ATLANTIS PRESS

Advances in Computer Science Research, volume 58

Modeling, Simulation and Optimization Technologies and Applications (MSOTA 2016)

Micro-blog Query Expansion Based on Lexical Time Distribution

Zheng Xiao

Department of Computer Science and Technology, Chengdu Neusoft University, Chengdu 611844, China

***Abstract*—This paper presents a method of query expansion based on lexical time distribution for micro-blog search. Similarity of distribution is used to measure the correlation between the extended words and the query words. A query model based on the time distribution of the words is established. The query expansion method is unable to accurately estimate the expansion words due to the short micro-blog content. The real graph on the evaluation data is retrieved by the TREC2011 and TREC2012 micro-blog. The results show that the model can effectively improve the performance of micro-blog search extension vocabulary query based on time distribution, not only significantly outperforms the classical content based query expansion model, which is better than other time using the method of query expansion.**

Keyword-query expansion; query model; vocabulary time distribution; social network

I. INTRODUCTION

In all kinds of query expansion methods, content based query expansion is the basic idea of[1].Strategy is the main way of using relevance feedback, the first query the initial retrieval results show to the user, and then through the interaction with the user feedback and query the relevant documents (called feedback document), extraction of query expansion word formation the feedback from the document, more effective query, finally use the new query again to retrieve, in order to obtain better retrieval effect. However, in the micro-blog search in affected extended content based query: as the feedback document micro-blog is too short, and most of the words appear only a second, sparse data cannot provide reliable information to establish the relationship between the query and words expansion[2].

II. QUERY EXPANSION BASED ON TIME

In relevance feedback, because of the lack of a priori probability estimation of $P(d)$ related document information, $P(d)$ is often used with the same value. In the micro-blog environment, micro-blog at the time of the distribution is not uniform[3], so the researchers use this feature to estimate $P(d)$, which was better query expansion words.

A. Recency Based Query Expansion

Based on recency query expansion, the representative work of Li and Croft[4] is proposed based on the correlation model of recency. The method that the new document is more likely to be the relevant documents, the time information they use document release, with exponential distribution a priori probability estimation of the document, issued from time to time closer to the query the document for greater prior probability:

$$P(d) = P(d|t_d) = \lambda e^{-\lambda|t_Q - t_d|} (1)$$

Among them, t_Q is the time when the user submits the query; t_d is the time of the document release; λ is the parameter of the exponential distribution.

By re estimating the prior probabilities of the related models, the method defines the query expansion model

$$P(\omega|\theta_{QE}) \propto \sum_{\theta_d \in R} P(\omega, Q|\theta_d)\lambda\, e^{-\lambda|t_Q - t_d|} (2)$$

Where θ_d is the document language model d. Based on recency query expansion, those in the "new" words in the document will have a higher probability of becoming expanded words.

B. Query Expansion Based on Explosive

The expansion method cannot query time or from the outbreak period with multiple outbreak of the query processing well based on recency query [5]. Therefore, researchers have proposed explosive release time using micro-blog to query expansion by Keikha frame time factor into the model, put forward the model based on the time. Methods the query generation process as a sampling process, first with the probability $P(t|Q)$ a time, then with probability $P(\omega|t, Q)$ from the sample selection word related documents this time in ω , $P(\omega|t, Q)$ after some transformation hypothesis is equivalent to the word in the document on the distribution of $P(\omega|d)$ and, as the final model (3) shown.

$$P(\omega|\theta_{QE}) \propto \sum_t P(t|Q) \sum_{d \in R_t} P(\omega|d) (3)$$

However, in the micro-blog search environment, as the feedback document micro-blog short, and most of the words appear only once, the document model $P(\omega|\theta_D)$ estimation is not sufficient, unable to provide reliable information [10] for the distribution of query expansion. Only depends on the time of $P(d)$ adjustment, can not solve the lack of feedback document model $P(\omega|\theta_D)$'s own estimates fundamentally.Therefore, this paper attempts to model $P(\omega|\theta_D)$ does not depend on the feedback documents, only through the time characteristic of word query relationships between words and words expansion, query expansion.

