Analyzing the behavior and text posted by users to extract knowledge

Soumaya Cherichi*, Rim Faiz**

* LARODEC, ISG Bardo
University of Tunis
Bardo, Tunisia

soumayacherichi@gmail.com

** LARODEC, IHEC Carthage
University of Carthage
Carthage Presidency, Tunisia

Rim.Faiz@ihec.rnu.tn

Abstract. Microblogging (e.g. Twitter1), as a new form of online communication in which users talk about their daily lives, publish opinions or share information by short posts, has become one of the most popular social networking services today, which makes it potentially a large information base attracting increasing attention of researchers in the field of knowledge discovery and data mining. Several works have proposed tools for tweets search, but, this area is still not well exploited. Our work consists of examining the role and impact of social networks, in particular microblogs, on public opinion. We aim to analyze the behavior and text posted by users to extract knowledge that reflect the interests and opinions of a population. 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: microblogs, relevant information, analyzing text posted, knowledge management system

adfa, p. 1, 2014.

1 Introduction

In the current era, People are becoming more communicative through expansion of services and multi-platform applications, i.e., the so called Web 2.0 which establishes social and collaborative backgrounds. They commonly use various means including Blogs to share the diaries, RSS feeds to follow the latest information of their interest and Computer Mediated Chat (CMC) applications to hold bidirectional communications. Microblogging is one of the most recent products of CMC, in which users talk about their daily lives, publish opinions or share information by short posts. It was first known as Tumblelogs on April 12, 2005, and then came into greater use by the year 2006 and 2007, when such services as Tumblr and Twitter arose. 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.

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.

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 [1] and [2], 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. For processing the content of tweets, we intend to use resources and linguistic methods Our experimental result uses a corpus of thou-

sand 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 related works. In section 3 we present an Analysis of short texts using NLP. In Section 4, we discuss experiments and obtained results. Finally, section 5 concludes this paper and outlines future work.

2 Related works

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), 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. 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.
Several works have focused on the analysis of data posted on microblogs, particularly in Twitter. [3] and [4] propose approaches for sentiment classification of Twitter messages i.e. determine whether tweets express a positive, negative or neutral feeling. Positive and negative polarities correspond respectively to a favorable and unfavorable opinion. To solve this task the authors have used natural language processing and machine learning techniques.
Many studies have found that there is a high correlation between the information posted on the web and actual results. [5] have used tweets to analyze awareness and anxiety levels of Tokyo habitants the events of earthquakes tsunami and states of nuclear emergencies in Japan in 2011. [6] have presented a method to measure the prevalence of H1N1 disease in the population of United Kingdom. They sought in the tweets the symptoms related to the disease. The obtained results were compared with real results from the Health Protection Agency. [7] also analyzed the tweets to predict public opinion and then compared the results with surveys.
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 [8], other studies like work of [9] 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.
Our work consists of examining the role and impact of social networks, in particular microblogs, on public opinion. We aim to analyze the behavior and text posted by users to extract knowledge that reflect the interests and opinions of a population. 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 Analysis of short texts using NLP

Data analysis of social networks has become a major trend in the field of natural language processing. Thus, large communities NLP gave its fair share to the analysis of data microblogs. In recent years, major conferences have created workshops for data analysis in social networks. Several studies concerned with the analysis of short texts do not aim only to determine the polarity of the messages, but to use the messages to detect events or predict results.
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:

- Content features refer to those features which describe the content relevance between queries and tweets.
- Tweet features refer to those features which represent the particular characteristics of tweets, as OOV and hashtags in tweet.
- Author features refer to those features which represent the authority of authors of the tweets in Twitter.

3.1 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 [10] 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.

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

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, it is not sensitive to the size of each tweet:

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

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.
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:

$$Quality(T) = 1 - \frac{NumberofOOV(Ti)}{Lg(Ti)} \quad (3)$$

3.2 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. Retweets they are supposed to contain no more information than the original tweets.
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.

3. Favor(Ti): this feature aims to calculate the number of times a tweet is classified as a favorite. According to [13], if a message is considered by many followers as a favorite, it means that it is relevant.
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.
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 [11][14] and [15], tweets containing URLs are more informative.

3.3 Author Relevance Features

Each blogger has specific characteristics such as number of follower and number of mention Accordind to [15,16], 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.

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:

$$\begin{aligned} Score(Ti, Q) = scoreContent(Ti, Q) \\ +\beta * scoreTweet(Ti, Q) \\ +\gamma * scoreAuthor(Ti, Q) \end{aligned} \tag{4}$$

With score (Ti,Q) on [0, 2] and β+γ=1. Where Ti and Q represent respectively, tweet and request. β and γ on [0,1] are a weighting parameter [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).

4.1 Experimental Evaluation

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.

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.

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.

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

1. We collected 3000 queries on recent actualities 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, we asked 450 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.

4.2 Results

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, 0,24462 when β = 0,4 and thus γ= 0,6.

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.

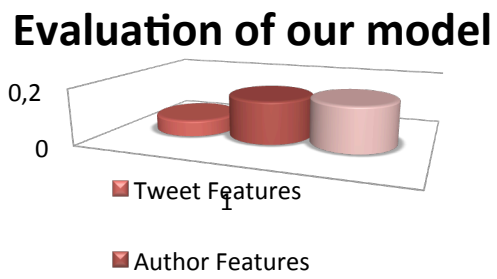

Fig. 1. - Comparing correlation coefficients

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.

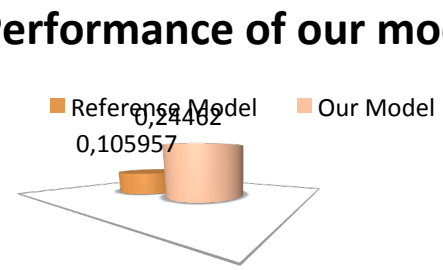

Fig. 2. - Comparing our model with reference model

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 0,24462. Can clearly be seen a 43% improvement in the satisfaction of our human judgment.

5 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.
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.

6 References

1. 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
2. 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
3. Barbosa. L and J. Feng. Robust sentiment detection on Twitter from biased and noisy data. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 36{44. Association for Computational Linguistics, 2010.

4. Jiang. L, M. Yu, M. Zhou, X. Liu, and T. Zhao. Target-dependent Twitter sentiment classification. Proc. 49th ACL: HLT, 1:151{160, 2011.
5. Doan. S, B.K.H. Vo, and N. Collier. An analysis of Twitter messages in the 2011 Tohoku earthquake. Arxiv preprint arXiv:1109.1618, 2011.
6. Lampos. V and N. Cristianini. Tracking the u pandemic by monitoring the social web. In Cognitive Information Processing (CIP), 2010 2nd International Workshop on, pages 411{416. IEEE, 2010.
7. OConnor. B, R. Balasubramanyan, B.R. Routledge, and N.A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In Proceedings of the International AAAI Conference on Weblogs and Social Media, pages 122{129, 2010.
8. 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.
9. 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.
10. Jones S., Walker K., Robertson S,. "A probabilistic model of information retrieval: Development and comparative experiments." Information Processing & Management, 36(6) :779–808, 2000
11. 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
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.
13. 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
14. 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.
15. 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
16. 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.