correlation coefficients*

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 have more impact on the search's results then Tweet features.

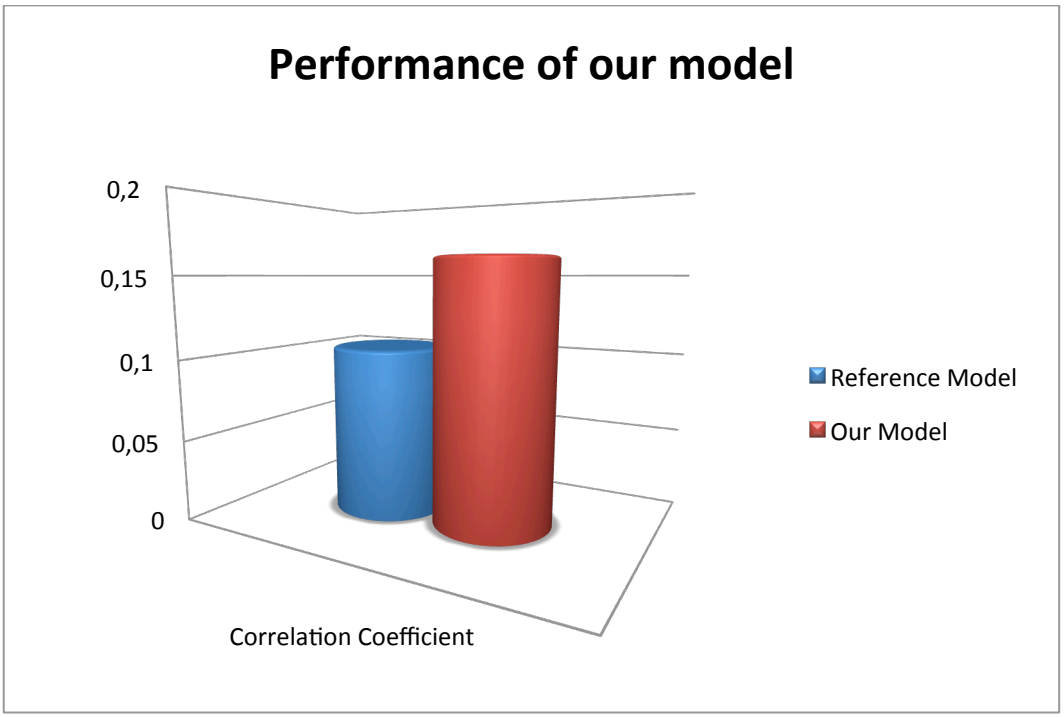

Figure.10. *Comparing our model with reference model*

The reference model combines only the features linearly without weighting. This model gave us the correlation coefficient equal to 0.10 and our model gave us the correlation coefficient of 0,16. Can clearly be seen a 35% improvement in the satisfaction of our human judgment.

6. Conclusion

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. Therefore

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 audience. The primary experimental evaluation that 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 Text REtrieval Conference (TREC) evaluation framework that will include a collection of many 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 of tweets.

References

[AGA 08] AGARWAL N., LIU H., TANG L., YU P. S. (2008), « Identifying the influential bloggers in a community », Proceedings of the international conference on Web search and web data mining, WSDM ’08, New York, NY, USA, 2008, ACM, p. 207-218.

[AKE 12] AKERMI I., AND FAIZ R., «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., 2012.

[AME 07] AMER-YAHIA S., BENEDIKT M., BOHANNON P., « Challenges in Searching Online Communities », IEEE Data Eng. Bull., vol. 30, no 2, , p. 23-31, 2007

[BAL 08] BALOG K., DE RIJKE M., WEERKAMP W., « Bloggers as experts : feed distillation using expert retrieval models », Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR ’08, New York, NY, USA, 2008, ACM, p. 753–754.

