Uprising microblogs: A Bayesian network retrieval model for tweet search

Lamjed Ben Jabeur
IRIT, Université Paul Sabatier
118 Route Narbonne
Toulouse, France
jabeur@irit.fr

Lynda Tamine
IRIT, Université Paul Sabatier
118 Route Narbonne
Toulouse, France
tamine@irit.fr

Mohand Boughanem
IRIT, Université Paul Sabatier
118 Route Narbonne
Toulouse, France
boughanem@irit.fr

ABSTRACT

We investigate in this paper the problem of accessing to real-time information and we propose a Bayesian network retrieval model for tweet search. The proposed model interprets tweet relevance as a conditional probability and estimates it using different sources of evidence. In particular, we introduce a social search model that considers, in addition to text similarity measures, the microblogger's influence, the time magnitude and the presence of hashtags. To evaluate our model, we conducted a series of experiments on the *TREC Tweets2011* corpus. Experiments with "Arab Spring" topic set show that both of social and temporal features improve tweet search for different types of queries. Final results show also that our model outperforms other traditional information retrieval baselines.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval—*Retrieval models*

General Terms

Algorithms, Experimentation, Performance

Keywords

Microblogs, Tweet search, Influence, Time Magnitude

1. INTRODUCTION

A microblogging service is a new communication medium for information broadcasting and network collaboration. Unlike other social networking services, microblogging posts are particularly short and submitted in real-time to report an actual topic. In this paper, we are basically interested in Twitter as the most popular and widely used microblogging service. 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 microbloggers' posts, called *tweets*. Moreover, Twitter is distinguished by the *retweet* feature which gives users the ability to forward an interesting tweet to their followers. Finally, a microblogger, called also *twitterer*, can annotate his tweet using *hashtags*, address it to a specific user using *mentions* or share a web resource by adding a URL. Figure 1 summarizes the main entities involved in the social network of Twitter and the possible relationships between them.

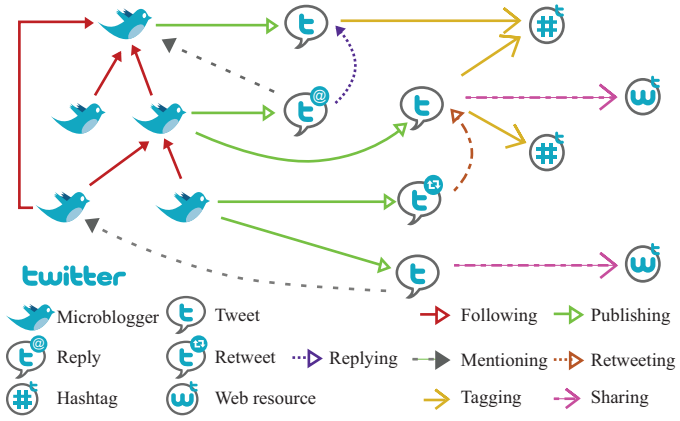


Figure 1: The social network of Twitter

According to the official statistics, about 50 millions tweets are submitted every day. With this important rate, microblogging data are more and more available while users are overwhelmed by the enormous quantity of new tweets and difficulty accessing to their interesting posts. A new retrieval task consisting on "tweet search" is therefore necessary. The tweet search task is defined by a retrieval status function $RSV(q, t_i, \theta_q)$ which estimates the relevance of tweet t_j considering the query q submitted at θ_q time. In the contrast of traditional Web search, tweet search aims at retrieving short, concise and real-time information about an actual topic or a recently occurred event. Due the specificity of microblogs, tweet search is confronted to several challenges compromising the indexing of the tweet stream [8], detecting spams [11], diversifying results [2] and evaluating the quality of tweets [4, 7, 6].

