Featured tweet search: Modeling time and social influence for microblog retrieval

Lamjed Ben Jabeur, Lynda Tamine and Mohand Boughanem
IRIT, Paul Sabatier University
118 Route de Narbonne
F-31062 TOULOUSE CEDEX 9
{*jabeur, tamine, boughanem*}*@irit.fr*

***Abstract*—This paper interests in social search over social networking services, typically in microblogging networks. We propose a new approach that integrates, within a Bayesian network model, new relevance factors such as the social importance of microbloggers and the temporal magnitude of tweets. In particular, the social importance of a microblogger is assimilated to his influence on the social network. This property is evaluated by applying PageRank algorithm on the social network of retweets and mentions. The temporal magnitude of microblogs is estimated based on temporal neighbors that present similar query terms. To validate our approach, we conducted a series of experiments on the TREC 2011 Microblog dataset. Results show that the integration of social and temporal features increases the retrieval effectiveness.**

***Keywords*-Microblogs; Tweet search; Social network; Influence; Time magnitude**

I. INTRODUCTION

Microblogs are popular networking services that enable users to broadcast an information. Unlike news headlines which is generated by mass media, microblogs address general topics that interest a large public as well as small communities and close social networks. In addition, microblogs enrich reported news with valuable information. For instance, some particular events are covered in real-time with instant updates and live photos from the event site. Moreover, microblogs identify the exact source of information (author) and describe its publishing context (time, geolocalisation, application, device, etc). Finally, microblogs extended the informative purpose of message broadcasting and enable people to express their opinion about real world events.

With the variety of supported features, microblogging services emerge as a promising tool to get acquainted with the latest news. However, seeking for information over microblogging spaces becomes a challenging task due the increasing amount of published information. In the case of Twitter[1] microblogging service, which is the focus of this work, about 340 million[2] messages (called *"tweets"*) are published every day. A part of these tweets are useless, ambiguous, redundant or incredible [1]. A new information retrieval task is therefore created. Its main purpose is to search for real-time information and to rank recent tweets. TREC 2011 Microblog track [2] defines tweet search as a real-time adhoc task where the users are interested in most recent and relevant information. In the spite of Web search, tweet search aims to find temporally relevant information, monitor content and follow current events and people activities [3].

Prior works addressing tweet search integrate a variety of textual features, microblogging features and social network features [4], [5]. These works consider that tweet relevance depends, on the one hand, from the importance of corresponding authors in the social network and, on the other hand, from the content quality such as URLs, mentions and hashtags. We investigate in this paper different motivations behind tweet search, namely topical, temporal and social motivations. We propose an integrated Bayesian network model that considers:

- the number of query terms in the tweet as an indicator of topical overlap between the query and the tweet;
- the social importance of the related microblogger as an indicator of tweet credibility;
- the topic activity periods which corresponds to the joint events in the real world.

In particular, we estimate the tweet relevance based on the microblogger influence and the time magnitude. The influence score is computed by applying PageRank algorithm on the social network of retweet and mentions. The time magnitude is estimated from the set of tweets in the same period that contains similar query terms.

This paper is organized as follows. Section 2 presents an overview of related work. Section 3 introduces the Bayesian network model for tweet search. Section 4 focuses on query evaluation process and the computation of conditional probabilities. Section 5 discusses experiments conducted on TREC 2011 Microblog dataset. Finally, section 6 concludes the paper and outlines future work.

II. RELATED WORK

The first work that investigated microblogging services by Java et al. [6], has focused on the microblogging practices and the social network structure. Recently, Teevan et al. [3] showed in their systematic overview of search behavior, that tweet search emerges as a new information retrieval task that differs from typical Web search. Several retrieval approaches have been proposed for tweet search task. We summarize below some of representative works.

The first categories of approaches combine different relevance indicators computed separately. Chen et al. [7] propose to combine variety of features such as the authority of the microblogger computed by applying PageRank

[1] http://www.twitter.com/
[2] http://blog.twitter.com/2012/03/twitter-turns-six.html

algorithm on the follower network *U-PageRank*; the popularity of microblogger involved in the discussion topic or thread $Pop(\mathcal{T})$; the similarity between the query and the tweet $sim(q,t)$; the time decay between the query and the tweet $(q.timestamp - t.timestamp)$. In the same approach, Nagmoti et al. [4] propose a linear combination of social network based measures and information quality indicators. Social network factors are computed based on the number of published tweets (*TweetRank*) as well as the number of followers (*FollowerRank*). The information quality is evaluated based on the tweet length (*LengthRank*) and outgoing hyperlinks (*URLRank*).

The second categories of approaches investigates a machine learning algorithm in order to combine the relevance features. Duan et al. [5] propose a learning to rank approach that uses three types of features. Content relevance features evaluate the tweet text (*BM25 score*, *Similarity of contents*, *Length*). Twitter specific features evaluate tweet quality (*URL*, *Retweets*, *hashtags*, *replies*). Account authority features evaluate the tweet author. The main score in this category (*Popularity Score*) is computed by applying PageRank algorithm on the social network of retweets. Metzler and Cai [8] propose a learning to rank approach that considers the textual similarity to the query (*text score*), the time difference between the query and the tweet (*tdiff*), the hashtag existence (*has hashtag*), the URL presence (*has url*), the percentage of words out of vocabulary (*OOV*) and the tweet length (*length*).