[BAR 11] BARRY S., “Web search: social & collaborative” in COnférence en Recherche d'Infomations et Applications CORIA 2011, 8th French Information Retrieval Conference CORIA : COnférence en Recherche d’Information et Applications Avignon, France, March 16-18, 2011. Proceedings CORIA, Editions Universitaires d'Avignon, 2011

[BAZ 05] Baziz M., Boughanem M., Aussenac-Gilles N ”Evaluating a Conceptual Indexing Method by Utilizing WordNet”, CLEF, 2005 p238-246 proceedings Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evalution

Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers, Springer, 2006

[BEN 12] BEN JABEUR L., TAMINE L., BOUGHANEM M., "Intégration des facteurs temps et autorité sociale dans un modèle bayésien de recherche de tweets",Conférence en Recherche d'Information et Applications (CORIA), Bordeaux, France, March 21-23, 2012. Proceedings, pp. 301–316, 2012.

[BEN 11] BEN JABEUR L., TAMINE L., BOUGHANEM M., « Un modèle de recherche d'information sociale dans les microblogs : cas de Twitter » Conférence sur les Modèles et l'Analyse des Réseaux : Approches Mathématiques et Informatique (MARAMI 2011), Grenoble, 2011

[BEN 10] BEN JABEUR L., TAMINE L., « Vers un modèle de Recherche d'Information Sociale pour l'accès aux ressources bibliographiques » (poster). Dans : Conférence francophone en Recherche d'Information et Applications (CORIA 2010), Sousse, Tunisie, 18/03/2010-20/03/2010, Centre de Publications Universitaires, p. 325-336, mars 2010.

[CHA 10] CHA M., HADDADI H., BENEVENUTO KRISHNA F., GUMMADI P., "Measuring User Influence in Twitter: The Million Follower Fallacy Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media 2010, ICWSM 2010

[CHE 12] CHERICHI S., FAIZ R., « Recherche d'information pertinente dans les microblogs: Mesure métrique de l'impact des critères pour améliorer les résultats de la recherche». Conférence Internationale sur l'Extraction et la Gestion des Connaissances – Maghreb, Hammamet, Tunisie, EGC-M 2012

[CHE 13] CHERICHI S. AND FAIZ R., New metric measure for the improvement of search results in microblogs. Proc. of the International Conference on Web Intelligence, Mining and Semantics (WIMS 2013), New York, NY, USA, 2013. ACM.

[CHE 13] CHERICHI S. AND FAIZ R., Relevant Information Discovery in Microblogs : New metric measure for the improvement of search results in microblogs. Proc. of INSTICC International Conference on Knowledge Discovery and Information Retrieval (KDIR 2013), Vilamoura, Portugal, 19-22 September 2013. ©SciTePress

[CHE 13] CHERICHI S., FAIZ R., "Relevant information management in microblogs". In International Conference on Knowledge Management, Information and Knowledge Systems (KMIKS 2013), Hammamet, Tunisia, Avril 2013.

[DAM 12] DAMAK F., PINEL-SAUVAGNAT K., CABANAC G., « Recherche de microblogs : quels critères pour raffiner les résultats des moteurs usuels de RI ? » COnférence en Recherche d'Information et Applications 2012 CORIA 2012 , Bordeaux, CORIA 2012 p. 371-328

[DIA 10] DIAKOPOULOS, N. A., & SHAMMA, D. A. (2010). Characterizing debate performance via aggregated twitter sentiment. In Proceedings of the 28th international conference on Human factors in computing systems (pp. 1195-1198). Atlanta, Georgia, USA: ACM. doi:10.1145/1753326.1753504

[DON 10] DONG A., ZHANG R., KOLARI P., BAI J., DIAZ F., CHANG Y., ZHENG Z., ZHA H., «Time is of the essence : improving recency ranking using Twitter data », Proceedings of the 19th international conference on World wide web, WWW '10, New York, NY, USA, 2010, ACM, p. 331–340.

[DUA 10] DUAN Y., JIANG L., QIN T., et al., "An empirical study on learning to rank of tweets", COLING Proceedings of the 23rd International Conference on Computational Linguistics Proceedings of the Conference, 23-27 August 2010, Beijing, China, pp. 295–303, 2010. Tsinghua University Press, 2010.

[EFR 10] EFRON M., « Hashtag retrieval in a microblogging environment », Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '10, New York, NY, USA, 2010, ACM, p. 787–788.