III. TEMPORAL DISTRIBUTION OF WORDS AND QUERIES

From the above analysis, it can be seen that the frequency of the word is varied from different time periods, from the angle of the discrete probability distribution.$P(T|\omega)$can be used to describe the time distribution of words,and to describe the convenience of the use of symbols to represent, whileT is a random variable that represents the time period. For a specific time period t_i, the use of $P(T = t_i|\omega).P(t_i|\omega)$ to represent the $P(t_i|\omega)$ description of the people in the time period t_i use the word. How frequently reflects the degree of popular words in the period of t_i. The t_i is in accordance with a fixed interval partitioning time fragment, namely micro-blog set of discrete time can be expressed as $\{t_1, t_2, \ldots\ldots\ldots, t_n\}$, t_i represents a specific time slice (taking natural day (24h) is a single bit). In this paper, the following content, no special instructions, the time in a day as time units are analyzed and discussed on the direct estimation of the temporal distribution of the words $P(t_i|\omega)$ is more difficult, according to Bias theorem:

$$P(t_i|\omega) = \frac{P(\omega|t_i)P(t_i)}{P(\omega)} (4)$$

Assuming that the time period is independent of each other and the prior probability $P(t)$ of each time interval is the same, the whole probability formula:

$$P(\omega) = \sum_{t_i \in T} P(\omega|t_i)P(t_i) (5)$$

Then apparently

$$P(t_i|\omega) = \frac{P(\omega|t_i)}{\sum_{t_j \in T} P(\omega|t_j)} (6)$$

In this paper, the language model all micro-blog each time period established called piecewise language model for time, expressed in θ_t. A language model is estimated by maximum likelihood estimation. The $P(\omega|t_i)$ can be estimated.

$$P(\omega|t_i) = P\left(\omega|\theta_{t_i}\right) = \frac{c(\omega,t_i)}{\sum_{\omega' \in V} c(\omega',t_i)} (7)$$

Where V is theV in the list; the word $c(\omega, t_i)$ is the total number of all micro-blog ω released at the time of the t_i.

According to the formula (10) and (11), the time distribution of the words can be estimated by the formula (8).

$$P(t_i|\omega) = \frac{P(\omega|\theta_{t_i})}{\sum_{t_j \in T} P(\omega|\theta_{t_j})} (8)$$

Words reflect the time distribution in the popular words, different time similar, we use the time distribution of the query to reflect this group query of the popularity of different time periods, denoted by $P(t_i|Q)$ and $P(t_i|\omega)$. This is the estimation method as shown in (9):

$$P(t_i|Q) = \frac{P(Q|t_i)P(t_i)}{P(Q)} = \frac{P\left(Q\middle|\theta_{t_i}\right)}{\sum_{t_j \in T} P\left(Q\middle|\theta_{t_j}\right)} = \frac{\prod_k P(q_k|\theta_{t_i})}{\sum_{t_j \in T} \prod_k P(q_k|\theta_{t_j})} (9)$$

We think that this similarity reflects the degree of correlation of words from a certain extent, can use a similar degree of correlation the time distribution to quantifiers. Considering the common KL distance measure between two distributions are asymmetric. The similarity isωand q_i time distribution of $P(T|\omega)$ and $P(T|q_i)$ can be calculated as.

$$S\left(P(T|\omega), P(T|q_i)\right) = \sum_{t_i \in T} |P(t_i|\omega) - P(t_i|q_i)| (10)$$

In the unit time, the formula (14) is equivalent to

$$S\left(P(T|\omega), P(T|q_i)\right) = \sum_{t_i} |P(t_i|\omega) - P(t_i|q_i)| \Delta t (11)$$

Where t_i represents a time period, Δt is a unit of time (in the micro-blog environment, usually with a natural day (24h) as a unit[7]).

Type (12) of the physical meaning of the expansion of the word and the time distribution of the query is the size of the.$S\left(P(T|\omega), P(T|q_i)\right)$value of the smaller the smaller the size of the two, which means that theωand q_i aremore similar.