The third category of approaches uses a language based model to combine tweet relevance features. Efron et al. [9] propose to integrate the topical relevance $Pr(Q|D)$ with both query and tweet temporal profiles $\log(\frac{m_{T_Q}}{m_{T_D}})$. The first factor m_{T_Q} is computed as the timestamps mean of retrieved tweets by the query Q. This score highlights new tweets if the query tends toward retrieving new documents. m_{T_D} is computed as the time-stamps mean of tweets retrieved by the pseudo-query of tweet D. This score highlights the temporal coherence between the query and candidate relevant tweets.

Besides the previously cited approaches that mainly investigate tweet properties, query-oriented approaches have addressed the problem of tweet shortness, vocabulary variation and term ambiguity. To tackle this problem, Bandyopadhyay et al. [10] propose to expand the query based on the title of Web search results. In the same approach, Li et al. [11] propose to extract the words with a strong connection to the topic in order to expand the query. Term similarly is estimated in this case based on the term association network and the term resistance network.

We propose in this work an integrated approach for tweet search that combines within a Bayesian network model different sources of evidence, namely the topical, the social and the temporal evidence. In particular, the relevance of a tweet is estimated based on its topical similarity to the query, the influence of corresponding microblogger and the time magnitude of the tweet. Our approach differs from related work in at least three respects :

- We model the tweet relevance within a integrated framework that support influenceable sources of evidences unlike previous work computing one or more scores for each source of evidence then combine them using a learning to rank approach [4].
- We model microbloggers using a weighted social network of retweets and mentions. Meanwhile, previous works model microblogger using a binary social network based only on followerships [7], retweets [5] or mentions. We compute a PageRank-like algorithm on the social network of retweets and mentions in order to identify active influencers in the network.
- We estimate the time magnitude of the tweet from the occurrence of query term configuration in temporal neighborhood unlike previous work [9] analyzing all tweet distributions regardless to the importance of each group of terms present in the query.

III. A Bayesian network model for tweet search

Tweet search is a particular information retrieval task driven by a variety of topical, social and temporal motivations. To perform this task, it is necessary to consider the sources of evidence behind these motivations. In fact, involved sources of evidence are mutually dependent. With this in mind, we propose to model tweet search using Bayesian network models that incorporate different sources of evidence into an integrated framework. This family of models supports the dependency between the integrated features. In addition, such Bayesian networks model ensures the retrieval process even though some data is unavailable such as a protected microblogger profile or when only a part of data is available. In this section, we first introduce some definitions and notations then we describe the proposed Bayesian network model topology.

A. Definitions and notations

Bayesian networks: A Bayesian network is a graphical model that represents random variables and conditional dependencies between them. Bayesian networks are modeled by a directed and acyclic graph $G(X,E)$, where the set of nodes X correspond to random variables and the set edges $E = X \times X$ represent conditional dependencies between them. Let X_i be a random variable and $Pa(X_i)$ the set of its parent nodes. The joint probability for all variables in the network is computed as $P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|Pa(X_i))$.

Term: Each term k_i in the index is associated to a random variable $k_i \in \{0,1\}$. The event of *"observing term k_i"* is noted $k_i = 1$ or shortly k_i. $k_i = 0$ denotes *"the term k_i is not observed"*. This event is noted also $\bar{k_i}$. We notice that the same notation k_i is used to represent the term k_i as well as the corresponding random variable and the network node. Let p be the number of index terms. It exists 2^p possible combinations between terms, called term configurations. For instance, an index of 2 terms (k_1 and k_2) presents 4 possible configurations represented by the set $\{(k_1,k_2),(k_1,\bar{k_2}),(\bar{k_1},k_2),(\bar{k_1},\bar{k_2})\}$. Each configuration

may represent a tweet or a query. A term configuration is noted $\vec{k}$. $on(k_i, \vec{k})$ associates to each k_i, the value of corresponding random variable in $\vec{k}$. $on(k_i, \vec{k}) = 1$ if term k_i is positively instantiated in $\vec{k}$. $c(\vec{k})$ represents the set of positively instantiated terms in $\vec{k}$.

Tweet: By analogy to the basic Bayesian network model, *tweets* are equivalent to documents. Each tweet t_j is associated to a random variable $t_j \in \{0,1\}$. The event $t_j = 1$ of *"observing tweet t_j"* is noted t_j. The complementary event $t_j = 0$ is noted $\bar{t}_j$. A tweet t_j is represented by a set of terms $t_j = k_1, ..., k_i, ..., k_n$ with k_i is a random variable indicating either term k_i is present in the tweet t_j or not. In addition, we propose to associate to each tweet t_j three other random variables t_{kj}, t_{sj} and t_{oj}. First variable t_{kj} models the event of observing t_j given an implicit knowledge of term occurrence in the tweet. The random variable t_{sj} models the event of observing t_j given an implicit knowledge of microblogger social influence. Finally, the random variable t_{oj} models the event of observing tweet t_j given an implicit knowledge of the time magnitude of tweet. These probabilities decompose the event of observing the tweet into three evidences: topical evidence, social evidence and temporal evidence.

Microblogger: Each microblogger is represented by a node u_f in the Bayesian network. A random variable $u_f \in \{0,1\}$ is associated to each microblogger. $u_f = 1$, shortly written as u_f, denotes *"microblogger u_f is observed"*. $u_f = 0$, noted $\bar{u}_f$, denotes *"microblogger u_f is not observed"*.

