International Journal of Database Theory and Application
Vol.8, No.6 (2015), pp.89-100
http://dx.doi.org/10.14257/ijdta.2015.8.6.08

A Hyperlink-Extended Language Model for Microblog Retrieval

Zhongyuan Han[1, 2], Muyun Yang[1], Leilei Kong[2], Haoliang Qi[2*] and Sheng Li[1]

1 School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China;
2 School of Computer Science and Technology, Heilongjiang Institute of Technology, Harbin, China;
** Corresponding Author haoliangqi163@163.com*

Abstract

Microblog retrieval has received much attention in recent years. In microblog retrieval, the content linked by URLs is one of the most important information of a microblog. We present a Hyperlink-extended model for microblog retrieval that combines content of microblogs and the content of embedded hyperlinks webpages using a probabilistic ranking function based on language model. Hyperlink-extended language model incorporates the users' information retrieval requirements and the microblog author's expression needs. Using standard TREC 2011 and TREC 2012 microblog retrieval collection, various aspects of our microblog retrieval model are evaluated. Results show our model significantly outperform the art-of-the-state URL-based approaches and the best performance of TREC 2012 microblog retrieval.

Keywords: *Information Retrieval, Language Model, Microblog Retrieval, Link*

1. Introduction

Microblog is one of the most powerful online communications media for people to learn what is happening around the world today. People post microblogs to broadcast messages in real time and express their views by using limited words quickly and freely. On the other side, more and more people search information via microblog retrieval because microblogs can provide users with real-time information not yet available in the mainstream media [1]. The study finds that Twitter acts not only as a social network, but also as a news source [2].The third-party search tools, including Google [3] and Bing [4], provide the microblog retrieval to access to tweets.

For microblog retrieval users, microblogs act as two media: social network and news media [2]. They retrieve the microblogs to get two kinds of information: information about some important events, comments, feelings or opinions from other people. Due to the limitation of texts length of microblog, the news media function is realized not only by the information the microblog posted, but also by the URL linking to other websites for the detailed information. From the authors' standpoint, the intention behind sharing a link is to post some potentially interesting information available somewhere else on the web, so the presence of a links can be seen as an indication of information [5]. Therefore, the document linked by URL is an important supplement to the microblog. Even in some cases, it is just what the author wants to share. For example, a tweet user comments on an essay, and he gives a links of this essay, what he wants to share is the content of hyperlink document.

In the existing microblog retrieval researches, in general, hyperlink is applied in two ways. In one case, the parts of hyperlink are used to expand microblog or query, for example, the contents of the title tag or the meta descriptor tags of the document to the

ISSN: 2005-4270 IJDTA

tweet itself are used to expand the microblog [6-10]. In the other case, the hyperlink is used as the features of learning to Rank algorithms. For example, in [11], whether the microblog contains the links or not is regarded as the most important feature for a microblog except the text content. The above two kinds of methods demonstrate the effectiveness of using links information in microblog retrieval research, but do not make the retrieval performance significantly increased. The reason lies that the links information is not fully utilized and the information of the linked document is discarded.

For making full use of linked document information, a retrieval model should be integrated with more hyperlink documents information. This will not only conform to the authors' expression goal, but also live up to the users' retrieval needs. Therefore, this paper proposes a new microblog retrieval model in language model framework, called *Hyperlink-extended Language Model*, to combine the content of microblogs with the whole content of URL linked webpages. By using Hyperlink-extended Language Model, the author's expression goals and the user's demands for information retrieval is integrated in one model. In this model, the user's retrieval demands expressed by their submitted queries are modeled as retrieval needs on the original microblog and that on the linked content, and the document is modeled via combining the microblog itself and hyperlink document.

Using standard corpus of TREC 2012 microblog real-time retrieval, we evaluate various aspects of our microblog retrieval Model. The results show the new model is significantly outperform the art-of-the-state URL-based approaches.

The rest of this paper is organized as follows. Section 2 presents the related work. Section 3 describes Hyperlink-extended Language Model in detail. Section 4 reports the experiments and performance comparisons. Discussion goes in Section 5, and the last section concludes our study.

2. Related Work

Tweet search introduces a new problem for information retrieval. The recent researches show that link information can improve the performance of microblog retrieval. Hyperlink information has been implemented in different approaches. This section reviews the links application researches in detail firstly. Then, since our work is carried out using language modeling, we will review the related studies in this framework in more detail.