As shown in Figure 1, two related words surrounded by the graphics area was less than two unrelated words surrounded by the graphics area(12) for Figure 2. The distribution of the correlation similarity estimation based on extended time vocabulary wordsωand q_i queries are given below:

$$rel\left(P(T|\omega), P(T|q_i)\right) = N(S\left(P(T|\omega), P(T|q_i)\right)) (12)$$

Where q_i is any Q in a word query; ωis the feedback of words in the document; $N(x)$is the normalized function, and $N(x) = \frac{(max-x)}{(max-min)}$, max and min respectively to the sample data of the maximum and minimum values. Because surrounded the two distribution area is less than or equal to 2 is greater than or equal to 0, so the max is set to 2 , min set to 0. $N(x)$on the x of linear transformation, converting x to [0,1] between the value and the time distribution of two words in the area surrounded by two words and the correlation is inversely proportional to the. 2 words of the time distribution of the area surrounding the small S the correlation between the two..

Similarly, the definition of the extension of the wordωand query Q is related to the degree of

$$rel\left(P(T|\omega), P(T|Q)\right) = N\left(S\left(P(T|\omega), P(T|Q)\right)\right) (13)$$

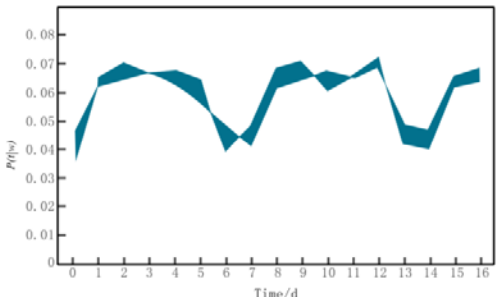

FIGURE I. THE TIME DISTRIBUTION OF THE QUERY WORDS AND RELATED WORDS IS A GRAPH

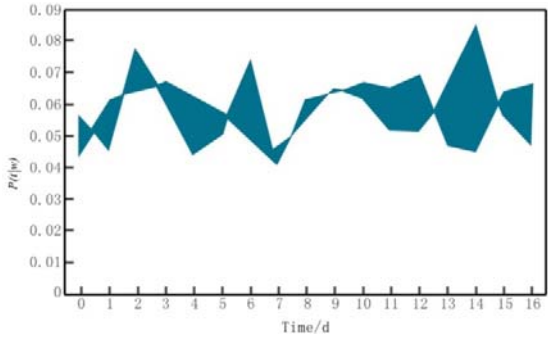

FIGURE II. THE DIFFERENCE OF THE TIME DISTRIBUTION OF THE QUERY WORDS AND THE UNRELATED WORDS IN THE GRAPH

IV. QUERY EXPANSION BASED ON LEXICAL TIME DISTRIBUTION

According to the above analysis, the Related words and query frequency in the different period of time with the same increase with the decrease trend. The time distribution of similarity measure can be extended and query (or queries). This section gives the relative degree of query expansion based on the distribution of time words, referred to as TTDM (Term Time Distribution Model).

Similar to other query expansion methods, TTDM combines the query expansion model with the original query to obtain the final query model, as shown in the formula (14):

$$\mathrm{P}\left(\omega\middle|\theta_Q\right) = (1-\lambda)P_{ml}(\omega|Q) + \lambda P(\omega|\theta_T) \quad (14)$$

Where θ_T is the extended model of vocabulary query based on time distribution; ω distribution is similar for the original query query and application of words with other words. Time extended query expansion method, ω is selected from the estimated feedback documents related to the highest degree of n.

θ_T are as follows:

$$\mathrm{P}(\omega|\theta_T) = \frac{\mathrm{score}(\omega,\mathrm{Q})}{\sum_{\omega' \in V} \mathrm{score}(\omega',\mathrm{Q})} \quad (15)$$

Among them, $\mathrm{score}(\omega, \mathrm{Q})$ is the correlation degree between the extended word and the query; ω is the extension; V is the set of query expansion words..

According to the degree of correlation between the word and the query for query expansion mainly includes two strategies: one strategy and one to many so-called strategies[8]. One strategy refers to query generation and sorting word dependent relationship in alternative extending words in the query, and one to many the method of query expansion, query expansion generation and sorting words depends on the candidate expansion terms and query the whole relationship. According to the two kinds of strategies, this paper defines two kinds of calculation ω and query expansion word relevance scoring method Q, denoted as $\mathrm{score}(\omega, \mathrm{Q})$, select expansion terms and estimation of query expansion model using the score.