Period: A period o_e corresponds a time window with a duration Δt. Each period covers a temporal interval defined by $\left[\theta_{o_e} - \frac{\Delta t}{2}, \theta_{o_e} + \frac{\Delta t}{2}\right]$, with the period timestamp θ_{o_e} corresponds to the center of the temporal interval. Successive periods can not be parallel or overlapped $\theta_{o_e} - \theta_{o_{e-1}} >= \Delta t$. A random variable $o_e \in \{0,1\}$ is associated to each period. $o_e = 1$, noted o_e, denotes *"the period o_e is selected"*. $o_e = 0$, noted $\bar{o}_e$, denotes *"the period o_e is not selected"*.

B. Belief network topology

We describe in figure 1 the topology of our Bayesian network model for tweet search. This model is inspired from work of Pinheiro et al. [12] that proposes to integrate topical and hyperlink-based authority evidences into a Bayesian belief network. Unlike inference Bayesian networks, where the query node represents the network root, terms as considered as network nodes in belief networks. Thus, query and tweets are modeled in two separated layers, allowing so to integrate additional sources of evidence in each layer. The Bayesian network model for tweet search is comprised of 3 connected networks:

1) Tweet network: The Bayesian network model represents each query term k_i with a node. The set of these nodes constitute the term layer K. A user query is modeled by a node corresponding random variable $q \in \{0,1\}$. It exists a directed edge (k_i, q) from the query node to parent term k_i if only $on(k_i, q) = 1$. A tweet t_j is represented at first time by three nodes t_{kj}, t_{sj} and t_{oj}. Respectively,

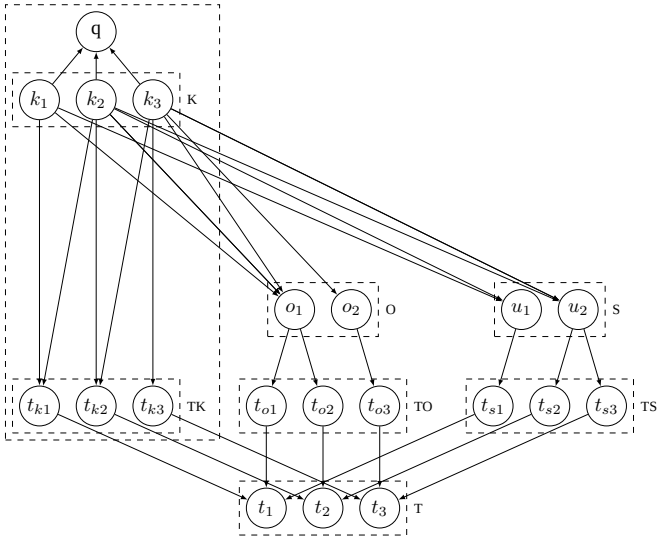

Figure 1: Belief network model for tweet search

these nodes belong to the topical evidence layer TK, the social evidence layer SO and the temporal evidence layer TS. t_{kj}, t_{sj} and t_{oj} are connected to a another node t_j. For $x \in \{k, s, o\}$, it exists an edge (t_{xj}, t_j) from t_{xj} to t_j. The set of nodes t_j constitutes the tweet layer T. We notice that t_{kj} is the only tweet node connected directly to term nodes. Thus, an edge (k_i, t_{kj}) connects t_{kj} to each included term k_i if $on(k_i, t_j) = 1$.

2) Microblogger network: Each microblogger u_f is represented by a node. These nodes constitute the social layer S. Microbloggers nodes are connected to correspondent tweet nodes in the social evidence layer TS. An edge (u_f, t_{sj}) is defined between a microblogger u_f and a tweet node t_{sj} if the tweet t_j is published by u_f. We notice that tweet t_v and a respective retweet t_w are represented by two independent nodes. In this case, retweet node t_w is connected to the retweeting microblogger instead of the original author of tweet t_v. In addition, microbloggers are connected to term nodes in layer K. An edge (k_i, o_e) connects a microblogger u_f to each term k_i appeared in one of his tweet at least $\{k_i \in K, (u_f, t_{sj}) \in E \wedge on(k_i, t_j) = 1\}$.

3) Period network: Each period o_e is represented by a node. Period nodes constitute the temporal layer O. Periods are connected to nodes from tweet temporal layer TO and term layer K. An edge (o_e, t_{oj}) connects a period o_e to a tweet node t_{oj} if t_j is published in the respective time window $\left|\theta_{t_j} - \theta_{o_e}\right| \leq \frac{\Delta t}{2}$. Once periods are not overlapped, a tweet is connected to one only period. Besides, a node o_e is connected to each term node k_i observed in the respective period $\{k_i \in K, on(k_i, t_j) = 1 \wedge \left|\theta_{t_j} - \theta_{o_e}\right| \leq \frac{\Delta t}{2}\}$.

IV. Tweet ranking

A. Query evaluation

The relevance of a tweet t_j with respect to a query q submitted at θ_q is computed by the probability $P(t_j|q, \theta_q)$.

Ignoring the query date, this probability is estimated by:

$$P(t_j|q) = \frac{P(t_j \wedge q)}{P(q)} \quad (1)$$

$P(q)$ have a constant value for all the tweets. $P(t_j|q)$ is then approximated with $P(t_j|q) \propto P(t_j \wedge q)$. Based on the topology of the Bayesian network for tweet search, the probability $P(t_j|q)$ is developed as follows:

$$P(t_j|q) \propto \sum_{\vec{k}} P(q|\vec{k})P(t_j|\vec{k})P(\vec{k}) \quad (2)$$

$\vec{k}$ is a term configuration. To simplify the computation of probability $P(t_j|q)$, only instantiated terms in the query are considered in the configuration $\vec{k}$.

