ITC 4/48

Information Technology
and Control
Vol. 48 / No. 4 / 2019
pp. 626-636
DOI 10.5755/j01.itc.48.4.22487

Incorporating Semantic Word Representations into Query Expansion for Microblog Information Retrieval

Received 2019/01/11 | Accepted after revision 2019/10/15

http://dx.doi.org/10.5755/j01.itc.48.4.22487

Incorporating Semantic Word Representations into Query Expansion for Microblog Information Retrieval

Bo Xu

Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, A923 Chuangxinyuan Building of Dalian University of Technology, Dalian, China;

State Key Laboratory of Cognitive Intelligence, iFLYTEK, P.R. China; e-mail: xubo@dlut.edu.cn

Hongfei Lin

Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, A923 Chuangxinyuan Building of Dalian University of Technology, Dalian, China; e-mail: hflin@dlut.edu.cn

Yuan Lin

WISE Lab, School of Public Administration and Law, A923 Chuangxinyuan Building of Dalian University of Technology, Dalian, China; e-mail: zhlin@dlut.edu.cn

Kan Xu

Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, A923 Chuangxinyuan Building of Dalian University of Technology, Dalian, China; e-mail: xukan@dlut.edu.cn

Lin Wang

Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, A923 Chuangxinyuan Building of Dalian University of Technology, Dalian, China; e-mail: nilgnaw@mail.dlut.edu.cn

Jiping Gao

Institute of Scientific and Technical Information of China, e-mail: gaojp@istic.ac.cn

Corresponding author: xubo@dlut.edu.cn

Microblog information retrieval has attracted much attention of researchers to capture the desired information in daily communications on social networks. Since the contents of microblogs are always non-standardized and flexible, including many popular Internet expressions, the retrieval accuracy of microblogs has much room for improvement. To enhance microblog information retrieval, we propose a novel query expansion method to enrich user queries with semantic word representations. In our method, we use a neural network model to map each word in the corpus to a low-dimensional vector representation. The mapped word vectors satisfy the algebraic vector addition operation, and the new vector obtained by the addition operation can express some common attributes of the two words. In this sense, we represent keywords in user queries as vectors, sum all the keyword vectors, and use the obtained query vectors to select the expansion words. In addition, we also combine the traditional pseudo-relevance feedback query expansion method with the proposed query expansion method. Experimental results show that the proposed method is effective and reduces noises in the expanded query, which improves the accuracy of microblog retrieval.

KEYWORDS: Microblog retrieval, query expansion, word embeddings, word vectors, information retrieval.

1. Introduction

With the rapid development of social networks and the Internet, it is increasingly difficult to accurately obtain information from a large amount of information resources, such as microblog texts, which involves massive information in the form of short texts. According to the statistics [17], the volume of microblog users has reached 309 million, and the usage rate is as high as 54.7%. Therefore, social communications on microblogs have gradually become one of the most important forms of communications in people's daily life [30].

The characteristics of microblogs include short information length, flexible language form, large data scale, strong timeliness and fast update speed. These characteristics make this task different from other information retrieval tasks. Microblogs, as short texts, contain less semantic information than other types of documents, which hinders the text matching in this retrieval task. Therefore, it remains a great challenge in matching user queries with the latent relevant short microblogs to the largest semantic degree. Moreover, microblogs, compared with other types of documents, contain more noises, such as the Internet and colloquial expressions. How to use these expressions in text matching remains a difficult problem in microblog retrieval. The expressions in short microblogs involve a lot of semantic information and preferences of users, which may enhance the retrieval performance.

However, the useful information cannot be accurately and fully obtained by directly modeling microblogs using traditional information retrieval models. The reason is that on the one hand, microblog has a single piece of information, and the writing method is arbitrary with many Internet popular expressions. On the other hand, microblog retrieval only uses the search keywords provided by the users, which cannot provide enough information to fully express the user's needs. The existing research shows that users' query words can only express a small part of the user's information needs [8]. One of the most direct and effective ways to solve these two problems is to apply the query expansion techniques.

Query expansion techniques refer to the process of reorganizing queries by using natural language processing, clustering and other methods to improve information retrieval performance [1]. At present, query expansion techniques have achieved good results in the traditional text retrieval tasks. However, directly applying these methods to microblog information retrieval cannot achieve the desired performance. Most query expansion techniques either use local document information for relevance feedback or use external dictionaries to cover new expansion terms, such as popular Internet language and homophonic terms in microblogs. The new expansion terms contained in microblogs play a crucial role in the user's retrieval request. Therefore, it is necessary to optimize query expansion for microblog retrieval by considering more semantic information in user queries.