In this paper, we are interested in tweet search and we particularly investigate the quality of tweets. We evaluate the quality of a tweet based on the social influence of microbloggers which is computed by applying the *PageRank* algorithm on the social network of retweets. For the same purpose, we evaluate the quality of a tweet based on the publishing time. In this case, tweets submitted at activity

SAC'12 March 25-29, 2012, Riva del Garda, Italy.

periods of query terms are considered more significant. The social importance of microblogger and the time magnitude are integrated with other textual and tweeting features using a Bayesian network model which propagates conditional dependencies between them.

The remainder of this paper is organized as follows. In section 2, we give an overview of related work addressing microblogs and tweet search. We describe in section 3 the topology of the Bayesian network model for tweet search then we focus on the query evaluation process and the computation of conditional probabilities. In Section 4, we discuss experiments and obtained results. Finally, section 5 concludes this paper and outlines future work.

2. RELATED WORK

We address in this paper the tweet search task and we are particularly interested in tweet ranking. First proposed approaches in this context have addressed several features related to the spatio-temporal context, the tweeting features and the social network structure. Approaches based on the spatio-temporal context propose to classify tweets into geographical clusters and then aggregate local news [8]. In the same purpose, other work propose to identify the activity bursts of the query topic then select the most descriptive tweets for each period of time [4]. In addition to the spatio-temporal context, tweeting features such as URLs, hashtags and retweets have been addressed by some tweet ranking approaches as an alternative to the chronological listing of tweets. Authors in [3] propose a learning to rank system where tweeting features are used as input parameters. Finally, some social-based approaches propose to use the social network structure and rank tweets based on the importance of corresponding microbloggers. In this context, authors in [6] propose to linearly combine the textual similarly with a social score estimated from the social network.

These social social-based approaches for tweet ranking evaluate the importance of microbloggers using several categories of measures. In the first category, statistical approaches propose to use tweeting specific features in order to estimate a social importance score. Authors in [1, 6] propose to rank microbloggers based on received retweets, tweets number (*TweetRank*) and followers number (*FollowerRank*). In the second category, similar microbloggers are classified by clustering-based approaches into topical clusters. A small set of largely authoritative authors are then extracted and ranked to identify topical authorities of each cluster [7]. In the third category, link-based approaches propose to use the structure of the social network in order to identify important microbloggers. Authors in [5] propose to compute an influence score for each microblogger by applying the *PageRank* algorithm on the followers' social network. This work is extended by *TwitterRank* algorithm [10] that computes a topical sensitive *PageRank* on the followers network. Similarly, work in [3] propose to compute a popularity score by applying the *PageRank* algorithm on the retweet network instead of the followers network.

We introduce in this paper a social-based approach for tweet search that integrates different factors namely the social importance of microbloggers, the temporal context, the tweeting features as well as the tweet content. Unlike related work, our model is characterized by the following features:
- Tweet relevance estimation is addressed using a Bayesian network model that integrates all used features. Previous work uses clustering-based approaches or learning to rank methods to combine separated features [8, 3, 6].
- Microbloggers are represented with the retweet network in the contrast of the followers' social network used in [5, 10]. In this case, the social importance of microbloggers is assimilated to their influence instead of authority. We consider only the sub-network generated by retrieved tweets avoiding so the dominance of some celebrities if the entire retweet network is considered [3].
- The time magnitude of a tweet is estimated from each term's occurrence in the temporal neighborhood in the contrast of work in [4] analyzing all tweets to locate activity burst periods of a specific topic.

3. A BAYESIAN NETWORK MODEL FOR TWEET SEARCH

Tweet search is a particular retrieval task which is carried out by several motivations. It uses different sources of evidence that may influence each others. Accordingly, Bayesian networks seems to be an adequate model to represent this task as it supports, though random variables and conditional dependencies, modeling influenceable sources of evidence. The second motivation behind the use of such model is that Bayesian networks support incomplete or sampled data and approximate the values of missed parameters. This helps to perform the tweet search task even if access to the full network is limited or some meta-data are not available.