In fact, the probability $P(t_j|\vec{k})$ depends on 3 sources of evidence: topical evidence, social evidence and temporal evidence. This probability $P(t_j|\vec{k})$ is rewritten as follows:

$$P(t_j|\vec{k}) = P(t_{kj}|\vec{k})P(t_{sj}|\vec{k})P(t_{oj}|\vec{k}) \quad (3)$$

By substituting $P(t_j|\vec{k})$ in formula 2, tweet relevance is estimated as:

$$P(t_j|q) \propto \sum_{\vec{k}} P(q|\vec{k})P(t_{kj}|\vec{k})P(t_{sj}|\vec{k})P(t_{oj}|\vec{k})P(\vec{k}) \quad (4)$$

In order to respect the temporal constraint in tweet search, we filter all the tweets with corresponding date θ_{t_j} is posterior to query date θ_q. We set relevance probability to $P(t_j|q) = 0$ for each tweet t_j where $\theta_{t_j} > \theta_q$.

B. Computing conditional probabilities

1) Probability $P(\vec{k})$: The probability $P(\vec{k})$ corresponds to the likelihood of observing the term configuration $\vec{k}$. We assume that all the configurations are independent and have an equal probability to be observed. Let n be the query q length which corresponds also to the number of terms represented in the configuration $\vec{k}$, the probability $P(\vec{k})$ is estimated as:

$$P(\vec{k}) = \frac{1}{2^n} \quad (5)$$

2) Probability $P(q|\vec{k})$: The probability $P(q|\vec{k})$ of generating the query q from a term the configuration $\vec{k}$ weights the different term configurations. First, we propose to weight each term k_i according to its appearance in the collection $w_{k_i} = \frac{df_{k_i}}{N}$ with df_{k_i} is the number of tweets containing k_i and N is the number of posterior tweets to the query q. According to the set of positively instantiated terms in the configuration $c(\vec{k})$, the probability $P(q|\vec{k})$ is computed using the Noisy-Or operator:

$$P(q|\vec{k}) = \begin{cases} \frac{1 - \prod_{k_i \in c(\vec{k}) \wedge q} w_{k_i}}{1 - \prod_{k_i \in q} w_{k_i}}, & \text{if } c(\vec{k}) \neq \varnothing \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

Thus, configurations with significant rare terms in the collection are highlighted in contrast of configurations that present commonly used terms.

3) Probability $P(t_j|\vec{k})$: The probability $P(t_j|\vec{k})$ that tweet t_j is generated by the configuration $\vec{k}$ measures the topical similarity between the tweet and the configuration. This probability could be estimated based on the term frequency tf_{k_i,t_j}. However, terms have less chance to be repeated once tweet length is limited. Therefore, we propose to weight each term k_i as following:

$$w_{k_i,t_j} = \begin{cases} \frac{tf_{k_i,t_j} - \beta}{tf_{k_i,t_j}}, & \text{if } on(k_i,t_j) = 1 \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

tf_{k_i,t_j} is the frequency of k_i in tweet t_j. $\beta = \frac{1}{1+n}$.

w_{k_i,t_j} map high frequencies into a small interval. We note that small value of β reduces the weight of frequent terms. Accordingly, we give less importance to term frequency rather than term presence in the case of long queries. With the value of β is dynamically configured in function of the query length n, term repetition would be less effective for short queries and vice versa.

The probability $P(t_j|\vec{k})$ is finally computed as:

$$P(t_j|\vec{k}) = \begin{cases} \frac{\sum_{k_i \in c(t_j) \wedge c(\vec{k})} w_{k_i,t_j}}{\sum_{\vec{k}} \sum_{k_i \in c(t_j) \wedge c(\vec{k})} w_{k_i,t_j}}, & \text{if } c(t_j) \wedge c(\vec{k}) \neq \varnothing \\ \delta, & \text{otherwise} \end{cases} \quad (8)$$

δ is a default probability.

4) Probability $P(t_{sj}|\vec{k})$: The probability $P(t_{sj}|\vec{k})$ of observing the tweet t_j having the social influence of corresponding microblogger and term configuration $\vec{k}$ is estimated as follows:

$$P(t_{sj}|\vec{k}) = P(t_{sj}|u_f)P(u_f|\vec{k}) + P(t_{sj}|\bar{u}_f)P(\bar{u}_f|\vec{k}) \quad (9)$$

The probability $P(t_{sj}|\bar{u}_f)$ of observing the tweet while corresponding microblogger u_f is not observed, is equal to 0. The probability $P(t_{sj}|\vec{k})$ is therefore transformed to:

$$P(t_{sj}|\vec{k}) = P(t_{sj}|u_f)P(u_f|\vec{k}) \quad (10)$$

Assuming that the two events of observing microblogger u_f and configuration $\vec{k}$ are independent, we write:

$$P(t_{sj}|\vec{k}) = P(t_{sj}|u_f)P(u_f) \quad (11)$$

First, the probability $P(t_{sj}|u_f)$ of observing tweet t_j having the microblogger u_f weights the tweets of each microblogger. This probability is computed equally for set of tweets $\tau(u_f)$ published by microblogger u_f.

$$P(t_{sj}|u_f) = \frac{1}{|\tau(u_f)|} \quad (12)$$

The prior probability $P(u_f)$ of observing microblogger u_f is related to his position in the social network. A microblogger would have more chance to be observed if he receives many retweets to his published tweets. This expresses the microblogger influence on the social network. On another hand, mentions express the authority of the microblogger as replies reflect the attention that other microbloggers give to his tweets. Mentions express also microbloggers motivation about an introduced topic.