2.1. Applying Hyperlinks in Microblog Search

El-Ganainy *et al*. propose a hyperlink-extended pseudo-relevance feedback method to expand queries by using titles and meta-description of embedded hyperlinks in retrieved microblogs [6]. Liang Feng enriches the microblog model by utilizing the webpage titles of embedded hyperlink documents [7]. The similar document expansion methods via the title tags, the contents of the keywords and description type meta tags of the linked documents are reported in [8-10]. Van Duc *et al*. investigation shows that in most of the cases the linked page titles appeared to bring the essence of the page's content in well-chosen terms [12].

Links information is also used as a ranking feature in many researches. El-Ganainy takes Has_URL and URL as the features in ranking models, in which Has_URL is a binary feature which is assigned 1 if the tweet contains at least one URL, and 0 otherwise, and URL indicates the number of URLs included in the tweet [13]. McCreadie *et al*. investigated how content of embedded hyperlinks in tweets could help to estimate the tweet's relevance [8]. They show that using hyperlink documents in tweets can improve retrieval effectiveness either by document expansion or learning to rank algorithm. The best method of TREC 2011 microblog task shows that the most important non-textual feature is a binary feature which is assigned 1 if the tweet contains at least one URL, and

0 otherwise [11]. In the TREC 2012 and 2013 microblog tasks, the best methods without manual influence show that the similarity score between the query and the linked document is a key feature [14-15].

Although these approaches are highly advanced and usually led to improvements, only part of hyperlink information in the microblog is utilized. We believe that the full information of the linked document can improve the performance significantly. Followed this idea, we get the best performance in TREC 2012 microblog real-time retrieval task. In this paper, the whole hyperlinks information is utilized from the point of view of both the authors of the microblog and the retrieval users, and then Hyperlink-extended Microblog Retrieval model is proposed.

2.2. Language Model

In the classic language model framework, queries and documents are modeled as query model θ_Q and document model θ_D. Specifically, the Kullback-Leibler divergence [16] is used to measure the difference between θ_Q and θ_D as follows:

$$KL(\theta_Q, \theta_D) = \sum_{w \subset V} P(w \mid \theta_Q) \log \frac{P(w \mid \theta_Q)}{P(w \mid \theta_D)} \quad (1)$$

Where V is the vocabulary, w is the word in V. $P(w|\theta_Q)$ and $P(w|\theta_D)$ are the word w's distribution in query model θ_Q and document model θ_D. θ_Q and θ_D are usually constructed as the unigram language model[17]. $P(w|\theta_D)$ is estimated by maximum likelihood estimate and smoothing technology, such as Dirichlet smooth, which is used to address the zero probability problem[18]. By applying Dirichlet smoothing method, $P(w|\theta_D)$ can be described as:

$$P(w \mid \theta_D) = \frac{c(w, D) + \mu P(w \mid C)}{|D| + \mu} \quad (2)$$

Where the definition of c(w,D) is the term frequency of w in document D. P(w|C) is the probability of w seen in the corpus C. |D| is the number of words in the document.

Query model θ_Q is regularly estimated by maximum likelihood estimate method as follows:

$$P(w \mid \theta_Q) = \frac{c(w, Q)}{|Q|} \quad (3)$$

Where c (w,Q) is the term frequency of w in query Q, and |Q| is the total number of words in query Q.

3. Hyperlink-Extended Language Model for Microblog Retrieval

3.1. Why Hyperlinks are Needed

Hyperlinks are the important features of microblogs. The motivations for making use of the content of hyperlink documents in microblog retrieval model to rank microblogs reflect in three aspects.

Firstly, the short length of microblogs limits the expression of full meaning, so the author tend to supplement the microblog content by using links. Indeed, since the amount of space in a tweet is very limited, the links is believed indispensable for the author to express his/her intention completely.

Secondly, recent researches show that microblogs with links can enhance the perceptions of the credibility judgment and the consumers oriented microblogs think that containing a URL leads to the higher quality [1]. So microblogs with hyperlinks have higher probability to be relevant and they are more suitable for users' retrieval needs.

Thirdly, users' retrieval intention of microblog is different from that of web. Teevan's research shows that the topically motivated searches on Twitter appear to contain themes related to timely and social information [19]. What the user need may be the timely and social information itself, that is, the links posted by the author, not the author's opinions and feelings. They merely resort to microblog retrieval to gain this information. Our statistic on the corpus of TREC2011 and TREC2012 microblog retrieval also shows that 95.1% and 87.5% highly relevant microblogs contain hyperlinks respectively.

Therefore, no matter from the author's point of view or from the user's point of view, the hyperlink documents play an important role in microblog retrieval: not only to express the information that the authors want to convey, but also to meet the users' information needs for information retrieval.