We present in this section the main features of our Bayesian network model for tweet search. After presenting the network topology and the query evaluation precess, we then focus on the computation of the conditional probabilities.

3.1 Network topology

The Bayesian network for tweet search is represented by a graph $G(X, E)$, where nodes $X = \mathcal{Q} \cup \mathcal{K} \cup \mathcal{T} \cup \mathcal{U}$ corresponds to the set of random variables and the set of edges $E = X \times X$ represents conditional dependencies among them. $\mathcal{Q}$, $\mathcal{K}$, $\mathcal{T}$ and $\mathcal{U}$ correspond respectively to the sets of queries, terms, tweets and microbloggers nodes. We detail in what follows the network nodes and edges.

3.1.1 Information nodes

As shown in figure 2, the Bayesian network nodes are classified into 4 layers.

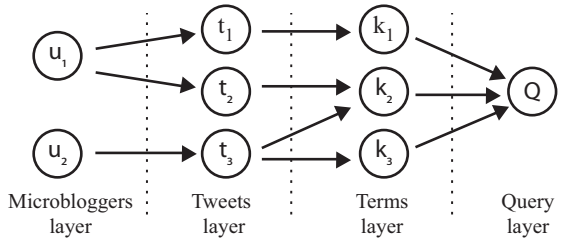


Figure 2: The Bayesian network of tweet search

- The query layer $\mathcal{Q}$: It includes the query node q which is associated with a binary random variable $q \in \{0, 1\}$. The short representation of $q = 1$ is noted q and denotes "the query q is observed". Conversely, $q = 0$ is noted $\bar{q}$ and denotes "the query q is not observed". We notice that the same notation q is used to refer to the query, the associated random variable and the corre-

sponding query node. The same notation is used for the rest of nodes in the network.

- **The terms layer $\mathcal{K}$:** It includes nodes of terms present in the query q and also included in the document index. A binary random variable $k_i \in \{0,1\}$ is associated to each term k_i.
- **The tweets layer $\mathcal{T}$:** includes nodes of tweets indexed at least with one parent term of the query q. A binary random variable $t_j \in \{0,1\}$ is associated to each tweet t_j.
- **The microbloggers layer $\mathcal{U}$:** includes nodes of microbloggers having published at least one tweet from the instantiated tweets layer. A binary random variable $u_k \in \{0,1\}$ is associated to each microblogger u_k.

3.1.2 *Information edges*

Each edge in the network express a conditional dependency between nodes. Edges connecting the query $q \in \mathcal{Q}$ with all parent terms $k_i \in \mathcal{K}$ represent the chance of generating the query from connected term. In its turn, an instantiated term k_i is connected to all parent tweets $t_j \in \mathcal{T}$ that it indexes. Edges from tweets to terms show that the event of observing a particular tweet impacts the observation of the connected term. Finally, a tweet node $t_i \in$ is connected to a single parent node corresponding to the microblogger $u_k \in \mathcal{U}$ having published t_j. This edge shows that the event of observing a tweet t_j depends on the observation event of the corresponding microblogger u_k. To avoid cycles in the graph, we assume that tweets and microbloggers are mutually independent between each other.

3.2 Query evaluation

The relevance of a tweet t_j considering a query q submitted at θ_q is assimilated to the joint probability that both events $t_j = 1$ and $q = 1$ appear. Ignoring the query submission time, this probability is computed as:

$$P(q \wedge t_j) = \sum_{\forall \vec{k}} P(q|\vec{k}) P(\vec{k}|t_j) P(t_j) \quad (1)$$

with $\vec{k}$ is one of the possible query parent configurations defined by a vector of random variables $\vec{k} = (k_1, k_2, ..., k_m)$, $k_i \in \{0,1\}$. Considering a query $q = \{k_1, k_2\}$ composed of two terms k_1 and k_2, the set of query parent configurations is represented by $\{(k_1,k_2),(k_1,\bar{k_2}),(\bar{k_1},k_2),(\bar{k_1},\bar{k_2})\}$.

