A Time-Aware Language Model for Microblog Retrieval

Bingjie Wei, Shuai Zhang, Rui Li, Bin Wang

Institute of Computing Technology, Chinese Academy of Sciences

Beijing, China, 100190

Email: weibingjie, zhangshuai01,lirui, wangbin@ict.ac.cn

Abstract

This paper describes our work (the IIEIR participation) in the TREC 2012 Microblog Adhoc Track. We proposed a ranking algorithm with temporal information based query. More and more research work proved that time is an important factor for improving the search result, especially for Microblog search. Based on Language Model, the representative work used time information as the document's prior information. Intuitively, there were two ways for making use of this feature. One way was query relevant while the other was query irrelevant. The hypothesis of the two models is "the newer of the document, the more important". However, different query had different hot time points (the top time of relevance documents' time distribution). Take this into consideration; we supposed four models based on hot time points (HTLM). On this basis, we considered the model which is not relevant with query as document's background information and the model which is relevant with query as document's independent information. We used smoothing operation and supposed a mix timed language model. The results suggested that, HTLM models are more effective for Microblog search and mix model further improved compared with the single model.

Keywords: Microblog search; Microblog retrieval; time-aware; language model; inform ation retrieval

1 Introduction

Microblog, as a type of social media, has become immensely popular in recent years. There exists many Microblog websites, such as twitter. Compared the twitter queries and web queries, user's intent is always like to find more freshness information when searching in Microblog[1]. These queries can be called as time-aware queries. In our work, we try to use time information based on language model to improve the effectiveness of Microblog retrieval.

In the background of language model, a way of using time feature is as document's prior by defining a functional relationship. These methods can be categorized into two kinds by whether it is dependent to query. Li and Croft [2] assumed that the document which is newer has more probability to be read by user. So Li and Croft incorporated an exponential decay into the language model as document prior with manually parameter. This model is a time-independent model. Efron and Golovchinsky [3] expanded Li and Croft [2]'s work and proposed a time-aware language model with query information. Efron and Golovchinsky thought that the parameter of Exponential distribution should be different when queries were different.

Previous works assumed that, when query is given, a document is more important while it's more freshness. However, this hypothesis is not suitable for the real situation when query is time-sensitive. As shown in Fig 1, different query has different relevance documents' time distribution. At the same time, [2] also indicated the phenomenon. In this paper, we build on these findings. These peak points of distribution are defined as "Hot Time Point of Query", which means relevance documents are more likely to occur in these moments. We proposed four time-aware language models that everyone relies on Hot Time Points (HTLM). HTLM models belong to query-dependent model. From another perspective, a document has two kinds of time information, one is about background and the other is about query-specific. Both are important for document ranking. So we build a mixed model by using smoothing method and proved that the mixed model is more effective.