To meet both aspects of the above information needs in retrieval models, we propose the *Hyperlink-extended Microblog Retrieval Model*. Section 3.2 briefly presents the Hyperlink-extended Microblog Retrieval Framework which describes how to integrate the information contained by microblog and the information linked by URL. Section 3.3 proposes hyperlink-extended microblog retrieval model. Section 3.4 provides the details of estimation and smoothing.

3.2. Hyperlink-Extended Microblog Retrieval Framework

Based on the analysis of 3.1, Figure 1 shows the links-based microblog retrieval framework:

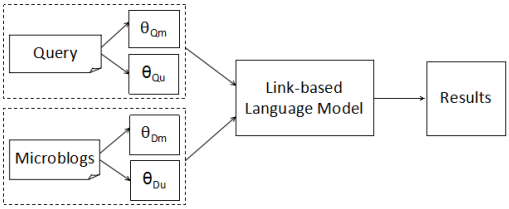


Figure 1. Hyperlink-Extended Microblog Retrieval Framework

In the microblog poster's eyes, to express the comprehensive information within length limitation, he will divide the microblog into two parts: the microblog's content and the supplementary information which expressed in the form of URL link.

From the microblog retrieval user's point of view, when searching the microblogs, he wants to get the information related to a specific topic expressed by the query he submits to the search engine. The information needs can be satisfied by microblog body itself or the webpages linked by the URL. In this way, the user's information needs can also be divided into two parts: the information needs get from microblog text and the information needs obtained from the URL linked webpages.

The user's standpoint can also be understood that some users might expect to search the microblog text relevant to his query and other users might expect to find webpages linked by the URL. For example, for the query "BBC World Service staff cuts", maybe the user wants to know people's views and comments on this event, or maybe he wants to search the news about BBC World Service staff cuts.

The next section presents a microblog retrieval model which incorporates a hyperlinked document into the ranking process for microblog retrieval.

3.2. Hyperlink-extended Microblog Retrieval Model

This section describes the *Hyperlink-extended Microblog Retrieval Model* which incorporates the original microblogs information and hyperlink documents.

The model is built from the two aspects: the author's information expression needs and user's information retrieval needs. Where D_m is microblog's text, D_u is the text of webpages linked by URL in the microblog D, Q_m is the user's information needs which might be satisfied by microblog's text, and Q_u is the user's information needs which are more likely met by the hyperlink documents. Obviously, Q_m and Q_u are independent and $P(Q_m|Q)+P(Q_u|Q)=1$. Then we define:

$$\mathrm{P}(D \mid Q) \propto P(D_m, D_u \mid Q_m, Q_u) \tag{4}$$

Supposing that D_m and D_u are independent, then we get:

$$P(D_m, D_u \mid Q_m, Q_u) = P(D_m \mid Q_m, Q_u) P(D_u \mid Q_m, Q_u) \tag{5}$$

Since Q_m represents the query that the user wants to search in microblogs text, Q_m is irrelevant to D_u. Likewise, Q_u is irrelevant to D_m. That is:

$$P(D_m \mid Q_m, Q_u) = P(D_m \mid Q_m) \tag{6}$$

$$P(D_{\mathrm{u}} \mid Q_m, Q_u) = P(D_{\mathrm{u}} \mid Q_{\mathrm{u}}) \tag{7}$$

When substitute equation (5), (6) and (7) into (4), we define the relevance between query Q and microblog D as:

$$R(D \mid Q) \propto P(D_{\mathrm{m}} \mid Q_m) \mathrm{P}(D_{\mathrm{u}} \mid Q_u) \tag{8}$$

In the next section, we will describe a way of estimating the probability distribution $P(D_m|Q_m)$ and $P(D_u|Q_u)$.

3.3. Final Estimation Details

According to Bayes' theorem, $P(D_m|Q_m)$ can be represented by Eq. (9):

$$\mathrm{P}(D_m \mid Q_{\mathrm{m}}) = \frac{P(Q_{\mathrm{m}} \mid \mathrm{D}_m) P(D_m)}{P(Q_{\mathrm{m}})} \tag{9}$$

$P(Q_m)$ is irrelevant to D_m, so it doesn't affect the document ranking. $P(D_m)$ is the prior probability of document D_m. In general, $P(D_m)$ is set as uniform. Therefore, the $P(D_m|Q_m)$ is equivalent to:

$$\mathrm{P}(D_m \mid Q_{\mathrm{m}}) \propto P(Q_{\mathrm{m}} \mid D_m) \tag{10}$$

Similarly, we can get the equation of $P(D_u|Q_u)$:

$$\mathrm{P}(D_u \mid Q_{\mathrm{u}}) = \frac{P(Q \mid D_u) \mathrm{P}(D_u)}{\mathrm{P}(Q)} \propto P(Q_{\mathrm{u}} \mid D_u) \mathrm{P}(D_u) \tag{11}$$

Note that the probabilities of D_u are different since some microblogs contain links but others not.