Substituting $P(t_j)$ probability by $P(t_j|u_k) \times P(u_k)$, with u_k corresponds to the microblogger having published the tweet t_j, and developing the $P(\vec{k}|t_j)$ probability, the equation 1 can be rewritten as follows:

$$P(q \wedge t_j) = \sum_{\forall \vec{k}} P(q|\vec{k}) \times P(t_j|u_k) \times P(u_k) \times \left(\prod_{\forall i | on(i,\vec{k})=1} P(k_i|t_j) \times \prod_{\forall i | on(i,\vec{k})=0} P(\bar{k_i}|t_j) \right) \quad (2)$$

with $on(i,\vec{k}) = 1$ if $k_i = 1$ according to $\vec{k}$ and $on(i,\vec{k}) = 0$ if $k_i = 0$.

To deal with the query time, we propose in this paper to filter tweets by publishing time and keep only those which are posted before submitting the query. The relevance probability of a tweet t_j posted at θ_{t_j} is therefore updated with $P(q \wedge t_j) = 0$ if $\theta_{t_j} > \theta_q$.

3.3 Probability estimation

We focus in what follows on the conditional probabilities introduced in equation 2 and we present corresponding computing formulas.

3.3.1 *Computing probability* $P(q|\vec{k})$

The probability $P(q|\vec{k})$ of observing the query q having the parent configuration $\vec{k}$ helps to weight the different combinations of the query terms. We estimate the probability of query q with m parent terms $\{k_1, k_2, \ldots, k_m\}$ as follows:

$$P(q|\vec{k}) = p_1 \times p_2 \times \cdots \times p_m \quad (3)$$

with $p_i = on(i,\vec{k})$. We notice that $P(q|\vec{k}) > 0$ only if all query terms are positively instantiated in the query parent configuration $\vec{k}$. This don't discard tweets containing partial terms of the query but gives an absolute importance to the query parent configuration where all terms are instantiated.

3.3.2 *Computing probability* $P(k_i|t_j)$

The probability $P(k_i|t_j)$ of observing term k_i in the tweet t_j depends, on the one hand, on the term's occurrence and on the other hand on the tweet's properties. This probability is computed using the term frequency $F(k_i,t_j)$, the hashtag presence $H(k_i,t_j)$, the time magnitude $T(k_i,t_j)$ and the tweet length $L(t_j)$:

$$P(k_i|t_j) = (1-\mu)F(k_i,t_j)\ H(k_i,t_j) + \mu\ T(k_i,t_j)\ L(t_j) \quad (4)$$

$$P(\bar{k_i}|t_j) = 1 - P(k_i|t_j) \quad (5)$$

with $\mu \in [0..1]$ is a smoothing parameter and the closer μ is to 0, a higher importance is given to term's appearance rather than the tweet's properties. We note that in the case where the term is not present, a default probability is assigned to the tweet depending on its length and time magnitude.

Functions introduced in equation 4 compute a relevance probability for each feature. These functions are detailed in what follows.

- *Term frequency* $F(k_i,t_j)$.

Due to the limited tweet length, a given term is almost used once in the same tweet. Repeating the term will emphasize it but don't attribute it an absolute highlight compared to other terms naturally occurring once. We propose so to substitute the common *tf* measure with a graduated function $F(k_i,t_j)$ that map high frequencies into a small interval as follows:

$$F(k_i,t_j) \begin{cases} 1 - \frac{a}{tf_{k_i,t_j}}, & \text{if } k_i \text{ is present in } t_j \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

with $a \in [0..1]$ and tf_{k_i,t_j} is the frequency of the term k_i in the tweet t_j.

- *Hashtag score* $H(k_i,t_j)$.