Accordingly, we consider in the social network both retweet and mentions association.

The social network of microbloggers is modeled by a mulitgraph $G = (U, E)$ where the set of nodes U represents instantiated microbloggers in the Bayesian network and the set of edges $E = U \times U$ denotes the set of relationships between them. R and M are respectively the set of retweeting associations and mentioning associations with $E = R \cup M$. A microblogger u_i is included in the network if he published one or more tweets containing at least one term of the query. A retweet relationship $(u_i, u_j) \in R$ is defined from microblogger u_i to microblogger u_j if u_i retweets a tweet from u_j. A mentioning relationship $(u_i, u_j) \in M$ is defined from the microblogger u_i to the microblogger u_j if u_i mentions u_j in at least one of his tweets. These relationships are weighted as follows:

$$w_{i,j} = \begin{cases} \frac{|\rho_{out}(u_i) \wedge \rho_{in}(u_j)|}{|\rho_{out}(u_i)|}, & \text{if } (u_i, u_j) \in R \\ \frac{|\sigma_{out}(u_j) \wedge \sigma_{in}(u_j)|}{|\sigma_{out}(u_i)|}, & \text{otherwise} \end{cases} \quad (13)$$

$\rho_{in}(u_i)$ and $\rho_{out}(u_j)$ are respectively the set of incoming retweets and the set of outgoing retweets of microblogger u_i. $\sigma_{in}(u_i)$ and $\sigma_{out}(u_j)$ are respectively the set of tweets where u_i has been mentioned and the set of tweets where u_i has mentioned another microblogger.

The social importance of the microblogger $P(u_f)$ is estimated by computing a weighted PageRank algorithm on the social network of retweets and mentions. An influence score is attributed to each microblogger as follows:

$$Inf^p(u_i) = \frac{d}{|U|} + (1-d) \sum_{u_j : e(u_j, u_i) \in E} w_{i,j} \frac{Inf^{p-1}(u_j)}{O(u_j)} \quad (14)$$

p is the number of the current iteration. $O(u_j)$ is the number of outgoing relationships from u_j. $d \in [0, 1]$ is the PageRank random surfer parameter. $w(j, i)$ is computed according the edge type as defined in formula 13. At each iteration the influence score $Inf^p(u_i)$ is normalized by the sum of all microblogger scores.

In the case where the access to all tweets is guaranteed, the probability of observing a microblogger is equal to his influence score $P(u_f) = Inf^p(u_f)$. Otherwise, this probability is computed proportionally to the percentage of the available tweet sample λ:

$$P(u_f) = \lambda Inf^p(u_f) + (1-\lambda) P_{default}(u_f) \quad (15)$$

$P_{default}(u_f)$ is the default probability of observing the microblogger u_f

5) Probability $P(t_{oj}|\vec{k})$: The probability $P(t_{oj}|\vec{k})$ of observing the tweet t_j knowing the tweet period o_e and the term configuration $\vec{k}$ is estimated as follows:

$$P(t_{oj}|\vec{k}) = P(t_{oj}|o_e) P(o_e|\vec{k}) + P(t_{oj}|\bar{o}_e) P(\bar{o}_e|\vec{k}) \quad (16)$$

The probability $P(t_{oj}|\bar{o}_e)$ of observing the tweet outside the respective period is equal to 0. Thus, $P(t_{oj}|\vec{k})$ is written as:

$$P(t_{oj}|\vec{k}) = P(t_{oj}|o_e) P(o_e|\vec{k}) \quad (17)$$

The probability $P(t_{oj}|o_e)$ of observing the tweet t_j, knowing period o_e, weights the different tweets published in o_e. We note that the visibility of a tweet increases with the number of received retweets. Consequently, this probability is computed proportionally to the number of retweets generated by t_j in the same period.

$$P(t_{oj}|o_e) = \frac{1 + |\rho_{o_e}(t_j)|}{|\tau(o_e)|} \quad (18)$$

$\rho_{o_e}(t_j)$ is the set of corresponding retweets of t_j in the same period o_e. $\tau(o_e)$ is the set of tweets published in o_e.

The probability $P(o_e|\vec{k})$ of selecting period o_e, having the configuration $\vec{k}$, weights the different periods. We estimate this probability based on two factors. First, we consider the time decay between period o_e and query date θ_q. In fact, recent tweets are more likely to interest microblog users. Second, we consider the percentage of tweets published in o_e and containing the configuration $\vec{k}$. This highlights active period of the configuration $\vec{k}$ that concurs with a real world event. Periods are weighted as following:

$$w_{o_e,\vec{k}} = \frac{\log(\theta_q - \theta_{o_e})}{\log(\theta_q - \theta_{o_s})} \times \frac{df_{\vec{k},o_e}}{df_{\vec{k}}} \quad (19)$$

θ_q, θ_{o_e} and θ_{o_s} are respectively the timestamps of the query q, the period o_e and the period o_s when the oldest tweet containing the term configuration $\vec{k}$ is published with $\theta_{o_s} \leq \theta_{o_e} \leq \theta_q$. $df_{\vec{k},o_e}$ is the number of tweets published in o_e and containing the configuration $\vec{k}$. $df_{\vec{k}}$ is the number of tweets with the term configuration $\vec{k}$.