When we substitute equation (10) and (11) into equation (8) and apply Maximum Likelihood Estimation, we get the following final estimation for P(D|Q) :

$$\mathrm{P}(D \mid Q) \propto \mathrm{P}(Q_m \mid D_m) \times P(Q_u \mid D_u) \times P(D_u) = \prod_{w \subset Q} (w \mid D_m)^{\mathrm{c}(\mathrm{w}, \mathrm{Q_m})} \times \prod_{w \subset Q} (w \mid D_{\mathrm{u}})^{\mathrm{c}(\mathrm{w}, \mathrm{Q_u})} \times P(D_u) \tag{12}$$

Applying logarithmic function to (12), then we obtain:

$$\mathrm{P}(D \mid Q) \propto \sum_{w \subset Q} \mathrm{c}(\mathrm{w}, \mathrm{Q_m}) \ln(w \mid D_m) + \sum_{w \subset Q} \mathrm{c}(\mathrm{w}, \mathrm{Q_u}) \ln(w \mid D_u) + \ln P(D_u) \tag{13}$$

Transform the first item in (3-10) into the following forms:

$$\sum_{w\subset Q} c(w,Q_m)\ln(w|D_m) = |Q_m| \sum_{w\subset Q} \frac{c(w,Q_m)}{|Q_m|}\ln(w|D_m) \quad (14)$$

$$= |Q_m| \sum_{w\subset Q} P(w|Q_m)\ln(w|D_m)$$

$$= |Q_m| [-\sum_{w\subset Q} P(w|Q_m)\frac{\ln(w|Q_m)}{\ln(w|D_m)} - \sum_{w\subset Q} P(w|Q_m)\ln(w|Q_m)]$$

$$\propto |Q_m| [-\sum_{w\subset Q} P(w|Q_m)\frac{\ln(w|Q_m)}{\ln(w|D_m)}]$$

In(14), $|Q_m|$ is the length of query Q_m, the right item is the negative KL-divergence. So, (14) can described as follows:

$$\sum_{w\subset Q} c(w,Q_m)\ln(w|D_m) = -|Q_m| KL(\theta_{Q_m},\theta_{D_m}) \quad (15)$$

The similar approach can be used for estimating Eq. (16):

$$\sum_{w\subset Q} c(w,Q_u)\ln(w|D_u) = -|Q_u| KL(\theta_{Q_u},\theta_{D_u}) \quad (16)$$

In our model, the query is represented by Q_m and Q_u. We suppose that $P(w|Q_m)$ and $P(w|Q_u)$ are identical distribution with P(w|Q) in the case of the absence of prior knowledge of Q. We define:

$$|Q_m| = P(Q_m|Q)|Q|, \ |Q_u| = P(Q_u|Q)|Q| \quad (17)$$

Where |Q| is the length of query Q. $P(Q_m|Q)$ and $P(Q_u|Q)$ is the probability of Q generates Q_m and Q_u respectively.

According to(13), (15), (16) and (17), we set P(D|Q) to be:

$$P(D|Q) \propto -P(Q_m|Q)KL(\theta_{Q_m},\theta_{D_m}) - P(Q_u|Q)KL(\theta_{Q_u},\theta_{D_u}) + \frac{\ln P(D_u)}{|Q|} \quad (18)$$

In (18), the first part is the product of $P(Q_m|Q)$ and KL divergence of microblog text D_m and query Q_m, the second is the product of $P(Q_u|Q)$ and KL divergence of linked document text D_u and query Q_u, and the last part can be understand to represent a confidence levels of hyperlinked document. There are two ways to estimate $P(D_u)$: if the microblog D contains links and we can get the linked pages content, we set $P(D_u)=1$; otherwise, we give it a very small value.

4. Experiments

4.1. Datasets

In 2011 and 2012, the Text REtrieval Conference (TREC) ran the Microblog real-time track that investigated adhoc tweet search [20, 21]. The aim of this task was to find the most relevant tweets for the user query in a real-time setting. To facilitate this track, the first legally redistributable Twitter test collection, named TWEETS11, was developed through the collaboration between TREC and twitter. We evaluate our method with the corpus of TREC 2012 real-time adhoc task on microblog track, called TWEETS11 corpus [20].