Marking a term k_i with a hashtag $\#k_i$ would puts a highlight on it and increases its importance compared to other terms in the tweet. We propose a hashtag function $H(k_i,t_j)$ that leverages the importance of the terms as follows:

$$H(k_i,t_j) \begin{cases} 1 - \frac{b}{tf_{\#k_i,t_j}}, & \text{if } \#k_i \text{ is present in } t_j \\ b, & \text{otherwise} \end{cases} \quad (7)$$

with $b \in [0..0.5]$ is the default hashtag score and $tf_{\#k_i,t_j}$ is the frequency of the hashtag $\#k_i$ in the tweet t_j. We note in equation 7 that $H(k_i, d_j) \geq (1-b)$ if the hashtag $\#k_i$ occurs a least once in the tweet t_j.

- *Time magnitude* $T(k_i, t_j)$.

The term's importance varies over the time, increasing and decreasing with its presence in the tweet feeds. Therefore, the probability of observing term k_i depends also on the time when tweet t_j is submitted. This probability would be more important when term k_i is frequently used. Considering a term k_i, we measure the time magnitude of tweet t_j as follows:

$$T(k_i, t_j) = \frac{df_{k_i,\Gamma_j}}{|\Gamma_j|} \tag{8}$$

$$\Gamma_j = \{t_k, |\theta_{t_j} - \theta_{t_k}| \leq \Delta t\} \tag{9}$$

with Γ_j refers to the set of temporal neighbors of the tweet t_j within the $2\Delta t$ time window. df_{k_i,Γ_j} is the number of tweets in Γ_j containing the term k_i.

- *Tweet length* $L(t_j)$.

Very short tweets are considered ambiguous and pointless. Unlike common document length measures maximizing the scores of short documents, we propose to favour tweets closer to the average tweets length avg_{tl}. A length score $L(t_j)$ is assigned to each tweet as follows:

$$L(t_j) = \frac{1}{1 + \left|avg_{tl} - tl_{t_j}\right|} \tag{10}$$

3.3.3 *Computing probability* $P(t_j|u_k)$

The probability $P(t_j|u_k)$ of arriving to tweet t_j having a microblogger u_k weights the different tweets of one microblogger. Considering the set $\mathcal{T}_{u_k}$ of instantiated tweets published by the microblogger u_k, this probability is computed as follows:

$$P(t_j|u_k) = \frac{1}{|\mathcal{T}_{u_k}|} \tag{11}$$

3.3.4 *Computing probability* $P(u_k)$

Since microblogger are root nodes, we attribute a prior probability to them which reflect their social importance. A microblogger would receive a high importance if he influences the network and if corresponding tweets are largely spread. In this work, the social importance is interpreted to the influence of the microblogger and estimated by applying the *PageRank* algorithm on the social network of retweets. To avoid the dominance of some celebrities characterized by a high retweet number, we propose to apply the *PageRank* algorithm only on the sub-network of microbloggers generated by instantiated retweets. This helps to evaluate microbloggers influence regarding to a specific topic.

The social network of microbloggers is modeled by a graph $G = (U, R)$ where U represents the set of instantiated microbloggers in the Bayesian network and $R = U \times U$ denotes the set of retweet relationships. A microblogger u_i is included in the network only if one of his tweets contains a query term at least. A retweet relationship $e(u_i, u_j) \in R$ is defined from a microblogger u_i to u_j if only it exists one tweet a least, published by u_j and retweeted by u_i. Influence weights are assigned to retweet associations using the set of tweets retweetd by u_j, noted $Rt(u_i)$, and the set of tweets $Pub(u_j)$ published by u_j:

$$w_{i,j} = \frac{|Pub(u_j) \wedge Rt(u_i)|}{|Rt(u_i)|} \tag{12}$$

The influence score of the the microblogger u_i is computed iteratively by applying the *PageRank* algorithm on the retweet network as follows:

$$Inf^p(u_i) = \frac{d}{|U|} + (1-d) \sum_{u_j \in U, u_j \to u_i} w_{i,j} \frac{Inf^{p-1}(u_j)}{O(u_j)} \tag{13}$$

with $Inf^p(u_i)$ is the influence score of the microblogger u_i at the iteration p. $O(u_j)$ denotes the number of outgoing retweet associations from node u_j. Microblogger probability is initialized to $Inf^0(u_i) = \frac{1}{|U|}$, d is the random walk parameter as defined in the *PageRank* algorithm.