The probability $P(o_e|\vec{k})$ is computed as:

$$P(o_e|\vec{k}) = \begin{cases} \frac{w_{o_e,\vec{k}}}{\sum_{\vec{k}} w_{o_e,\vec{k}}}, & \text{if } df_{\vec{k},o_e} > 0 \\ \gamma, & \text{otherwise} \end{cases} \quad (20)$$

γ is a default probability.

In case where only a sample of tweets is available, $P(o_e|\vec{k})$ is estimated by:

$$P(o_e|\vec{k}) = \lambda P(o_e|\vec{k}) + (1-\lambda) P_{default}(o_e) \quad (21)$$

λ is the percentage of tweet sample. $P_{default}(o_e)$ is the default probability of observing the period o_e.

V. Experimental evaluation

We conduct a series of experiments on TREC 2011 Microblog dataset in order to study the effectiveness of our model. We focus in this study on the query level and we analyze the impact of each integrated feature.

A. Experimental setup

Tweet and query dataset: These experiments are carried out on the *tweets2011* dataset distributed by TREC 2011 Microblog Track [2]. The dataset includes about 16 million tweets published over 16 days. Table I presents general statistics about the collection. We observe that 0.07% of tweets in the collection are retweets. Mentioning tweet represent 0.45% of total tweets. We notice that this dataset is built based on Twitter API which provides a

representative sample of 1% of the tweet stream [13]. Other tweets published in the same period are not included in the collection.

Tweets	16141812	Microbloggers	5356432
Retweets	1128179	Network nodes	5495081
Mentions	7193656	Network retweets	1061989
Terms	7781775	Network mentions	9503013

Table I: Dataset statistics

We extracted the social network from the tweets in the dataset. About 5.3 million microbloggers are found. We notice that the number of network nodes exceed the number of microbloggers inside the collection as presented in table I. This is explained by the fact that some retweets and mentions point to other users outside the collection. Each microblogger in the network is involved in about 0.19 retweet associations and 1.73 mention associations.

Real-time ad-hoc task: The real-time ad-hoc task of TREC 2011 Microblog includes 49 time stamped queries. In contrast of other TREC tasks where results are ranked by score, real-time search task ranks results by the inverse chronological order. p@30 precision is reported as the official measure. It evaluates the ability of a system to return relevant tweets in the top 30 results. The Mean Average Precision *MAP* is also referenced as a non official measure.

Tweet filtering and model parameters: In these experiments, we do not integrate any future data or external resource. Only tweet in English are included in the result set. In addition, retweets are removed once they are presumed irrelevant in this track. We propose also to remove all replies from the final result as discussion tweets would be irrelevant for this task. We notice that the filtering process is applied after the final ranking of tweets. The model parameters are set to: Δt = $1day$; δ = 10^{-n}; λ = 0.1; $P_{default}(u_f)$ = 0.5; d = 0.15; γ = $\frac{\Delta t}{\theta_q - \theta_{o_s}}$; $P_{default}(o_e)$ = 0.5.

B. Evaluating retrieval effectiveness

We compare our Bayesian Network model for Tweet Search *BNTS* to some models from TREC Microblog track:

- *isiFDL* (1^{st})*:* Learn to Rank model based on *MRF* model [8].
- *DFReeKLIM* (2^{nd}): *Kullback-Leibler* based model [14].
- *KAUSTRerank* (17^{th}): Learn to Rank model that considers user authority [15]. Basic run is noted *KAUSTBase*.
- *gust* (20^{th}): Language model that considers the query temporal profile [9].
- *Disjunctive:* Official track baseline (Boolean model).

Table II presents a comparison of p@30 and *MAP* with different thresholds on the result set size (*cutoff*). First, we note that the threshold choice impacts the retrieval effectiveness. In fact, time-ranked result set presents low error risk if only some few tweets are included.

BNTS presents an improvement of 33% compared to TREC p@30 median and an improvement of 24% compared to *MAP* median. A difference of about −25% is noted compared to 1^{st} model *isiFDL*. Considering the social-based models *KAUSTBase* and *KAUSTRerank*, *BNTS* shows inferior results, expect for *KAUSTBase MAP*. We notice that this model integrates URL-based feature. Compared to time-based model *gust*, *BNTS* presents higher p@30 values with the threshold set to 30. The *gust* model shows however higher p@30 with a cutoff at 300, and vice versa for *MAP* values. Considering p@30 for full result set (1000 tweets), our model presents an improvement of 37% compared to the *Disjunctive* baseline. We note also an improvement of 17% compared to the 2^{nd} ranked model *DFReeKLIM*.