In this paper, we propose a novel query expansion method using semantic word representations for microblog information retrieval so as to improve the accuracy of understanding the user's query intentions in microblog search. Word representations can map words into a new space and represent them in a multidimensional continuous real vector. Studies have shown that word vectors trained by neural networks have the property of additivity [20], and many applications have used this property to obtain good results in different natural language processing tasks, such as phrase recognition [20]. Since these word vectors are trained by the entire text set, all the contents of the text set are considered, and no external data sets are needed during the training process. Therefore, we believe that semantic word representations may be a valuable source for incorporating high quality expansion words. The contributions of this paper are listed as follows:

1. We analyze user's search terms, and use the additivity of the word vectors to understand the user's query intents to select the most relevant words as the candidates for query expansion;
2. We integrate the word representation based query expansion into the pseudo-relevance feedback to enrich the topics of results and reduce topic intersections;
3. We conduct extensive experiments using the data from Sina Microblogs, and experimental results show that the proposed method is effective.

2. Related Work

There are currently three approaches to query expansions: query expansion based on relevance feedback, query expansion based on local analysis, and query expansion based on global analysis.

Query expansion based on relevance feedback utilizes feedback from the initial retrieval to enrich the original query. Specifically, retrieval system returns a set of search results to the user's original query, and the user checks the set of results and labels the relevant and irrelevant ones. Then, the retrieval system uses the important words in the relevant documents to expand the query, and returns the optimized search result. The advantage of this method is that the query expansion words can be accurately obtained, and the accuracy of the search can be greatly improved. The disadvantage is that the user needs to participate, and a large amount of data is needed for parameter training. If the feedback of the user is wrong, the system performance will drop sharply. As classic and important query expansion methods, relevance model was proposed to incorporate probabilistic models into query likelihood language models [13]. Moreover, term dependency was modeled via Markov random field to incorporate dependence and independence relationship of terms for query expansion [19]. For example, Abberley et al. [1] showed that this method helps to improve the accuracy of the query in a small scope, but increases the query burden of users.

Query expansion based on local analysis is also known as pseudo-relevance feedback method. Specifically, retrieval system performs the initial retrieval on the user query and assumes that the first N documents returned are relevant documents, and then the important words therein are used as query expansion words for query expansion. This method overcomes the shortcomings of relevance feedback that require user participation, but slightly reduces the accuracy of information retrieval. Kelly and Teevan [11] proposed to add the user's query log based on the pseudo-relevance feedback to infer the user's query intents, and automatically perform query expansion based on relevant information. Shen et al. [26] proposed to add user clicks as implicit feedback for secondary sorting of the document, and obtained better results. Kurland et al. [12] also proposed to repeat query expansion by iterations, but experiments show that the method suffered from problems such as query theme drifts.

Query expansion based on global analysis aims to mine the relevance difference among words, and treats the most relevant words as complements to the query. Specific methods include word clustering, latent semantic analysis [23], co-word analysis [28] and semantic dictionary (WordNet) [25]. Qiu et al. [24] proposed using words similar to all query words as query expansion words. These methods enrich the representation of query words semantically, but did not try to understand the user's query intent, leading to problems such as subject shifting and introducing noises. In order to prevent problems such as subject shifting, Xu and Croft [32] proposed a local context analysis method, which demonstrated that high-quality expanded words tend to co-occur with all query words in the front documents returned by the initial search. Balog et al. [4] proposed an expansion method that considers both query dependent and query independent terms to improve the recall rate of retrieval.

Meij et al. [18] proposed a query expansion method that combines local analysis and global analysis for better retrieval performance.

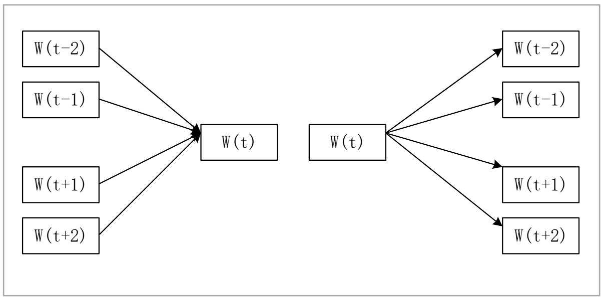
The above mentioned query expansion methods have been applied in traditional text retrieval field. However, it is difficult to achieve desired performance by directly using these methods in microblog retrieval. This is because there is always a large number of network vocabularies in microblogs, as well as a lot of junk text, without any useful information. Because of these factors, if top-ranked microblogs returned by the initial search are not relevant, microblog query expansion through pseudo-relevance feedback will be of little use. Moreover, because external dictionaries contain few emerging online words, which appear frequently in the microblogs, it is difficult to meet the needs of users by using global dictionary based query expansion methods such as statistical dictionaries and semantic dictionaries. To improve microblog retrieval performance, Wang et al. [29] proposed a multimedia expansion framework using microblog clustering based on social networks and semantic associations. Zhou et al. [34] proposed to personalize query expansion by using user-generated markup content such as social tags. Anagnostopoulos et al. [3] used the tags on Twitter to extract the semantic information and form a semantic network graph for semantic query expansion. Li et al. [16] incorporated time factors into the language model as document prior probability for considering relevant documents at certain time points. Since the method in [16] achieved the state-of-the-art performance, we treat it as a strong baseline in our experiments.