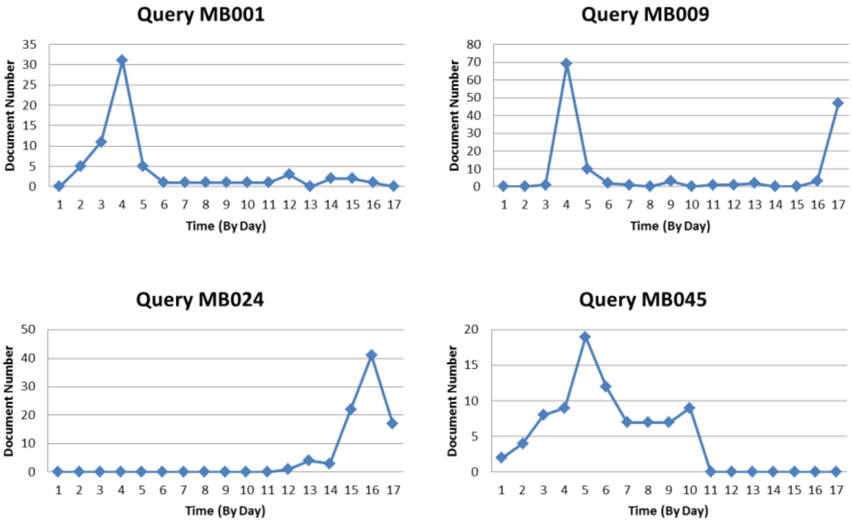

Fig.1 Relevant Documents' Time Distributions of Queries from TREC 2011 Microblog

2 Relevance Work

2.1 Language Model

Information retrieval area used language modeling frameworks first time in 1998 by Ponte and Croft [4]. In this paper, we rely on the query likelihood model. When a document d and a query q are given, the ranking function (1) is the posterior probability that the document multinomial language model generated query[5]. $p(d)$ in Eq.1 usually set to a fixed value or be ignored.

$$p(d\,|\,q) \propto \log p(d) + \sum_{w\in V} tf(w,q) \log p(w\,|\,m_d) \tag{1}$$

In order to avoid zero probability, Zhai and Lafferty [6] proposed multi smoothing methods. The simplest smoothing method is called Jelinek-Mercer smoothing, which gives:

$$p(w\,|\,m_d) = (1-\lambda) P_{ml}(w\,|\,m_d) + \lambda P_{ml}(w\,|\,m_c) \tag{2}$$

2.2 Timed Language Model

Incorporating document's creation date to language model as document prior has two kinds of ways. One is independent about query, which means that time distribution of the collection is used as document's background information. Li and Croft [2] used an exponential distribution of document's creation date in Eq.3, which shows newer documents have higher score.

$$P(d)=P(d\,|\,t_d)=re^{-r(T_c-T_d)} \tag{3}$$

Here T_c is the newest date in the collection, γ is the rate parameter of the exponential distribution. We will use LC as shorthand.

Efron and Golovchinsky [3] expanded Li and Croft's work. Efron and Golovchinsky pointed that document's prior is not only related with publication time but also related with specific query. So they change γ to γ_q for using query information. For a query q, we let pseudo document set $\mathrm{P}=\{d_1,d_2,\cdots,d_k\}$ and the time set of these documents $T=\{t_1,t_2,\cdots,t_k\}$. The function of computing γ_q is :

$$\gamma_q{}^{ml}=1/\overline{T} \tag{4}$$

Here $\overline{T}$ is the sample mean of T. In the later, we call this algorithm EGML as abbreviated.

3 Time-aware Mixed Language Model

3.1 Query Analysis

Li and Croft [2] split query to two categories. One is time-sensitive query which refers to these queries has more relevant documents in a specific period obviously, the other is not. Likewise microblog queries are almost time-sensitive queries[1]. We used 50 topics of TREC2011 Microblog Adhoc Track to analyze. Fig.1 shows four topics' time distribution of relevant documents as examples (MB001, MB009, MB024, and MB045). EGML assumed that the document which is newer is more important for specific query. However, some times, this assumption may not work. In Figure 1, we can find that the most relevant documents of the four queries do not distributed in the newest day, for example, query MB001 as the fourth day, and query MB045 as the fifth day. Specifically, for query MB001, EGML will improve the documents which publication time is in fourth and seventeen day where there have little relevant documents. If we performed in accordance with this assumption, the real relevant document will be punished leading to bad search result.

In response to this phenomenon, we hypothesize that the document is more likely to be related when its time is more closed to query's hot time:

Query's Hot Time: Given a query q, the peak points set of its relevant documents is called query's hot time. In particular, these points which are relatively higher will compose the set.

After defining query's hot time, the parameter T_c in Eq.3 is different value for different queries instead of the newest time of the collection. The intuition of this assumption is that documents which near query's hot time will get higher ranking location.

3.2 Hot Time Language Model (HTLM)

The analysis of query's relevant documents distribution in section 3.1 shows that for the time-sensitive queries, the peak points are different, which be called as query's hot time. In this hypothesis, the challenge is to get the query's hot time when the re levant documents cannot be known. In the past time, using pseudo relevance feedback in information retrieval is a popular way to be used in place of the relevant documents. The time distributions of queries' top 500 documents are shown in Fig.2. In the present study, we get first retrieval documents as the pseudo documents by using the query likelihood language model (Eq.1).