[EVG 09] EVGENY, M. (2009). Iran: Downside to the "Twitter Revolution". Dissent, 56(4), 10-14.

[GRO 09] GROSSMAN, L. (2009, June 17). Iran Protests: Twitter, the Medium of the Movement. Time.com. Retrieved August 24, 2010, from http://www.time.com/time/world/article/0,8599,1905125,00.html

[[HIE 01] HIEMSTRA D., « Using Language Models for Information Retrieval», PhD thesis, Enschede, January 2001.

[ING 09] INGWERSEN P., "Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction", Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 101–110, 1994 ACM/Springer 1994

[JAN 09] JANSEN, B. J., ZHANG, M., SOBEL, K., & CHOWDURY, A. (2009). Twitter power: Tweets as electronic word of mouth. J. Am. Soc. Inf. Sci. Technol., 60(11), 2169-2188.

[JAV 07] JAVA, A., SONG, X., FININ, T., & 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 (pp. 56-65). San Jose, California: ACM. doi:10.1145/1348549.1348556

[JAV 09] JAVA A., SONG X., FININ T., TSENG B., « Advances in Web Mining and Web Usage Analysis », Chapter Why We Twitter : An Analysis of a Microblogging Community, p. 118–138, Springer-Verlag, Berlin, Heidelberg, 2009.

[JON 00] JONES S., WALKER K., ROBERTSON S,. "A probabilistic model of information retrieval: Development and comparative experiments." Information Processing & Management, 36(6) :779–808, 2000.

[KIR 06] KIRSCH M., GNASA M., CREMERS A. B., « Beyond theWeb : Retrieval in Social Information Spaces », In Proceedings of the 28 th European Conference on Information Retrieval (ECIR 2006, Springer, 2006.

[KRI 08]KRISHNAMURTHY, B., GILL, P., & ARLITT, M. (2008). A few chirps about twitter. In Proceedings of the first workshop on Online social networks (pp. 19-24). Seattle, WA, USA: ACM. doi:10.1145/1397735.1397741

[KWA 10] KWAK H., LEE C., PARK H., MOON S., « What is Twitter, a social network or a news media ? », Proceedings of the 19th international conference on World wide web, WWW '10, New York, NY, USA, 2010, ACM, p. 591–600.

[LONG 09] LONGUEVILLE, B. D., SMITH, R. S., & LURASCHI, G. (2009). "OMG, from here, I can see the flames!": a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In Proceedings of the 2009 International Workshop on Location Based Social Networks (pp. 73-80). Seattle, Washington: ACM. doi:10.1145/1629890.1629907

[MAS 11] MASSOUDI K., TSAGKIAS E., DE RIJKE M., WEERKAMP W., ((Incorporating Query Expansion and Quality Indicators in Searching Microblog Posts)), ECIR 2011 : 33rd European Conference on Information Retrieval, Springer, Dublin, p. 362–367, 2011.

[MAR 60] MARON M. E., KUHNS J. L., "On relevance, probabilistic indexing and information retrieval". J. ACM, 7(3) :216–244, 1960.

[MUT 01] MUTSCHKE P., « Enhancing Information Retrieval in Federated Bibliographic Data Sources Using Author Network Based Stratagems », Reserach and Advanced Technology for Digital Libraries: 5th European Conference, ECDL 2001, Darmstadt, Germany, September 4-9, 2001.

[NAG 10] NAGMOTI R., TEREDESAI A., "Ranking Approaches for Microblog Search" Proceeding WI-IAT '10 Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01 Pages 153-157

[NOL 09] NOLL G., AU YEUNG C.-M., GIBBINS N., MEINEL C., SHADBOLT N., « Telling experts from spammers : expertise ranking in folksonomies », Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '09, New York, NY, USA, 2009, ACM, p. 612–619.

[OUN 10] OUNIS I., MACDONALD C., LIN J., SOBOROFF I., « TREC 2011 Microblog Track », Text REtrieval Conference TREC, November 2010.