		Cutoff	p@30		*MAP*	
isiFDL	*	30	**0.4551**	(−25%)	0.2439	(−27%)
DFReeKLIM	*	30	0.4401	(−22%)	**0.2811**	(−37%)
BNTS		30	0.3422		0.1774	
Median	*		0.2575	(+33%)	0.1426	(+24%)
gust	*	30	0.3218	(+6%)	0.1812	(−2%)
KAUSTRerank	*	50	**0.3456**	(−9%)	**0.2390**	(−17%)
KAUSTBase	*	50	0.3347	(−7%)	0.1902	(+5%)
BNTS		50	0.3129		0.1990	
gust		300	**0.3220**	(−31%)	0.1970	(12%)
BNTS		300	0.2231		**0.2201**	
BNTS		1000	**0.1844**		**0.1929**	
DFReeKLIM	*	1000	0.1136	(+62%)	0.1651	(+17%)
Disjunctive	*	1000	0.0986	(+87%)	0.1411	(+37%)

Table II: Comparison of p@30 and *MAP* (* official result)

Analyzing *BNTS* difference from p@30 median per topic in figure 2, we note an improvement on 37 topics out of 49. Highest negative difference concerns topic 18 which includes only one relevant tweet. Positive difference is noted for instance in topic 1 (*"BBC World Service staff cuts"*) which is characterized by a high number of tweet containing query terms (82581).

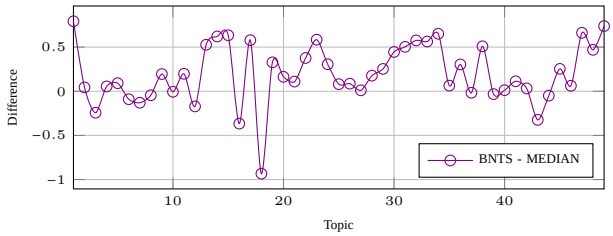

Figure 2: BNTS difference from p@30 median per topic

C. Feature based analysis

In order to study the impact of each relevance feature separately, we evaluate in these experiments different configurations of our model. The next results are computed with a threshold of 1000 tweets for each result set.

1) Topical relevance: We compare the topical configuration of our model *BNTS.K'* to 2 baselines. The first baseline *Boolean Frequency BF* computes a relevance as the number of present terms in the tweet. The second baseline *Term Frequency TF* considers the term frequency. We note that only the topical evidence is activated in our model: $P(q|\vec{k}) = 1$, $P(t_{sj}|\vec{k}) = 1$ and $P(t_{oj}|\vec{k}) = 1$. Figure 3 presents *BNTS.K' MAP* difference per topic. *BF* model shows higher precision than *TF* model. Term presence is therefore more significant for tweet search in contrast of long document retrieval. Main improvement of *BNTS.k'* compared to *TF* model, and at the same time main decrease compared to *BF* model, is observed for topic 17 *"White Stripes breakup"*. This is explained by the fact that all relevant tweets of this topic present the full name of the related music band *"White Stripes"*. On the other hand, commonly used term *"White"* is highly repeated in irrelevant tweets.

	BF	*TF*	*BNTS.K'*
MAP	0.1449	0.1432	**0.1474**

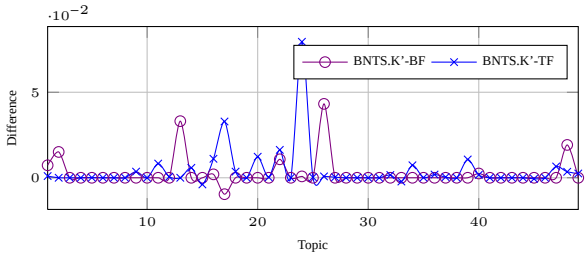

Figure 3: *BNTS.K'* difference from *BF* & *TF MAP* per topic

2) Social relevance: We compare the topical configuration of our model *BNTS.K*, $P(t_{sj}|\vec{k}) = P(t_{oj}|\vec{k}) = 1$, to the social configuration *BNTS.KS* where the temporal feature is desactivated $P(t_{oj}|\vec{k}) = 1$. Figure 4 presents *MAP* difference of *BNTS.KS* model from *BNTS.K* model. The

	BNTS.K	BNTS.KS
MAP	0.1579	**0.1642**

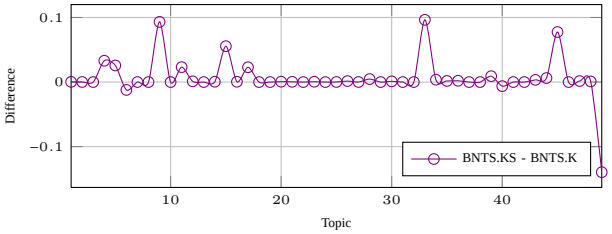

Figure 4: BNTS.KO difference from BNTS.K *MAP* per topic

overall improvement of *BNTS.KS* approve the significance of the social context for tweet ranking. An important positive change is noted for instance in the topic 9 *"Toyota Recall"*. In this case, relevant tweets are produced by network influencers such as *@tunkuv* (editor) and *@tjmarx* (filmmaker).

3) Temporal relevance: We compare the topical configuration of our model *BNTS.K* to the temporal configuration *BNTS.KO* where the social feature is deactivated $P(t_{sj}|\vec{k}) = 1$. Figure 5 presents *MAP* difference of *BNTS.KS* model from *BNTS.K* model. Considering all queries, *BNTS.KO* model shows an improvement of 17% compared to *BNTS.K*. We conclude that the temporal distribution is an indicator of tweet relevance. Main improvement of *BNTS.KO* configuration is observed for topic 4 *"Mexico drug war"*.