In order to obtain high-quality query expansion terms for microblog, it is necessary to understand the semantic information behind the user queries. One of the most effective methods for obtaining semantic information of terms is semantic word representations. Recently, there are several ways to map terms to low-dimensional space, including neural networks [21], matrix factorization based on word co-occurrence [14], and data representations based on word context information [15]. Recent studies have addressed microblog retrieval using various methods [2, 7, 10, 27, 31, 33, 35]. Chen et al. [7] proposed a temporal predictor to capture the temporal relevance of words using kernel density estimation over the feedback documents. Hasanain et al. [10] investigated 37 state-of-the-art query performance predictors in combination with different retrieval models for microblog search. Zhang et al. [33] seek for efficient top-k query processing in a huge microblog dataset for compact indexing and judicious search. Models that use neural networks to train word vectors include continuous bag of words (CBOW) and the skip-gram models. These two models have been demonstrated effective in different natural language processing tasks [5]. Moreover, Mikolov et al. [21] found that the word vectors trained by these methods subject to algebraic vector addition operations, namely, the word vector for “Russia” plus the vector “river” equals to a similar vector with the vector for “Volga river”. This property has been applied in the phrase recognition task [1]. In this paper, we will use the semantic word representations to select high-quality query expansion terms. Since all word representations are trained by the entire microblog corpus, this method belongs to the query expansion method based on global analysis. In addition, in order to prevent topic shifting, we comprehensively combine the pseudo-relevance feedback algorithm and the global analysis query expansion method by synthesizing their own advantages for select the query expansion terms that can best express user’s intentions.

Next, we briefly introduce the basic descriptions about word embedding training [6, 9, 22]. In order to train high-quality word representations, the representation vectors are updated in iterations, so that the objective function are consistently optimized for improving the quality of the vectors. There are two types of models for training word vectors: continuous bag-of-word (CBOW) and the skip-gram model, which are illustrated in Figure 1.

Figure 1

Word Representation Models: (1) CBOW (2) skip-gram



In [6], a three-layer neural network model is used to update the word representations and continuously improve the accuracy of the CBOW or skip-gram model in the training process. The goal of the CBOW model is to learn high quality word representations and predict the current word based on the adjacent words in the contexts.

$$\max\{P(\mathrm{w}_t \mid \mathrm{w}_{t-1},...,\mathrm{w}_{t-n+2},\mathrm{w}_{t-n+1})\}, \tag{1}$$

subject to

$$P(\mathrm{w}_t \mid \mathrm{w}_{t-1},...,\mathrm{w}_{t-n+2},\mathrm{w}_{t-n+1}) > 0, \tag{2}$$

$$\sum_{w}^{|\mathrm{V}|} P(\mathrm{w} \mid \mathrm{w}_{t-1},...,\mathrm{w}_{t-n+2},\mathrm{w}_{t-n+1}) = 1. \tag{3}$$

The input layer contains N word vectors, which reaches the output layer through a hidden layer. The output layer is composed of |V| nodes, and each node represents the probability of the current word. The probability is formulated as follows:

$$P(\mathrm{w} \mid \mathrm{w}_{t-1},...,\mathrm{w}_{t-n+2},\mathrm{w}_{t-n+1}) = \frac{\mathrm{e}^{y_w}}{\sum_{t}^{|\mathrm{V}|} e^{y_{w_t}}}. \tag{4}$$

where

$y_w = b + Wx + U\,\mathrm{tanh(d+Hx)}$

$\mathrm{x} = (\mathrm{C}(\mathrm{w}_{t-1}),...,\mathrm{C}(\mathrm{w}_{t-n+2}),\mathrm{C}(\mathrm{w}_{t-n+1}))$.

By optimizing the maximum likelihood probability, all parameters, together with the word representations, are updated using gradient descent to obtain the optimal word vector representations. In practical applications, it is often assumed that the direct connection network from the input layer to the output layer is removed. In theory, the above model can be optimized by the gradient descent formula, but there are still many problems that have not been fully solved. It is noted that the probability denominator of the current word generation needs to consider all the words in the corpus, which will result in extremely low efficiency of the entire training model. In order to solve this problem, the word vector can be trained using the following two methods: hierarchical Softmax [22] or negative sampling [9], so that the complexity of training can be reduced from $|V|$ to $\log_2|V|$. Studies have shown that negative sampling has a better effect than hierarchical Softmax. Therefore, we use the negative sampling method to train word vectors for microblog query expansion.