The Microblog track examines search tasks and evaluation methodologies for information seeking behaviors in microblogging environments such as Twitter. The TWEETS11 corpus consists of an approximately 1% sample of tweets from January 23, 2011 to February 7, 2011 (inclusive), totaling approximately 16 million tweets. Major events that took place within this time frame include the massive democracy demonstrations in Egypt as well as the Super Bowl in the United States [20]. Different types of tweets were presented, including replies and retweets. The corpus was designed

to be a reusable, representative sample of the twittersphere, for example, both important and spam tweets are included.

Using twitter crawler provided by track organizers, 10,397,336 tweets are downloaded.

Moreover, we extracted URL links in tweets and downloaded them as an external resource. The number of successfully downloaded webpages is 814,817.

As TREC requested, the corpus is processed as follows:

1) The null tweets were removed.
2) Deal with the forwarding microblog. For the forwarding microblog with "RT", if there is no content in front of RT, then delete the microblog. Otherwise, keep the content in front of RT and regard it as the useful information which is added by the forwarding user. The information behind the RT is deleted.
3) Filter out all the non-English tweets using language identifier tool provided by Nutch[1].
4) Porter stemmer is used for stemming and stop words are filtered by Indri toolkit[2].

After preprocessing, the effective microblogs number is 3,754,077.

Note that the future source of evidence is forbidden: information that would not have been available to the system at the timestamp of the query. For example, IDF scores computed using tweets not already posted at the timestamp of the query [20]. So we completed the real-time index in our experiments. Actually, we built a separate index for every query [22]. Table 1 shows the statistics on the 3,754,077 microblogs.

Table 1. Statistics on TWEETS11

		2011 topics		2012 topics	
Description	Microblog Corpus	Relevant Microblog	High Relevant Microblog	Relevant Microblog	High Relevant Microblog
Number of Microblog	3,754,077	2,821	529	6,059	2,468
Average Links Rate	26.5%	81.6%	95.1%	77.1%	87.5%

4.2. Performance Measures

Referring to TREC real-time microblog retrieval task, we use P@30 and MAP as the evaluation measures.

Precision@k is a measure for evaluating top k positions of a ranked list using two levels (relevant and irrelevant) of relevance judgment. P@30 gives a simple measure of search effectiveness on the top 30 search result page. Precision@k is defined as:

$$\mathrm{P@k} = \frac{1}{\mathrm{k}} \sum_{j=1}^{k} r_j \tag{19}$$

Where k denotes the truncation position, r_j equals one if the document in the jth position is relevant and zero otherwise.

Mean average precision (MAP) is one of the most widely-used measures for information retrieval. MAP is the mean of average precision (AP) scores over a set of queries. Average precision for a single query is calculated by taking the mean of the precision scores obtained after each relevant document is retrieved:

$$\mathrm{AP} = \frac{1}{\mathrm{RN}} \sum_{\mathrm{k}=1}^{\mathrm{n}} (\mathrm{P}(k) \times rel(k)) \tag{20}$$

[1] http://nutch.apache.org/
[2] http://www.lemurproject.org/indri.php

Where RN is the number of relevant documents, n is the number of retrieval document. P(k) is the precision at k. The rel(k) is 1 if the document at rank k is relevant, otherwise, 0. Note that the relevant documents that are not retrieved receiving a precision score of zero.

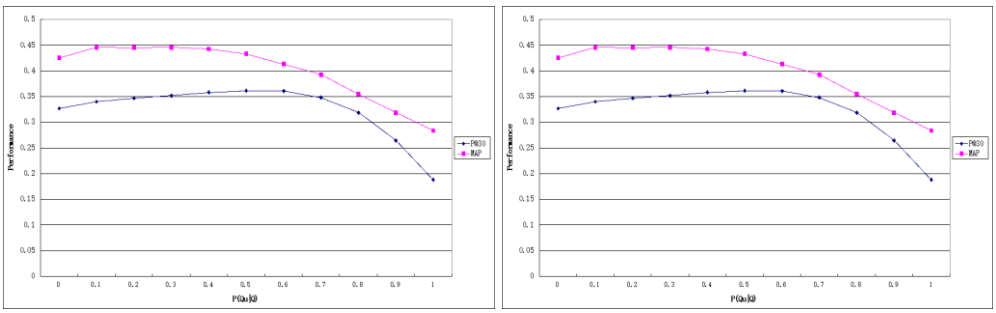
4.3. Experimental Results

4.3.1. Experimental Results on TREC 2012 Microblog Test Sets: In our experiments, the language model is selected as the baseline. The *Mixture Retrieval Model of Document Expansion, Query Expansion and URL* (MRM) [15], which is the best run in TREC 2012 microblog retrieval track, is selected as the strong baseline. The results of *Hyperlink-Extended Pseudo Relevance Feedback* (HPRF) reported in [6] are listed for reference. Following the TREC 2012 microblog task, the results of the TREC 2012 microblog topics are reported with the parameters trained on the TREC 2011 microblog topics.