4. EXPERIMENTAL EVALUATION

We conducted a series of experiments in order to evaluate the performance of our model. The two main objectives of this evaluation consist on studying the impact of used features on tweet search effectiveness and comparing the proposed model with other information retrieval baselines. We present in this section the experimental setup then we discuss the obtained results

4.1 Experimental setup

- *Tweet corpus.*

We use in these experiments the *Tweets2011* corpus distributed for TREC 2011 Microblog track[1]. This collection includes about 16 millions tweets sampled from Twitter over 2 weeks during the uprising movement in 2011. We present in table 1 general statistics about this dataset.

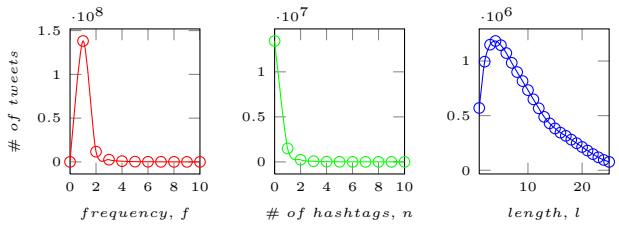
Tweets	16,141,812
Retweets	1,128,179
Tweet with hashtags	1,860,112
Unique terms	7,781,775
Unique hashtags	455,179
Microbloggers	5,356,432
Retweet associations	1,060,551
Retweet Net. nodes	5,495,081
Retweet Net. edges	1,024,914
Giant Component	11.12%

Table 1: Dataset statistics

Table 1 shows also general statistics about the social network. This dataset involves more than 5 millions microbloggers with an average of 3 tweets published by each one. The extracted social retweet network includes extra microbloggers as some retweets point outside the collection. This explains why the total number of nodes in the retweet network exceeds the number of microbloggers.

Figure 3 presents the distribution of term frequency, hashtags and document length. Figure 3.(a) shows that terms are often used once in the same tweet. Figure 3.(b) shows that only 2% of tweets contains 2 or more hashtags while the majority of tweets (88%) don't include any hashtag. Figure 3.(c) shows that the length distribution presents a peak at 4. We also report that 53% of tweets include 8 terms at least with an average tweet length estimated to 11.05.

[1]http://sites.google.com/site/microblogtrack/



(a) Term frequency (b) Hashtags (c) Tweet length

Figure 3: Term frequency, hashtags and length distributions

- *Topic set.*

As TREC 2011 Microblogs relevance assessments are not yet released, we used in these experiments the "Arab Spring" topic set that we asked 2 regular Twitter users to built it without a prior knowledge of the corpus. Queries of this topic set are about the "Arab Revolution" in which social media services have helped to accelerate the uprising movements. We note that queries express social, topical and timely information needs [9]. Socially motivated queries aim at tracking a person's activity (e.g. "Wael Ghonim"), find people with similar interests (e.g. "Tunisia Sidibouzid") and collect public sentiments about a topic (e.g. "Mubarak dissolves government"). Topical queries aim at finding informations about a specific topic (e.g. "Number of protesters in Tharir") or a general interest (e.g. "Tunisian revolution"). Temporal queries aim at retrieving tweets about news (e.g. "ElBaradei arrives in Egypt"), current events (e.g. "Clashes in Tahrir") and a service's status in real-time (e.g. "SMS down Egypt"). Finally, 25 time-stamped queries are collected. Top 20 tweets returned by compared models and configurations are evaluated by 4 volunteers who are also familiarized with Twitter and interested in the uprising movements. The evaluation process consists on a binary judgment of relevance based on the query type and the submission time. Among 3750 analyzed tweets, 849 relevant tweets are identified for the 25 queries.