3. Microblog Query Expansion Based on Word Representations

We adopt the above-mentioned neural network based model to learning word representations for query expansion. The vector representations of query words in the same query (excluding the stop words) are added to obtain a query vector. Specifically, we perform word segmentation processing on the user's query text and remove the stop words to obtain a query word set. The word vectors in the query word set are then normalized and the vectors are added to obtain a new vector as the query vector. We choose the terms closest to the query vector from the corpus dictionary as complements to enrich user queries, namely, the set formed by the words is used as a query expansion candidate set. Because this query vector is determined by all the query terms, it can best express the user's search intents. By choosing the closest words as candidate expansion terms, it has a high probability to understand the user's query intention in depth, and improves the accuracy of information retrieval. We call this method the word vector model (VEC).

In addition, because a large number of microblog languages are not as formal as traditional texts, they are more random, and contain a large number of network terms. These network terms cannot be expanded using query expansion methods such as WordNet. The proposed method is based on the word vectors trained by microblog itself, which learn the knowledge of network terms. It may have unique advantages compared to other query expansion methods.

To further enhance our method, we combine the proposed method with pseudo-relevance feedback. The pseudo-relevance feedback algorithm belongs to the query expansion method of local analysis. The main idea is to use the top-ranked document returned by the search engine as the feedback document, and find the non-stop words with higher word frequency as the expanded words and add them to the original queries for a second search. This method assumes that the search engine first retrieves the returned top-ranked document as a relevant document, but this condition is often not satisfied, especially on microblog retrieval tasks. Studies have shown that this method can improve the average accuracy of information retrieval to some extent. Therefore, we propose to optimize pseudo-relevance feedback with semantic word representations. For these two methods, the pseudo-relevance feedback method uses the partial results returned by the search engine for initial query expansion, and relies much on the results of the initial retrieval, while the representation-based method uses the neural network model to perform global word vector training, and finally understands the user's intention through the vector sum of the query words so as to find the vocabulary that best expresses the user's search intention. These two methods belong to the local analysis and global analysis, respectively. Moreover, based on our preliminary experiments, we observe that these two methods obtain different expansion words through different perspective with different noisy expansion terms. In order to enhance the retrieval performance, we use the candidate expansion words obtained by the two methods as the final expansion words. We call this method (PSVEC), and it can greatly avoid the subject shift problem caused by the introduction of noises. Specifically, we use the intersections of these two sets of expansion terms formulated in Eq. (5):

$$V = \{\mathrm{x} \mid \mathrm{x} \in \mathrm{V}_1 \bigcap \mathrm{x} \in \mathrm{V}_2\}. \tag{5}$$

We illustrate the proposed framework in Fig. 2.As shown in Fig. 2, the process of the proposed method is as follows:

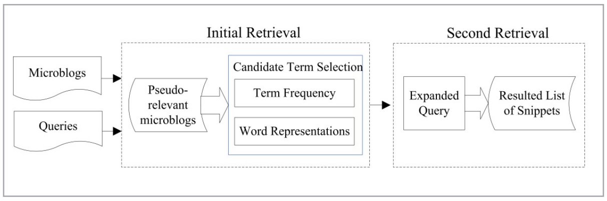
1 The original query of the user is directly searched by a certain search engine framework such as Lucene (http://lucene.apache.org/), and the stop words are removed, thereby obtaining the first high frequency *m* keywords of the top-ranked documents.

2 The word vector-based query expansion method is used to obtain the user query vector and calculate the query vector.

3 The similarity between the query vector obtained in step (2) and the *m* keyword vectors obtained in step (1) is calculated. We then select the expansion words with the highest similarity as the query expansion words, and conduct a second retrieval to obtain the results for users.

Figure 2

The combination of word representation based query expansion and pseudo-relevance feedback



4. Experiments

4.1. Experimental Settings

We crawl the experimental dataset from Sina Microblog, the largest microblog website in China. The corpus covers the microblogs related to the hot events that occurred from September 2013 to November 2017. We continuously detected the hot events in Sina Microblog and used the crawler program to crawl the relevant microblog, and obtained a total of 2.4G Sina Microblog. Since our research focused on microblogs, we removed the "author", "@ mention", "url" and other information. In addition, since there is a large amount of forwarded text in microblogs with little useful information, we removed all the forwarded microblogs and only kept the original microblogs. Since too short microblog often lacks information, we consider microblogs with less than 30 words as spams and filter them in the preprocessing steps. The total number of microblogs used in the experiment was 5,658,291, involving 456,234 users. After using the public Ansj tokenizer (http://www.oschina.net/question/tag/ansj) to segment the microblog corpus, the number of words is 330,792,290 in total.

We searched for four types of query terms in different fields, including economics, sports, electronics, and music. Ten graduate students were invited to annotate the top 100 microblogs returned using the default retrieval model of the Lucene search engine for evaluating the retrieval performance. The rating levels include "irrelevant", "slightly relevant", and "relevant". Since the Normalized Discounted Cumulative Gain (NDCG) indicator can handle multi-level evaluation results, the NDCG value is mainly used as an evaluation metrics in the experiment. We adopt NDCG@30 as the main evaluation metric, which measures the

NDCG values in terms of top 30 documents. This is because top-ranked documents are the most important indicators that reflect the retrieval performance and user satisfaction. In addition, if the slightly relevant documents are regarded as relevant documents in a unified manner, the Mean Average Precision (MAP) evaluation index can be used as the auxiliary evaluation metrics of the retrieval performance of the proposed method.