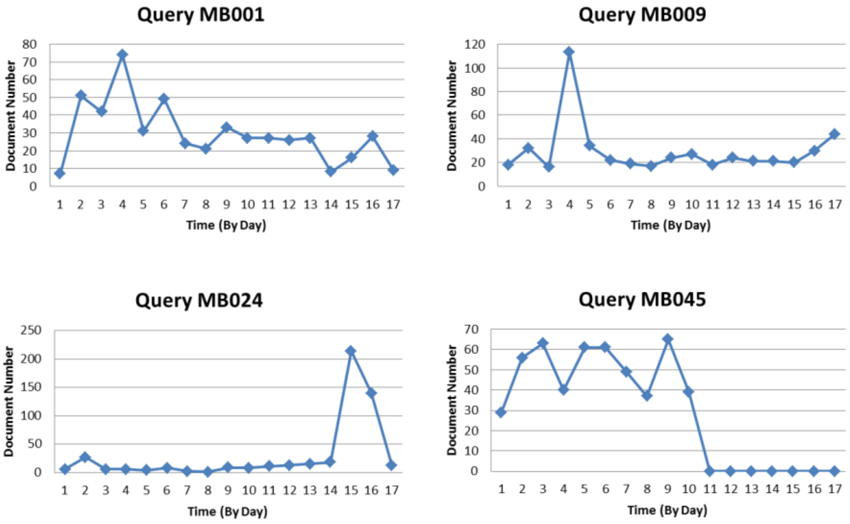

Fig.2 Pseudo Relevant Documents' Time Distributions of Queries from TREC 2011 Microblog

By seeing Fig.2, we can find that the peak points of pseudo relevant documents' time distribution is almost same with real relevant documents' distribution. This means we can use these pseudo peak points instead of the real peak points. According to statistics, there are 21 queries' pseudo peak point is same with real point, and 14 queries only one day. So we decide that using the peak point as query's real hot time approximately.

Some symbols are defined as follows: given a query q, the pseudo document set is $\mathrm{P}_q=\{d_1,d_2,\cdots,d_k\}$, $T_q=\{t_1,t_2,\cdots,t_k\}$ means the time set of P_q. So according to two methods of computing parameter γ, while $T_c=T_{cq}=\max(T_q)$, we will get two models. One is manually specified in Eq.3 (HTLM-LC). The other is given in Eq.4 (HTLM-ML).

Further, by comparing the four topics in Fig.2, we find that different topics have different number of peak points. In particular, query MB009 has one absolute high point, and query MB045 has four relatively high points. This finding is also in line with the mentioned content in Li and Croft's paper [2]. So we introduce query's hot time set as $HotT_{cq}s=\{ht_1,ht_2,\cdots,ht_{hn}\},ht_1<ht_2<\cdots<ht_{hn}$ by rule:

Rule: Set the highest number of the time distribution is $MaxDN$, if and only if next higher number is greater than $\alpha * MaxDN$, this time can be added to the hot time set $HotT_{cq}s$.

We introduce a variable α, values in 0.0~1.0, which means the degree of choosing hot time. .After getting the $HotT_{cq}s$ of query q, we define T_c:

$$T_c=T_{cqd}=\{ht_i : T_{cq}>ht_{i-1}\ and\ T_{cq}\le ht_i\}.$$

Then we also get two models by the computing parameter γ method (HTLM-AdaptiveMultiLC and HTLM-AdaptiveMultiML).

Corresponding to the characteristics of the query, we proposed four models based on query's hot time. The training of the parameters will be shown in section 4.

3.3 Mixed Time Language Model

As previously mentioned, we can classify the existing work into two parts, one is only based on the whole collection which are recorded as $P(d,t)$, and the other is relevant with specific query as $P(d,q,t)$. The algorithm without query information is defining the relationship about the document and time. This information can be seen as document's background time message. $P(d,q,t)$, corresponding to $P(d,t)$, considers the specific information which be included in query. This can be seen as document's query time message.