Table 2. Experiment Results for TREC 2012 Microblog Test Sets

	P@30	MAP
Language Model	0.3983	0.2804
HPRF	0.4339	0.3044
MRM	0.4695	0.3536
Hyperlink-extended Language Model	**0.4734**	**0.3544**

4.3.2. Influence of $P(Q_m|Q)$ and $P(Q_u|Q)$: In Eq. (18), $P(Q_m|Q)$ denotes the probability that a user want to obtain the information from the microblogs, and $P(Q_u|Q)$ is the probability that a user expect to get the information from the linked documents. $P(Q_m|Q)$ and $P(Q_u|Q)$ represent the user's search interest. We tested several values of $P(Q_u|Q)$ ranging from 0 to 1.0, and monitored the retrieval effectiveness measured by P@30 and MAP, see Figure 2. (a) is the performance on TREC 2011 topics and (b) is on TREC 2012 topics. As shown in Figure 2, $P(Q_u|Q) = 0.3$ achieved the best results for both P@30 and MAP on TREC 2011 and TREC 2012. It means that 7:3 weighting between $P(Q_m|Q)$ and $P(Q_u|Q)$ can get the best result. Here $P(D_u)$ is set as a constant 1.0 to ignore the influence of the last item of Eq. 18.



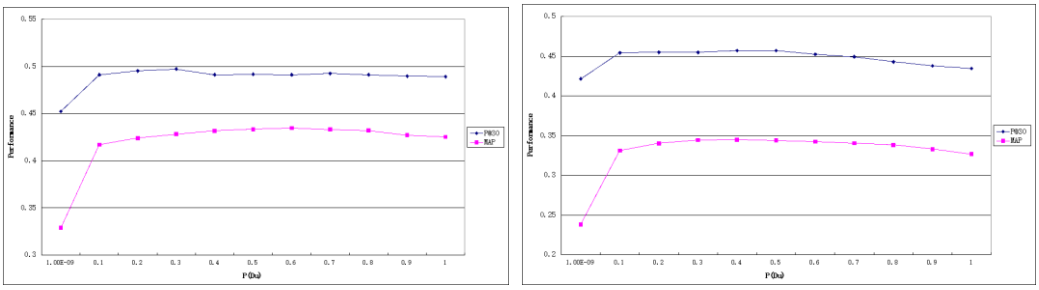
(a) TREC 2011 (b) TREC 2012

Figure 2. Retrieval Performance for different Weight Of $P(Q_u|Q)$ With $P(D_u)$ Is Set As A Constant 1.0 .

4.3.3. Influence of Prior Probability of Links: $P(D_u)$ is used to denote the confidence level of the links. When the hyperlinks exists, the $P(D_u)$ is set 1.0. Otherwise, when the links could not be downloaded or does not exist, a smaller $P(D_u)$ value(from 0 to 1.0, step

0.1) is given to show the punishment to links. Figure 3 illustrates the effect of parameter $P(D_u)$ on retrieval performance. At this time, $P(Q_m|Q)$ is set 1.0 and the hyperlink documents are not used. (a) is for TREC 2011 microblog test set and (b) is for TREC 2012.

The results reflect that even without the content of hyperlink document in Hyperlink-extended Language Model, the information whether the microblog contains links can improve the retrieval performance. The conclusion is in accord with the experiments that used in other algorithm which regard the links as an important feature [14,15].



(a) TREC 2011 (b) TREC 2012

Figure 3. Retrieval Performance for different Weight of P (D_u). P ($Q_u|Q$) Is Set To 0.0.

4.3.4. Combination Influence of P(Du) and P(Qu|Q): We analyze the performance when the parameter is used alone in the previous section. Table 3 shows the combination influence of $P(D_u)$ and $P(Q_u|Q)$. When $P(Q_u|Q)$ is set to 0.0 and we consider only certain aspects of $P(D_u)$, the value of $P(D_u)$ is seen as a punishment for the microblog without hyperlinks. When combine the two parameters $P(D_u)$ and $P(Q_u|Q)$, $P(D_u)$ is regarded as the confidence level of hyperlinked document. At this time, we tend to set a larger value for $P(D_u)$ since the virtual hyperlink documents has contained the punishment. That's demonstrated in Table 3, when $P(D_u)$ is set a larger value, the better retrieval performance can be achieved.