The number of expansion terms is a hyper-parameter in query expansion-based retrieval. Based on the preliminary experiments on a development set, we find that both the pseudo-relevance feedback method and the word vector-based query expansion method achieve the best performance when setting the number of expansion terms to eight in our experiments. Therefore, in the following experiments, we set the number of expanded words to eight words. In other circumstances such as Twitter search, the number of expansion terms can be tuned using a development dataset. The optimal number of expansion terms can be used to improve the retrieval performance to the largest extent in practical applications. Since the proposed method is general, it can also be extended to the microblog retrieval from other sources, such as Twitter. Since these exist large differences between Chinese and English in terms of microblog expressions and Internet languages, our interest and further work remains Twitter-oriented retrieval.

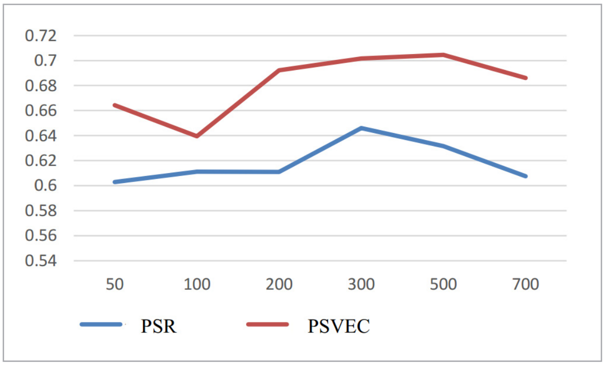
4.2. Parameter Selection for seudo-Relevance Feedback

We index the Sina Microblog corpus using the Lucene 3.5 search engine released by Apache. The pseudo-relevance feedback method selects the candidate words with the highest term frequency as the expanded words after removing the stop words from the top-ranked documents retrieved by Lucene. The expansion words are combined with the original query as the expanded query for second retrieval. We illustrate the retrieval performance with respect to different numbers of documents in terms of NDCG@30 in Fig. 3.

It can be observed from Fig. 3 that when the number of documents is 300, we can obtain the best NDCG@30 value. Therefore, in the subsequent experiments, we set the number of documents as 300.

Figure 3

Retrieval performance in terms of NDCG@30



4.3. Models Compared

We first compare our method with the vector space model (VSM). VSM assumes that documents and queries are parts of a *t*-dimensional vector space, where *t* is the number of index terms. In the vector space, each document *D* can be represented as a vector of index terms D_i=(d_{i1}, d_{i2},... , d_{it}), where d_{ij} denotes the weight of the j^{th} word in document D_i. In a similar way, each query *Q* can be expressed as a *t*-dimensional vector: $Q = (q_1, q_2, ..., q_t)$, and q_i is the weight of the i^{th} word in the query vector. The documents are sorted based on the cosine similarities of the document vector and the query vector. We use this method as the first baseline model. The cosine formula is defined as follows:

$$\cos(\mathrm{D}_i,\mathrm{Q}) = \frac{\sum_{j=1}^{t} d_{ij} * q_j}{\sqrt{\sum_{j=1}^{t} d_{ij}^{2} * \sum_{j=1}^{t} q_j^{2}}}. \qquad (6)$$

We adopt a modified language model HTLM [16] as the second baseline model. The model considers that each document d is generated from a polynomial distribution, and ranked using the language model, shown as follows:

$$P(\mathrm{D}_i \mid \mathrm{Q}) = \mathrm{logp}(\mathrm{D}_i) + \sum_{w \in V} tf(\mathrm{w},\mathrm{Q})\,\mathrm{logp}(\mathrm{w} \mid D_i). \qquad (7)$$

We also directly use the pseudo-relevance feedback algorithm to obtain the feedback from the top-ranked 300 documents, and select the eight most frequent words as the query expansion words as the third baseline model (denoted as PSR short for PSeudo-Relevance feedback). Moreover, we also compare the proposed model with two classic and effective query expansion methods, term dependency model [19] (denoted as TD) and relevance model [13] (denoted as RM) in our experiments.

4.4. Retrieval Performance of the Combination of Word Representation-based Query Expansion and Pseudo-Relevance Feedback

We use the publicly available word embedding source code provided by Google (http://code.google.com/p/word2vec/) to train the word vector for the microblog corpus. We use the skip-gram model with hierarchical Softmax optimization method, and set the vector dimension as 200, and the window size as 8. We finally obtain 419,063 word vectors. The query expansion method based on the word vectors first removes the stop words from the query, then normalizes the remaining word vectors and performs the sum operation, and finally selects the eight words with the highest similarity to the obtained vector as the expansion words for query expansion. The expanded words are combined with the original query at the ratio of 1:3 and retrieved as a new query. The experimental results are shown in Table 1.