When given a query q and a document d, background and query time message of the document is indispensable. So we use smoothing strategy to combine the two kinds of information and propose a mixed time language model (MixTimedLM):

$$P(d)=\omega P(d,q,t)+(1-\omega)P(d,t)$$

where ω is a mixing parameter that controls the degree of smoothing. Experimental results show that mix model further improved than single model.

4 Experimental Design and Results

4.1 Experimental Data and Parameter Value

After preprocessing of Twitter Data, including removing RT tweets and no-English tweets and removing the @ and url information of tweet content and using Porter stemmer for content stemming, 9679710 tweets constitute the whole collection. Also there are 50 topics in TREC 2011 and 38079 labeled tweets.

We use MAP and P@30 of top 1000 results as evaluation indicator. In this paper, P@30 is mainly indicator instead of MAP specially. Baseline is the query likelihood language model (QL) with smoothing parameter β's value 0.4. Based on the test collection after preprocessing, we tuned the parameters that mentioned in the above models by using 5-fold cross validation. The values of these parameters are shown in Tab.1.

Tab.1 Model Parameters' Value chosen for best performance(P@30)

Model	Parameter	Description	Value
LC	γ	the rate parameter of exponential distribution	0.3
HTLM-LC	γ	the rate parameter of exponential distribution	0.2
HTLM-AdaptiveMultiLC	γ	the rate parameter of exponential distribution	0.5
	α	the degree parameter for choosing hot time	0.8
HTLM-AdaptiveMultiML	α	the degree parameter for choosing hot time	0.9

4.2 Timed Language Model Experiment

Table 2 shows that search results of these timed models on the training set. Comparied with Baseline, LC and EGML both improve MAP but decline P@30, with HTLM series models improving P@30 significantly and almost P@30. This performance of P@30 indicator improved may be due to the higher importance of these documents in query's hot time near. Because closed to more relevant tweets, the ranking locatio ns of these tweets are prompted in the rank list. Meanwhile, some non-relevant documents exists in the specific time, this measure may also change these documents' location and we will try to study in future work.

Tab.2 Retrieval Effectiveness on TREC 2011 Microblog Dataset

Model		*P@30*	*MAP*
BaseLine	*QL*	0.3252	0.3099
$P(d,t)$	*LC*	0.3244	0.3168
$P(d,q,t)$	*EGML*	0.3238	**0.3178**
	HTLM-LC	0.3327	**0.3146**
	HTLM-ML	0.3347	0.3023
	HTLM-AdaptiveMultiLC	0.3354	0.3038
	HTLM-AdaptiveMultiML	**0.3367**	0.3142

Next we will test the performance of the mixed time language model. According to table 1 and equation 5, we respectively choose the parameters value of the top 10 P@30 as the candidate models set of $P(d,t)$ and $P(d,q,t)$. Table 3 shown the top 3 models based on MAP and P@30 indicator, where ω is the smoothing degree in equation 5.

Tab.3 Mix Time-Aware Model's Retrieval Effectiveness on TREC 2011 Microblog Dataset

	$P(d,t)$	$P(d,q,t)$	ω	***P@30***	***MAP***
P@30Top3	LC $\gamma=0.2$	HTLM-AdaptiveMultiLC $\gamma=0.8$, $\alpha=0.9$	0.2	0.3374	0.3172
	LC $\gamma=0.3$	HTLM-AdaptiveMultiML $\alpha=0.6$	0.5	0.3374	0.3127
	LC $\gamma=0.3$	HTLM-AdaptiveMultiML $\alpha=0.9$	0.5	0.3374	0.3075
MAPTop3	LC $\gamma=0.3$	HTLM-AdaptiveMultiLC $\gamma=0.3$, $\alpha=0.8$	0.1	0.3320	0.3217
	LC $\gamma=0.3$	HTLM-AdaptiveMultiLC $\gamma=0.3$, $\alpha=0.4$	0.1	0.3306	0.3216
	LC $\gamma=0.3$	HTLM-AdaptiveMultiLC $\gamma=0.3$, $\alpha=0.5$	0.1	0.3299	0.3216

In Table 3, the retrieval performance of mixed model is better than QL, LG and EGML based on MAP or P@30 respectively. The $P(d,q,t)$ models are both HTLM-AdaptiveMulti series, which display that the hot time models are effectiveness.