- *Evaluation measures.*

We focus in this work on the retrieval effectiveness of our model and its ability to return relevant tweets in top results. We therefore study the p@10 and p@20 precisions, which correspond to the proportions of relevant tweets respectively over the top 10 and 20 retrieved tweets.

- *Model parameters.*

We configure the model parameters as follows: $\mu = 0.25$, $a = 0.25$, $b = 0.4$, $\Delta t = 1hour$ and $d = 0.15$.

- *Tweet processing.*

Tweets are indexed using NESTOR platform, a microblogging search engine developed in our team. This system recognizes tweeting features such as mentions and hashtags and detects declarative retweet associations "RT @username". It also separates URLs and email addresses from tweet and indexes only textual content. Finally, this system supports multi-language tokenisation and uses the Porter stemming algorithm for recognized English text.

4.2 Evaluation of features impact on tweet search

We compare in these experiments several configurations of our model. This study consists on disabling a specific feature for each configuration and compare its performance to the initial model. A short description of the different configurations is presented in table 2. We note that a precision decline compared to the initial model helps to conclude that the corresponding ignored feature improves the retrieval effectiveness of tweet search and vice versa.

BNTS	Bayesian network model for tweet search
BNTS-L	BNTS model with *Tweet Length* feature ignored $L(t_j) = 1$
BNTS-T	BNTS model with *Time magnitude* feature ignored $T(k_i, t_j) = 1$
BNTS-H	BNTS model with *Hashtag feature* ignored $H(k_i, t_j) = b$
BNTS-S	BNTS model with *Social influence* feature ignored $Inf^p(u_i) = 1$
BM25	Okapi BM25 probabilistic model
VSM	Vector Space Model
BM	Boolean Model with tweets are ranked by present terms then reverse chronological order

Table 2: Models notations

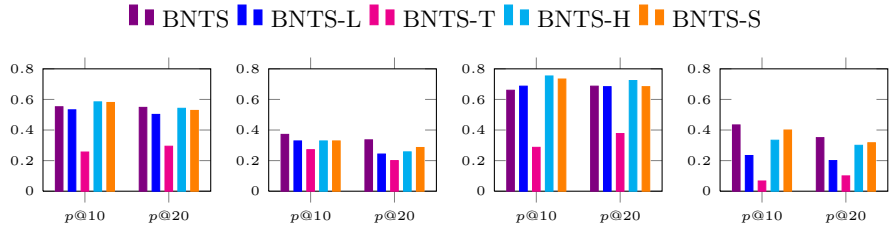
Figure 4.(a) shows that the different configurations have close precisions except the *BNTS-T* model presenting a considerable decline (-54%) compared to the *BNTS* model. We conclude so that the time mgntitude is a primordial feature for tweet search. The impact of the oher features varies depending on the query type. Analyzing the performance of the *BNTS-S* model, we note a general precision decline in figures 4.(b) and 4.(d) while in figure 4.(c) precisions of the *BNTS-S* model are similar or overpass their analogues for the *BNTS* model. We conclude so that the social influence of microbloggers is an important feature for social and temporal queries where users are interested in persons and fresh informations. This feature is less important for topical queries where users search for a specific information independently of the person who reports it. A precision decrease is also noted for the *BNTS-H* model in the case of social and temporal queries in contrast of topical queries where precision rises. We conclude that the hashtag feature is not helpful for specific topic search particularly when one of the query terms is frequently used as a hashtag. Considering the document length feature represented by the *BNTS-L* model, a significant precision decrease is observed in figure 4.(d). This explains that users are interested in short tweets in the case of temporal queries which mainly address news and real-time information.