Through the above mentioned experimental results, we can find that the two methods of VEC and PSR are much better than the other two methods, which can improve the query quality to a certain extent, and the two query expansion methods have the same effect. We observe that by calculating the similarity between the top 300 high-frequency keywords and the query vector after removing the stop words in the top-ranked documents, the eight words with the highest similarity scores can be obtained as the query expansion words. The expanded words are combined with the original query as a new query. It can be seen from Fig. 3 that the effect is better when N=300. The experimental results are shown in Table 2.

Table 2

The experimental results of the five models

Model	NDCG@30	NDCG@60	MAP
VSM	0.514	0.491	0.387
HTLM	0.526	0.494	0.379
PSR	0.646	0.638	0.597
VEC	0.659	0.608	0.570
RM	0.625	0.583	0.525
TD	0.639	0.608	0.547
PSVEC	0.702	0.645	0.656

Through the experimental results, it can be easily found that the fusion of the two methods is better than any of the methods. It can be seen that the fusion of the two methods is a good choice.

In order to more clearly compare the advantages and disadvantages of the three algorithms, namely PSR, VEC and PSVEC, we select the best results of the three algorithms for comparison. As shown in Fig. 4, it is not difficult to find that VEC and PSR have the same effect. One is an expansion method based on local queries, and the other is a query expansion method based on global analysis. PSR's noise mainly comes from irrelevant documents, while VEC's noise mainly

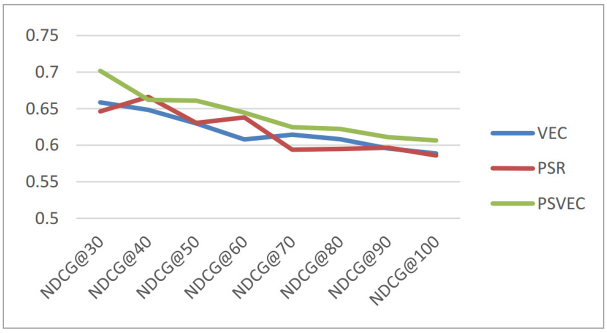
Table 1

Retrieval Performance in terms of NDCG

Model	NDCG@30	NDCG @40	NDCG @50	NDCG @60	NDCG @70	NDCG @80	NDCG @90	NDCG @100
VEC	0.659	0.648	0.630	0.608	0.614	0.608	0.595	0.589
VSM	0.514	0.554	0.528	0.491	0.474	0.434	0.413	0.397
HTLM	0.526	0.542	0.517	0.494	0.474	0.446	0.414	0.393
PSR	0.646	0.666	0.630	0.638	0.594	0.595	0.597	0.586
TD	0.639	0.622	0.615	0.608	0.610	0.582	0.574	0.551
RM	0.625	0.609	0.612	0.583	0.576	0.579	0.552	0.536

Figure 4

Retrieval performance in terms of NDCG



comes from word vector training error. By combining the two methods, it is possible to complement each other for better performance.

The experimental results show that the pseudo-relevance feedback query expansion method based on local analysis achieves comparable results with the word vector based query expansion method based on global analysis. The main reason is that the pseudo-relevance feedback query expansion method relies too much on the results returned by the first retrieval. If the quality of the results returned by the first retrieval is high, the pseudo-relevance feedback query expansion method will get high quality query expansion words. Conversely, if the results returned by the first retrieval are not ideal, the pseudo-relevance feedback method is prone to subject offsets. The microblog content organization is not standardized, and the writing method is arbitrarily leading to the result of the first search engine returning. On the other hand, because the average length of the microblogs is short, the optimal value of *N* tends to be relatively large, further reducing the result. Moreover, the query expansion method based on the word vector model relies too much on the word vector expression of words. Nowadays, the research on the word vector training method in natural language processing is not mature, so the word vector model can easily lead to the introduction of noise vocabulary. On the other hand, microblog contents are often organized in an unstandardized way and do not conform to the syntactic structure of the formal language. Therefore, the effect of using this method alone is not particularly desirable.

Although the results of the above two methods introduce different types of noises, the experimental results show that the noise categories introduced by the two methods are different. One is the noise caused by the local analysis, which is easy to cause the subject shift, and the other one is the noise caused by the vector representation error through the global analysis. Therefore, by combining the two methods, the two query expansion methods complement each other and successfully reduce the subject shift caused by the introduction of noises. Through the experimental results, it is not difficult to find that after the two algorithms are combined, the obtained results are effective, and the NDCG value can be greatly improved. We also compared the time cost of different methods. We observe that query expansion based methods (TD, RM and our methods) cost a bit more time than other baseline methods in the retrieval process, although query expansion enhance the retrieval performance. This is because query expansion is executed through two stages, the initial retrieval and the second retrieval, while other methods only execute one time retrieval. Since the total time costs of different methods are comparable to each other, we believe query expansion based microblog retrieval is more effective in improving the overall retrieval performance.