4.3 Parametric of Models Sensitivity

There are two parameters that we have defined and set in this paper (Seen in Table 1). One is the rate parameter of exponential distribution: γ . The other is the degree parameter for choosing hot time: α . We will discuss one by one in the follow paper.

For parameter γ, its value range is 0.01~0.1 with 0.01 increase, and 0.1~1 with 0.1 increase. Figure 3 shows the change curve in the parameter.

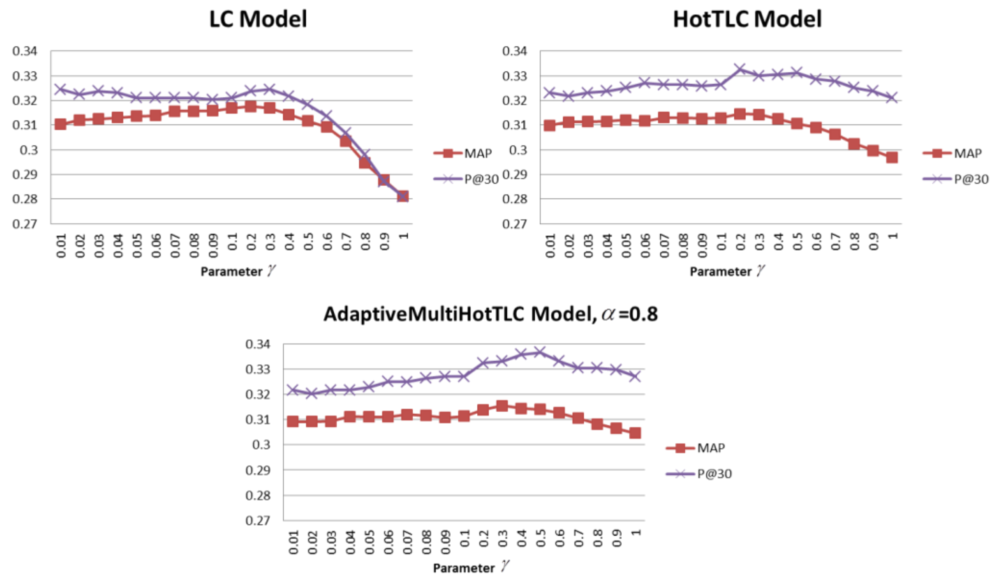

Fig.3 Performance of Models (LC, HotTLC, AdaptiveMultiHotTLC) when parameter γ changes

From Figure 3, we can see that almost every model reaches the highest point in the value of 0.1~0.5. When γ is in the value of 0.1~1, the performance of LC algorithm decline rapidly, with HotTLC and AdaptiveMultiHotTLC algorithms slowly. Hence, we can conclude that the models based hot time is parameter-insensitive algorithm.

For parameter α, its value range is 0.1~1 with 0.1 increase. The experimental results can be seen in Figure 4.

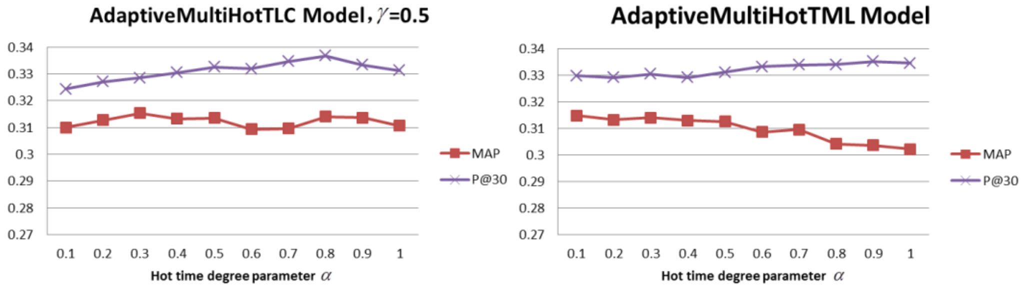

Fig.4 Performance of Models (AdaptiveMultiHotTLC, AdaptiveMultiHotTML) when parameter α changes