[RAM 10] RAMAGE D., DUMAIS S. T., LIEBLING D. J., ((Characterizing Microblogs with Topic Models.)), ICWSM'10, p. -1–1, 2010.

[RAM 11] RAMAGE D., DUMAIS S.T., LIEBLING D.J., "Characterizing microblogs with topic models", ICWSM'10, pp. 1–1, 2010 Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010 The AAAI Press, 2010

[ROB 76] ROBERTSON S., SPARCK JONES K., "Relevance weighting for search terms". Journal of The American Society for Information Science, 27(3) :129–146, 1976.

[ROB 98] ROBERTSON S., WALKER S., HANCOCK-BEAULIEU M., « Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive », Text REtrieval Conference TREC, 1998, p. 199-210.

[Sak 10] SAKAKI T., OKAZAKI M., MATSUO Y., « Earthquake shakes Twitter users : realtime event detection by social sensors », Proceedings of the 19th international conference on World wide web, WWW '10, New York, NY, USA, 2010, ACM, p. 851–860.

[SAL 83] SALTON G., "A comparison between manual and automatic indexing methods. Journal of American Documentation", 20(1) :61–71, 1971.

[SAL 83] SALTON G., MCGILL M,. "Introduction to Modern Information Retrieval". McGraw- Hill, New York, 1983.

[SAN 09] SANKARANARAYANAN J., SAMET H., TEITLER B. E., LIEBERMANN M. D., SPERLING J., « TwitterStand : news in tweets », Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS '09, New York, NY, USA, 2009, ACM, p. 42–51.

[SAR 01] SARWAR B., KARYPIS G., KONSTAN J., AND RIEDL J., "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, 2001. ACM.

[SHA 09] SHAMMA, D. A., KENNEDY, L., & 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). Beijing, China: ACM.

[SON 10] SONG S., LI Q., ZHENG N., ((A spatio-temporal framework for related topic search in microblogging)), Proceedings of the 6th international conference on Active media technology, AMT'10, Springer-Verlag, Berlin, Heidelberg, p. 63–73, 2010.Teevan J.,

[VIE 10] VIEWEG, S., HUGHES, A. L., STARBIRD, K., & PALEN, L. (2010). Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In Proceedings of the 28th international conference on Human factors in computing systems (pp. 1079- 1088). Atlanta, Georgia, USA: ACM. doi:10.1145/1753326.1753486

[WEN 10] WENG J., LIM E.-P., JIANG J., HE Q., « TwitterRank : finding topic-sensitive influential twitterers », Proceedings of the third ACM international conference on Web search and data mining, WSDM '10, New York, NY, USA, 2010, ACM, p. 261–270.

[YAJ 10] YAJUAN DUAN, LONG JIANG, TAO QIN, MING ZHOU, HEUNG-YEUNG SHUM "An Empirical Study on Learning to Rank of Tweets" COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics Pages 295-303

[YAN 10] YANG Z., GUO J., CAI K., TANG J., LI J., ZHANG L., SU Z., « Understanding retweeting behaviors in social networks », Proceedings of the 19th ACM international conference on Information and knowledge management, CIKM '10, New York, NY, USA, 2010, ACM, p. 1633–1636.

[ZAI 10] ZAIHAN Y., L. H., DAVISON B. D., « Topic-driven Multi-type Citation Network Analysis », 9th International Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information (RIAO), Paris, France, 28/04/2010-30/04/2010, April 2010.

[ZHA 07] ZHANG J., TANG J., LI J., « Expert Finding in a Social Network », KOTAGIRI R., KRISHNA P., MOHANIA M., NANTAJEEWARAWAT E., Eds., Advances in Databases : Concepts, Systems and – Applications, vol. 4443 de Lecture Notes in Computer Science, p. 1066-1069, Springer Berlin / Heidelberg, 2007

[ZHA 09] ZHAO, D., & ROSSON, M. B. (2009). How and why people Twitter: the role that micro-blogging plays in informal communication at work. In Proceedings of the ACM 2009 international conference on Supporting group work (pp. 243-252). Sanibel Island, Florida, USA: ACM. doi:10.1145/1531674.1531710