5. Conclusions

This paper proposes two methods to establish an efficient and accurate retrieval model for microblog retrieval. Both methods use query expansion techniques. The first method is based on the semantic word representations. The word embedding technique is used to map all the words of the text corpus to low-dimensional vectors, and the user's query vector is obtained by analyzing the user query words, so as to use the similarity between the vectors. The second method is to combine the pseudo-relevance feedback and word vector based query expansion method to filter out the irrelevant words obtained in the pseudo-relevance feedback. Experiments show that the accuracy of the pseudo-relevance feedback algorithm in the first method is the same as that in the traditional method. It shows that any single method introduces a large amount of noise information while enriching the user query, and even generates the subject shift phenomenon like the pseudo-relevance feedback algorithm. Since pseudo-relevance feedback algorithm belongs to the local analysis algorithm and the word vector-based query expansion method is the global analysis method, the noise information introduced by the two methods tends to be different. Querying the intersection of the expansion words can effectively avoid noises. The combination method achieves much improvement in terms of NDCG. Future research will be carried out by assigning appropriate weights on expansion terms and fully utilizing the user and tag information from the microblogs. We will also extend our work to other types of microblogs, such as Twitter.

Acknowledgement

This work is partially supported by a grant from the Ministry of Education Humanities and Social Science Project (No.19YJCZH199), the Foundation of State Key Laboratory of Cognitive Intelligence, iFLYTEK, P.R. China (COGOS-20190001, Intelligent Medical Question Answering based on User Profiling and Knowledge Graph), the Natural Science Foundation of China (No.61632011, 61572102, 61976036, 61702080, 61602079), Postdoctoral Science Foundation of China (No. 2018M641691), and the Fundamental Research Funds for the Central Universities (No. DUT18ZD102).

References

1. Abberley, D., Kirby, D., Renal, S., Robinson, T. The THISL Broadcast News Retrieval System. In Proceedings of ESCA Workshop on Accessing Information in Spoken Audio, 1999, 19-24.
2. Albishre, K., Li, Y., Xu, Y. Query-Based Automatic Training Set Selection for Microblog Retrieval. Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2018, 325-336. https://doi.org/10.1007/978-3-319-93037-4_26
3. Anagnostopoulos, I., Kolias, V., Mylonas, P. Socio-semantic Query Expansion Using Twitter Hashtags. 2012 Seventh International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), 2012, 29-34. https://doi.org/10.1109/SMAP.2012.15
4. Balog, K., Weerkamp, W., Rijke, M. A Few Examples Go a Long Way: Constructing Query Models from Elaborate Query Formulations. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2008, 371-378. https://doi.org/10.1145/1390334.1390399
5. Bengio, Y., Bengio, S. Modeling High-Dimensional Discrete Data with Multi-Layer Neural Networks. NIPS, 1999, 99, 400-406.
6. Bengio, Y., Ducharme, R., Vincent, P., Janvin, C. A Neural Probabilistic Language Model. The Journal of Machine Learning Research, 2003, 3, 1137-1155.
7. Chen, Q., Hu, Q., Huang, J. X., He, L. TAKer: Fine-Grained Time-Aware Microblog Search with Kernel Density Estimation. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(8), 1602-1615. https://doi.org/10.1109/TKDE.2018.2794538
8. González-Caro, C., Calderón-Benavides, L., Baeza-Yates, R., Tansini, L., Dubhashi, D. Web Queries: The Tip of the Iceberg of the User's Intent. Workshop on User Modeling for Web Applications, 2011, 2011.
9. Gutmann, M. U., Hyvärinen, A. Noise-contrastive Estimation of Unnormalized Statistical Models, with Applications to Natural Image Statistics. Journal of Machine Learning Research, 2012, 13(Feb), 307-361.
10. Hasanain, M., Elsayed, T. Query Performance Prediction for Microblog Search. Information Processing and Management, 2017, 53(6), 1320-1341. https://doi.org/10.1016/j.ipm.2017.08.002
11. Kelly, D., Teevan, J. Implicit Feedback for Inferring User Preference: A Bibliography. ACM SIGIR Forum, 2003, 37(2), 18-28. https://doi.org/10.1145/959258.959260
12. Kurland, O., Lee, L., Domshlak, C. Better than the Real Thing: Iterative Pseudo-Query Processing Using Cluster-based Language Models. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2005, 19-26. https://doi.org/10.1145/1076034.1076041
13. Lavrenko, V., Croft, W. B. Relevance-based Language Models. International ACM SIGIR Conference on Research and Development in Information Retrieval, 2001, 51(2), 120-127. https://doi.org/10.1145/3130348.3130376
14. Lebret, R., Collobert, R. Word Emdeddings Through Hellinger PCA. arXiv preprint arXiv:1312.5542, 2013. https://doi.org/10.3115/v1/E14-1051
15. Levy, O., Goldberg, Y., Ramat-Gan, I. Linguistic Regularities in Sparse and Explicit Word Representations.