Figure 4 tells us that, with the changes of the parameter α, the curve relative changes smoothly. Concerned about the P@30 curve, the performance almost becomes better. The reason of this is relatively intuitive. When the value of α increasing, documents published in no-hot time are exclude so that to reduce the noise and improve the search performance. It is noticed that MAP curve change unlike our expection. In our opinion, the performance should increase when the value of parameter α adding and will achieve optimal when α is 0.8 or 0.9. But in fact, it doesn't. Possible reason caused for this different result is the degree of curve fitting. In detail, for AdaptiveMultiHotTLC algorithm with same γ for all topics, the different hot degree time will get same score which brings differed results for different topics. So the curve of the AdaptiveMultiHotTLC changes ups and downs. For AdaptiveMultiHotTML, it decease when α increase. Possible reason is that the MAP improves when α is 0.1, because the distribution of time score is almost fit the time distribution of pseudo documents, and at the same time the distribution of pseudo documents is almost same with the distribution of real relevant documents.

In summary, the rate parameter of exponential distribution γ has less affected MAP and P@30. The degree parameter for choosing hot time α impacts P@30 stable, with no stable for MAP.

5 Conclusion and Future Work

One of the core problems for microblog retrieval is how to incorporate time information into information retrieval. Based on Language Model, defining the create time as document's prior is a generally used method. There are two kinds of models Distinguished by whether it is related to queries. Both the models are based on the hypothesis that that the document which is published more recent is more important. However, by observing the queries, we find that the assumption is not suitable for every topic. Different queries have different hot time, which means that there are more relevant documents at that moment. If we assume that newer document has a higher prior score, this will harm relevant document's ranking position if more relevant docs are not published at the newest time. So we propose hot time language series model (HTLM series), which consider different topics with different number hot time. Finally, we define four models with two parameters.

As said above, there are two kinds of models, one is based on the whole collection, and the other is relevant with query. HTLM series belongs to the second model. The model only uses whole collection information can be seen as document's background message. The model using query information is document's specific message about specific query. So given a document d and a query q, the relationship between documents and time should be combined these two factor information. At last, we propose a mixed time language model by using smooth strategy and proved by experiment that mixed model is better than single model.

In the future, there are some points can be study, including: 1) exponential distribution is the most popular way to define the relationship between document and time, but whether it is the best distribution need more work; 2) tweet has many features that the webpages don't have, for example, hashtag. Our next work is trying to incorporate these tweet features with time information to define microblog prior. 3) as mentioned above, improving one time's importance will bring relevant documents but also some non-relevant documents. We will try to do something to decrease the non-relevant documents' existing.

References

1. Teevan, J., 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*2011, ACM: Hong Kong, China. p. 35-44.
2. Li, X. and W.B. Croft, *Time-based language models*, in *Proceedings of the twelfth international conference on Information and knowledge management*2003, ACM: New Orleans, LA, USA. p. 469-475.
3. Efron, M. and G. Golovchinsky, *Estimation methods for ranking recent information*, in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*2011, ACM: Beijing, China. p. 495-504.
4. Ponte, J.M. and W.B. Croft, *A language modeling approach to information retrieval*, in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*1998, ACM: Melbourne, Australia. p. 275-281.
5. Song, F. and W.B. Croft, *A general language model for information retrieval*, in *Proceedings of the eighth international conference on Information and knowledge management*1999, ACM: Kansas City, Missouri, United States. p. 316-321.
6. Zhai, C. and J. Lafferty, *Model-based feedback in the language modeling approach to information retrieval*, in *Proceedings of the tenth international conference on Information and knowledge management*2001, ACM: Atlanta, Georgia, USA. p. 403-410.