4.3 Evaluation of the retrieval effectiveness

We compare in table 3 the performance of the *BNTS* model with some traditional baselines. We notice that the *VSM* model, which is a frequency-based retrieval system, show low precisions values compared to the *BM* baseline which completely ignores the term frequency. This confirms our assumption that tweet search strongly depends on the simple appearance of terms regardless from their frequency typically when the all the terms of the query are present.

	p@10	% Change		p@20	% Change	
BNTS	0.552			**0.548**		
BM25	**0.576**	-4%		0.494	11%	
BM	0.416	33%	**	0.382	34%	***
VSM	0.376	47%	**	0.360	52%	**

Table 3: Comparing the retrieval effectiveness. T-Test significance: $* \; p < 0.05$, $** \; p < 0.01$, $*** \; p < 0.001$.



(a) All queries (b) Social queries (c) Topical queries (d) Temporal queries

Figure 4: Features impact on the tweet search

Compared to the *BM* baseline, the *BNTS* model shows an important improvement that exceeds 30%. This improvement is achieved thanks to used features, namely, the tweet length, the hashtags, the time magnitude and the social influence of microbloggers. Compared to the *BM25* model, our *BNTS* model presents higher p@20 precisions but slightly lower p@10 values. With improvements for *BM* and *VSM* models are statistically significant, we can conclude that our *BNTS* improves the retrieval effectiveness compared to these traditional baselines. We note also through the slight decrease of precisions from p@10 to p@20 that our proposed model is able to maintain a stable performance.

5. CONCLUSION

We proposed in this paper a social model for tweet search that evaluates the quality of tweets using different sources of evidence. Theses features are integrated into a Bayesian network model that supports the conditional dependencies between them. One of the main integrated feature consists on the social influence of microbloggers estimated by applying the *PageRank* algorithm on the retweet network. In order to evaluate the tweet quality, we consider also the time magnitude of tweets which is estimated using the query terms occurrence in the temporal neighborhood.

Our experimental evaluation on the *Tweets2011* corpus shows that the time magnitude is a primordial feature for tweet search. We conclude also that the social importance of microbloggers and hashtags occurrence have an impact on the retrieval effectiveness particularly for temporal and social queries. Analyzing the term frequency, we conclude that tweet search depends on the simple occurrence of the query terms. Final comparison shows that our model significantly outperforms traditional information retrieval baselines.

For future work, we plan to extend the Bayesian network topology and represents the tweet publishing time and hashtags as network nodes. We will investigate also model parameters typically the time window Δt in order to automatically detect optimal value once some topics may evolve more or less faster. Finally, we will conduct experiments using TREC 2011 Microblogs topic set and compare our model with other tweet search models.

6. REFERENCES

[1] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

[2] M. D. C. S. C. M. Czerwinskiz. Find me the right content! diversity-based sampling of social media spaces for topic-centric search, 2011.

[3] Y. Duan, L. Jiang, T. Qin, M. Zhou, and H.-Y. Shum. 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, 2010. Association for Computational Linguistics.

[4] M. Grinev, M. Grineva, A. Boldakov, L. Novak, A. Syssoev, and D. Lizorkin. Sifting micro-blogging stream for events of user interest. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 837–837, New York, NY, USA, 2009. ACM.

[5] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

[6] R. Nagmoti, A. Teredesai, and M. De Cock. Ranking approaches for microblog search. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, pages 153–157, Washington, DC, USA, 2010. IEEE Computer Society.

[7] A. Pal and S. Counts. 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, 2011. ACM.

[8] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, pages 42–51, New York, NY, USA, 2009. ACM.

[9] J. Teevan, D. Ramage, and M. R. Morris. #twittersearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 35–44, New York, NY, USA, 2011. ACM.

[10] J. Weng, E.-P. Lim, J. Jiang, and Q. He. 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, 2010. ACM.

[11] S. Yardi and D. Boyd. Tweeting from the town square: Measuring geographic local networks. In *International Conference on Weblogs and Social Media*, 2010.