	BNTS.K	BNTS.KS
MAP	0.1579	**0.1909**

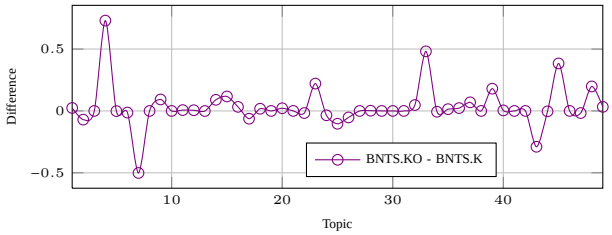

Figure 5: BNTS.KO difference from BNTS.K *MAP* per topic

Analyzing the related distribution of tweets over the time in figure 6, we observe that relevant tweets are mainly concentrated in the 5^{th} day before the query. Similar distribution is presented by tweets containing *"Mexico drug"* or *"drug war"* with some decay. Meanwhile, the distribution of all tweets or tweets containing only the term *"Mexico"* is regular. This confirms our choice to study the temporal distribution per term configuration instead of the global distribution of tweets which may be impacted by commonly used terms.

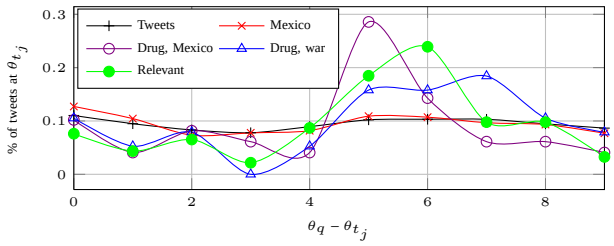

Figure 6: Temporal distribution of tweets (topic 4)

VI. Conclusion

We proposed in this paper a social model for tweet search that integrates, within a Bayesian network model, the topical relevance of tweets, the social relevance of microbloggers and the temporal relevance of tweet period. In particular, the topical relevance score highlights tweets

presenting all terms of the query rather that some repeated ones. The social score underlines tweets published by influencer microbloggers. Finally, the temporal score emphasizes tweets published in activity periods of the query topics. Experiments conducted on TREC 2011 Microblog dataset shows that the integration of the different sources of evidence enhances the quality of tweet search.

In future work, we plan to automatically detect the query profile and adjust the score of integrated features according to the sensibility of the query to the social and temporal contexts. In addition, we plan to represent hashtags and URLs entities in the Bayesian network model.

REFERENCES

[1] 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*, ser. GIS '09, 2009, pp. 42–51.

[2] I. Ounis, C. Macdonald, J. Lin, and I. Soboroff, "Overview of the trec-2011 microblog track," in *Proceeddings of the 20th Text REtrieval Conference (TREC 2011)*, 2011.

[3] 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*, ser. WSDM '11. New York, NY, USA: ACM, 2011, pp. 35–44.

[4] 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*, ser. WI-IAT '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 153–157.

[5] 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*, ser. COLING '10, 2010, pp. 295–303.

[6] A. Java, X. Song, T. Finin, and B. Tseng, "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*, ser. WebKDD/SNA-KDD '07, 2007.

[7] C. Chen, F. Li, B. C. Ooi, and S. Wu, "Ti: an efficient indexing mechanism for real-time search on tweets," in *Proceedings of the 2011 international conference on Management of data*, ser. SIGMOD '11. New York, NY, USA: ACM, 2011, pp. 649–660.

[8] D. Metzler and C. Cai, "Usc/isi at trec 2011: Microblog track," in *Proceeddings of the 20th Text REtrieval Conference (TREC 2011)*, 2011.

[9] M. Efron, A. Kehoe, P. Organisciak, and S. Suh, "The university of illinois' graduate school of library and information science at trec 2011," in *Proceeddings of the 20th Text REtrieval Conference (TREC 2011)*, 2011.

[10] A. Bandyopadhyay, M. Mitra, and P. Majumder, "Query expansion for microblog retrieval," in *Proceeddings of the 20th Text REtrieval Conference (TREC 2011)*, 2011.

[11] Y. Li, Z. Zhang, W. Lv, Q. Xie, Y. Lin, R. X. W. Xu, G. Chen, and J. Guo, "Pris at trec2011 micro-blog track," in *Proceeddings of the 20th Text REtrieval Conference (TREC 2011)*, 2011.

[12] M. A. P. de Cristo, P. P. Calado, M. de Lourdes da Silveira, I. Silva, R. Muntz, and B. Ribeiro-Neto, "Bayesian belief networks for ir," *International Journal of Approximate Reasoning*, vol. 34, no. 2-3, pp. 163 – 179, 2003, soft Computing Applications to Intelligent Information Retrieval on the Internet.

[13] I. Soboroff, D. McCullough, J. Lin, C. Macdonald, I. Ounis, and R. McCreadie, "Evaluating real-time search over tweets," in *Proceedings of the 6th international AAAI Conference on Weblogs and Social Media*, ser. ICWSM '12, 2012.

[14] G. Amati, G. Amodeo, M. Bianchi, A. Celi, C. D. Nicola, M. Flammini, C. Gaibisso, G. Gambosi, , and G. Marcone, "Fub, iasi-cnr, univaq at trec 2011," in *Proceeddings of the 20th Text REtrieval Conference (TREC 2011)*, 2011.

[15] J. Jiang, L. Hidayah, T. Elsayed, and H. Ramadan, "Best of kaust at trec-2011: Building effective search in twitter," in *Proceeddings of the 20th Text REtrieval Conference (TREC 2011)*, 2011.