CoNLL-2014, 2014, 171. https://doi.org/10.3115/v1/W14-1618

16. Li, L., Xu, G., Yang, Z., Dolog, P., Zhang, Y., Kitsuregawa, M. An Efficient Approach to Suggesting Topically Related Web Queries Using Hidden Topic Model. World Wide Web, 2013, 16(3), 273-297. https://doi.org/10.1007/s11280-011-0151-3

17. Ma, J. X., Buhalis, D., Song, H. ICTs and Internet Adoption in China's Tourism Industry. International Journal of Information Management, 2003, 23(6), 451-467. https://doi.org/10.1016/j.ijinfomgt.2003.09.002

18. Meij, E., Weerkamp, W., Rijke, M. A Query Model Based on Normalized Log-Likelihood. Proceedings of the 18th ACM Conference on Information and Knowledge Management, 2009, 1903-1906. https://doi.org/10.1145/1645953.1646261

19. Metzler, D., Croft, W. B. A Markov Random Field Model for Term Dependencies. International ACM SIGIR Conference on Research and Development in Information Retrieval, 2005, 472-479. https://doi.org/10.1145/1076034.1076115

20. Mikolov, T., Chen, K., Corrado, G., Dean, J. Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781, 2013.

21. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. Advances in Neural Information Processing Systems, 2013, 3111-3119.

22. Morin, F., Bengio, Y. Hierarchical Probabilistic Neural Network Language Model. Proceedings of the International Workshop on Artificial Intelligence and Statistics, 2005, 246-252.

23. Park, L. A., Ramamohanarao, K. Query Expansion Using a Collection Dependent Probabilistic Latent Semantic Thesaurus. Knowledge Discovery and Data Mining, 2007, 224-235. https://doi.org/10.1007/978-3-540-71701-0_24

24. Qiu, Y., Frei, H. P. Concept Based Query Expansion. Proceedings of the 16th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, 1993, 160-169. https://doi.org/10.1145/160688.160713

25. Richardson, R., Smeaton, A. Using WordNet in a Knowledge-Based Approach to Information Retrieval, CA-0395, 1995.

26. Shen, X., Tan, B., Zhai, C. Context-Sensitive Information Retrieval Using Implicit Feedback. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development, 2005, 43-50. https://doi.org/10.1145/1076034.1076045

27. Siddiqua, U. A., Ahsan, T., Chy, A. N. Combining A Rule-Based Classifier with Ensemble of Feature Sets and Machine Learning Techniques for Sentiment Analysis on Microblog. International Conference on Computer and Information Technology, 2017, 304-309. https://doi.org/10.1109/ICCITECHN.2016.7860214

28. Voorhees, E. M. Query Expansion Using Lexical-Semantic Relations. International ACM SIGIR Conference on Research and Development in Information Retrieval, 1994, 61-69. https://doi.org/10.1007/978-1-4471-2099-5_7

29. Wang, S. Y., Liao, W. S., Hsieh, L. C., Chen, Y. Y., Hsu, W. H. Learning by Expansion: Exploiting Social Media for Image Classification with Few Training Examples. Neurocomputing, 2012, 95, 117-125. https://doi.org/10.1016/j.neucom.2011.05.043

30. Weerkamp, W. Finding People and Their Utterances in Social Media. Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2010, 918-918. https://doi.org/10.1145/1835449.1835691

31. Wei, J., Liao, X., Zheng, H., Chen, G., Cheng, X. Learning from Context: A Mutual Reinforcement Model for Chinese Microblog Opinion Retrieval. Frontiers of Computer Science, 2018(1-2), 1-11.

32. Xu, J., Croft, W. B. Query Expansion Using Local and Global Document Analysis. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1996, 4-11. https://doi.org/10.1145/243199.243202

33. Zhang, D., Nie, L., Luan, H., Tan, K. L., Chua, T. S., Shen, H. T. Compact Indexing and Judicious Searching for Billion-Scale Microblog Retrieval. ACM Transactions on Information Systems, 2017, 35(3), 1-24. https://doi.org/10.1145/3052771

34. Zhou, D., Lawless, S., Wade, V. Improving Search via Personalized Query Expansion Using Social Media. Information Retrieval, 2012, 15(3-4), 218-242. https://doi.org/10.1007/s10791-012-9191-2

35. Zingla, M. A., Latiri, C., Mulhem, P., Berrut, C., Slimani, Y. Hybrid Query Expansion Model for Text and Microblog Information Retrieval. Information Retrieval Journal, 2018(5), 1-31. https://doi.org/10.1007/s10791-017-9326-6