Table 3. Results for Different $P(D_u)$ and $P(Q_u|Q)$

		2011					2012				
	$P(Q_u\|Q)$ / $P(D_u)$	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
P@30	0.6	0.4986	0.5068	0.5054	0.5007	0.4918	0.4638	0.4763	0.4780	0.4808	0.4791
	0.7	0.4939	0.5054	0.5075	0.5020	0.4939	0.4627	0.4757	0.4785	**0.4842**	0.4797
	0.8	0.4939	0.5041	0.5082	0.5054	0.4993	0.4542	0.4734	0.4774	0.4836	0.4819
	0.9	0.4946	0.5041	**0.5122**	0.5054	0.5000	0.4497	0.4695	0.4734	0.4802	0.4802
	1.0	0.4952	0.5034	0.5116	0.5075	0.4973	0.4458	0.4571	0.4689	0.4757	0.4780
MAP	0.6	0.442	0.4422	0.4402	0.4337	0.4165	0.3509	0.3546	0.3568	0.3593	0.3587
	0.7	0.4437	0.4433	0.4427	0.4372	0.4204	0.3490	0.3534	0.3572	0.3607	0.3603
	0.8	0.4453	0.4440	0.4442	0.4392	0.4235	0.3467	0.3517	0.3565	0.3608	0.3613
	0.9	0.4438	0.4450	0.4450	0.4411	0.4310	0.3443	0.3496	0.3544	0.3597	**0.3620**
	1.0	0.4455	0.4452	**0.4458**	0.4428	0.4333	0.3401	0.3468	0.3519	0.3582	0.3614

5. Conclusions

For exploiting the content of hyperlinked documents in retrieval model, we propose a Hyperlink-extended Retrieval Framework which incorporates the content of the microblog and the hyperlinked documents. The Hyperlink-extended Retrieval Framework is realized by a Hyperlink-extended Language Model in this paper. In this model, the

users' retrieval interest is divided into two aspects: the interest to the content of microblog and the interest to the documents linked in microblogs. In detail, KL-divergence is used to measure the similarity between query model and document model. Furthermore, a factor based on the prior probability of links and the query length is used to control the confidence of the linked documents. The results show that our model is significantly outperform the art-of-the-state hyperlink-extended approaches on the standard TREC microblog real-time retrieval collection.

The hyperlink-extended retrieval framework can be regarded as a more general framework for microblog retrieval and there are many researches can be extended into it. For example, the probability $P(Q_m|Q)$ which describes the interest to the content of microblog and the probability $P(Q_u|Q)$ which denotes the interest to the linked documents are allocated a fixed value for all user in this paper. But if the user's click history is available, the $P(Q_u|Q)$ can be given a higher probability when a user always clicks the URLs in the microblog. In the future, we will use the machine learning algorithms to learn which query words trend to look for the content of the microblogs and which one want to read the linked documents.

Meanwhile, we set the same word distribution of query model (means $P(w|Q_m)=P(w|Q_u)=P(w|Q)$) in this paper. In fact, there are some difference between the Q_m and Q_u. For example, users often express their standpoints when posting microblogs in social network. So the microblog content usually contains a larger proportion strong feeling words than the linked documents. So, if there are some strong feeling words in the user's query words, what he concerns may be the other people's opinions. Therefore, to meet users' needs better, $P(w|Q_m)$ with emotion words in query can be higher than $P(w|Q_u)$. In addition, the $P(w|Q_m)$ and $P(w|Q_u)$ can also be estimated by the query expansion technology on the microblog corpus and linked documents corpus respectively to catch the difference between the Q_m and Q_u.

In the future, more weight estimation methods of microblog model and linked document model will be explored based on our Hyperlink-extended Microblog Retrieval Framework to improve the microblog retrieval performance.

Acknowledgements

This work is supported by the NSF China (No. 61370170 & 61272384 & 61402134 & 61173074) and the National Social Science Fund China (No. 14CTQ032).