The second approach uses a pair of multiple strategies to compute the temporal correlation of the extended word and the query. The application of the formula (16) defines the $\mathrm{score}(\omega, \mathrm{Q})$ as

$$\mathrm{score}(\omega, \mathrm{Q}) = \mathrm{rel}(\mathrm{P}(\mathrm{T}|\omega), \mathrm{P}(\mathrm{T}|\mathrm{Q})) \quad (16)$$

Among them, $\mathrm{P}(\mathrm{T}|\omega)$ is the time distribution of wordω, which is estimated by the formula (12); $\mathrm{P}(\mathrm{T}|\mathrm{Q})$ is the time distribution of the whole query, which is estimated by the formula (13).

V. RESULTS AND ANALYSIS

A. Experimental Result

Table 1 shows the results of our method and the baseline method described on Tweets2011 data sets, our method, using a single query words related to query expansion strategy marked TTDM-q, with the overall query related query expansion strategy marked TTDM-Q. are listed in Table 1 for our method and the baseline method in the evaluation index on the P@30 value and various methods with respect to language model improvement. The significance by Wilcox on test ($p < 0.05$). "&" query model in this paper is better than LM, "*" that is better than RM, "+" said was superior to that of RBRM, "said was superior to that of BBRM.

TABLE I. COMPARISON OF MODEL RETRIEVAL PERFORMANCE

	TREC 2011		TREC 2012	
	P@30	P@20	P@30	P@20
LM	0.4136	0.4490	0.3350	0.3610
	1	1	1	1
RM	0.4503	0.4939	0.3661	0.4017
	（+8.9%）	（+10%）	（+9.3%）	（+11.3%）
RBRM	0.4714	0.5204	0.3836	0.4297
	（+14.0%）	（+15.9%）	（+14.5%）	（+19.0%）
BBRM	0.4510	0.5041	0.3757	0.4068
	（+9.0%）	(+12.3%)	(+12.1%)	(+12.7%)
TTDM-Q	0.4741&*	0.5133&	0.4096&*+#	0.4475&*#
	（+14.6%）	（+14.3%）	（+22.3%）	（+24.0%）
TTDM-q	（+15.3%）	（+18%）	（+25.0%）	（+24.7%）

With consistent above, with a single query words related to query expansion strategy is superior to the TTDM-q and query the whole query expansion strategy of TTDM-Q. in TTDM-q is better than the TTDM-Q method, the detailed analysis of the following only for TTDM-q.

B. Effect of Query Expansion on Retrieval Performance

In query expansion will frequently query expansion model and the original query model based on query model to estimate the final, its weight was regulated by the λ (see (3)). This section of the value ofλanalysis was carried out on the retrieval performance, the results are shown in Figure 3.

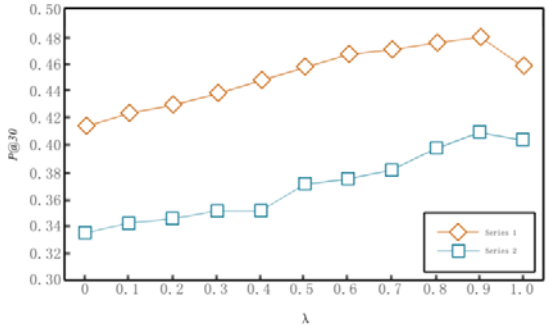

FIGURE III. EFFECT OF QUERY EXPANSION ON RETRIEVAL PERFORMANCE

It can be seen from Figure 3, with the gradual increase of the weight of theλquery expansion model, steadily improve the retrieval performance, when the weight of query expansion model is 0.9 best retrieval performance, the weight of the original query is only 0.1.

The experimental results also show that using only the original query ($\lambda = 0$) is difficult to achieve satisfactory results, and only using the query expansion model ($\lambda = 1$) to replace the original query to a certain extent, to complete the retrieval task. However, an effective query model should be combined with the original query and query expansion, and the reasonable distribution of weight two the.

C. *The Effect and Analysis of the Number of Extended Words on TTDM*

And other query expansion method, this method also has two important parameters: the number of query expansion and feedback document number. The remainder of this paper will analyze the influence of the two parameters of TTDM-q.

The number of expansion terms determines the number of expansion terms to expand the query. Figure 4 shows the effect of model query expansion words retrieval performance in the different number of documents under the condition.

It can be seen from Figure 5, the different couple back documents, along with the increasing number of expansion, the performance of TTDM-q increased gradually and tends to be stable, and the content of the query expansion method based on large differences exist: query expansion words content query expansion usually use fewer based (e.g. Ref. 14 in micro-blog TREC the data set to 20).

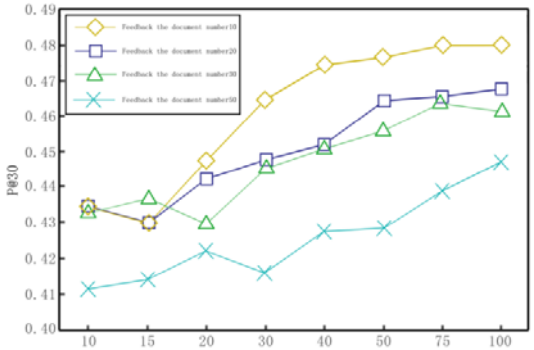

FIGURE IV. TERC2011 QUERY1-49

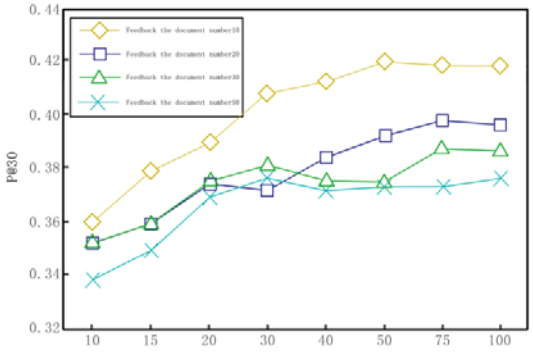

FIGURE V. TERC2012 QUERY 51-110

In addition, the time of mining the relation between words is not limited to the expansion of the application in the query, to study the relationship between mining and other words the word also has a certain reference value.

REFERENCES

[1] Bing Li. A New Query Expansion Method for Statements Query Retrieval Based on Event Structure[A]. Information Engineering Research Institute, USA.Proceedings of 2013 2nd International Symposium on Computer,Communication,Control and Automation(3CA 2013)[C].Information Engineering Research Institute, USA:,2013:4.

[2] Shafiq Ahmad Khan,M.M.Qurashi. Historical Variations in the Specialized Subjects of the Elected Fellows of the Pakistan Academy of Sciences[A]. Pakistan Academy of Sciences.Proceedings of the 4th Session of 2011 Workshop of Pakistan Academy of Sciences[C].Pakistan Academy of Sciences:,2011:10.

[3] Efron M, Lin J, He J, et al. Temporal feedback for tweet search with non-parametric density estimation[C]// International ACM SIGIR Conference on Research & Development in Information Retrieval. ACM, 2014:33-42.

[4] Li X, Croft W B. Time-based language models[C]// Twelfth International Conference on Information & Knowledge Management. 2003:469-475.

[5] Zhai C, Lafferty J. A study of smoothing methods for language models applied to Ad Hoc information retrieval[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2004:179-214.

[6] Keikha M, Gerani S, Crestani F. Time-based relevance models[C]// Proceeding of the, International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July. 2011:1087-1088.

[7] Duan L, Guo W, Zhu X, et al. Constructing Spatio-Temporal Topic Model for Microblog Topic Retrieving[J]. Geomatics & Information Science of Wuhan University, 2014.

[8] Peetz M H, Meij E, De Rijke M, et al. Adaptive temporal query modeling[M]// Advances in Information Retrieval. Springer Berlin Heidelberg, 2012:455-458.

[9] Casanovas J M MM. A New Minkowski Distance Based on Induced Aggregation Operators[J]. International Journal of Computational Intelligence Systems, 2012, April 2011(2):123-133.

[10] Carpineto C, Romano G. A Survey of Automatic Query Expansion in Information Retrieval[J]. Acm Computing Surveys, 2012, 44(1):159-170.