References

[1] Morris M. R., Counts S. and Roseway A., "Tweeting is believing? Understanding Microblog Credibility Perceptions", Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work. ACM, Washington, USA, February **(2012)**, pp. 441-450.
[2] Kwak H., Lee C. and Park H, "What is Twitter, a social network or a news media"? Proceedings of the 19th International Conference on World Wide Web, ACM, Raleigh, USA, April **(2010)**, pp. 591-600.
[3] Cassidy M. K. M. and Kulick M., "An update to Google social search. The Official Google Blog, **(2011)**, pp. 17.
[4] Schwartz B., "Bing Adds Twitter Smart Answers", Search Engine Land, July **(2009)**.
[5] Nagmoti R., Teredesai A. and De Cock M., "Ranking approaches for microblog search", In 2010 IEEE/WIC/ACM International joint conference on Web Intelligence-Intelligent Agent Technology (WI-IAT), New York, USA, **(2010)**, pp. 153-157.
[6] El-Ganainy T., Magdy W. and Rafea A., "Hyperlink-Extended Pseudo Relevance Feedback for Improved Microblog Retrieval", SoMeRA, Gold Coast, Australia, July **(2014)**, pp. 7-12.
[7] Liang F., Qiang R. and Yang J., "Exploiting real-time information retrieval in the microblogosphere", Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries. ACM, New York, USA, June **(2012)**, pp. 267-276.
[8] McCreadie R. and Macdonald C., "Relevance in microblogs: Enhancing tweet retrieval using hyperlinked documents", Proceedings of the 10th conference on open research areas in information retrieval. Le Centre De Hautes Etudes Internationales D'informatique Documentaire, Paris, France, May **(2013)**, pp. 189-196.

[9] Limsopatham N., McCreadie R. and Albakour M., "University of Glasgow at TREC 2012: Experiments with Terrier in medical records, microblog, and web tracks", Proceedings of The 21th Text Retrieval Conference, TREC, Gaithersburg, USA, **(2012)**.
[10] Kim Y., Yeniterzi R. and Callan J., "Overcoming vocabulary limitations in twitter microblogs", The 21th Text Retrieval Conference, TREC2012, Gaithersburg, USA, **(2012)**.
[11] Donald M. and Congxing C., "USC/ISI at TREC 2011: Microblog Track.The 20th Text Retrieval Conference, TREC2011, Gaithersburg, USA, **(2011)**.
[12] Van Duc T. H., Demeester T. and Deleu J., "Urgent participation in the microblog track 2012", The 21th Text Retrieval Conference, TREC2012, Gaithersburg, USA, **(2012)**.
[13] El-Ganainy T., Wei Z. and Magdy W., "QCRI at TREC 2013 Microblog Track", The 22th Text Retrieval Conference, TREC2013, Gaithersburg, USA, **(2013)**.
[14] S. Zhu, Z. Gao, Y. Yuan, H. Wang and G. Chen, "PRIS at TREC 2013 Microblog Track", The 22th Text Retrieval Conference, TREC2013, Gaithersburg, USA, **(2013)**.
[15] Han Z., Li X. and Yang M., "Hit at trec 2012 microblog track", The 21th Text Retrieval Conference, TREC2012, Gaithersburg, USA, **(2012)**.
[16] Lafferty J. and Zhai C., "Document language models, query models, and risk minimization for information retrieval", Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 9-12111-119, September **(2001)**.
[17] Song F. and Croft W. B., "A general language model for information retrieval", Proceedings on the 22nd Annual International ACM SIGIR Conference, **(1999)**, pp. 279-280.
[18] Zhai C. and Lafferty J., "A study of smoothing methods for language models applied to ad hoc information retrieval", Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, **(2001)**, pp. 334-342.
[19] Teevan J., Ramage D. and Morris M. R., "#TwitterSearch: a comparison of microblog search and web search", Proceedings of the fourth ACM international conference on Web search and data mining. ACM, Hong Kong, China, February **(2011)**, pp. 35-44.
[20] Ounis I., Macdonald C., Lin J. and Soboroff I., "Overview of the TREC-2011 Microblog Track", The 20th Text Retrieval Conference, TREC2011, Gaithersburg, USA, **(2011)**.
[21] Soboroff I, Ounis I., Macdonald C. and Lin J., "Overview of the TREC-2012 Microblog Track", the 21th Text Retrieval Conference, TREC2012, Gaithersburg, USA, **(2012)**.
[22] Han Z., Li X. and Yang M., "Feature Analysis in Microblog Retrieval Based on Learning to Rank", Natural Language Processing and Chinese Computing, **(2013)**, pp. 410-416.

Author

Zhongyuan Han, he was born in 1977, Ph. D. candidate, Associate Professor. His research interests include information retrieval, information filtering, and natural language processing.

