International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.12 (2015), pp.341-354
http://dx.doi.org/10.14257/ijmue.2015.10.12.33

Query Subtopic Mining by Combining Multiple Semantics

Lizhen Liu, Wenbin Xu, Wei Song, HanshiWang and Chao Du

Information and Engineering College, Capital Normal University, Beijing 100048, China
wsong@cnu.edu.cn

Abstract

Query subtopic mining aims to find aspects to represent people's potential intents for a query. Clustering query reformulations is the most common approach for subtopic mining these days. However, there are some challenges that the existing approaches have to face in finding both relevant and diverse subtopics, such as term mismatch and data sparseness. In this paper, a novel semantic representations for query subtopics is introduced, which including phrase embedding representation and query category distributional representation, to solve those problems mentioned above. Furthermore, we also combine multiple semantic representations into vector space model and compute a joint similarity for clustering query reformulations. To evaluate our theory an experiment is conducted on a public dataset offered by NTCIR subtopic mining project, the experimental results show that phrase embedding representation is the most effective representation while combining multiple semantics benefits short text clustering and improves the performance of query subtopic mining.

Keywords: *subtopic mining, query understanding, semantic representation, information retrieval*

1. Introduction

Currently, people are highly relying on search engines to find information they are interested in on the internet. Most people used to express their search intents through short keywords queries, primarily because current search engines deal with such queries better than queries expressed in natural language. In some cases, it is difficult for people to interpret their complex information needs with appropriate keywords. Therefore, short keyword queries are usually ambiguous or multi-faceted. For example, a simple query "eclipse" is not only likely to be a simple word which literally refers to an astronomical natural phenomenon, but also has other interpretations such as a song, a film, publications, even a software development platform named "eclipse" and so on. Due to the above reason, the information which the people get from the search engines may not match the intention of the information seeker. To understand peoples' intents behind queries has become a key challenge for information retrieval, especially in the investigation of search results diversification and personalization [1-3].

Recently, query subtopic mining has gained some attention in the research community [4-5]. It aims to automatically discover a set of subtopics from a given query and each of them is defined as a short string, which specifies and disambiguates the search intent of the original query [6]. Query subtopics could be extracted from different resources including query suggestion [7-9], top-ranked search results [3,10-12], and external resource(*e.g.* Wikipedia) [13-14]. Nevertheless, most existing approaches rely on clustering short texts, such as query reformulations, for discovering potential query subtopics. In addition, the representations of subtopic candidates are usually based on bag of words or co-occurency in search results or query logs. Such simple representations suffer from word mismatch and data sparseness problems. The same query intent could be described using different words, for example, “eclipse software” and “eclipse platform”. If we cannot merge them properly, the resulting query intent would be very redundant. It is not convenient to show them to users and perhaps could even harm users’ experiences. Co-occurrence in top ranked results is an optional solution for solving word mismatch problem. However, according to our statistics, the average number of documents that one aspect phrase occurs in is only10. This means that co-occurrence statistic information based on local data should be very sparse as well.

In this paper, we attempt to overcome the above problems by incorporating rich global semantic information at different levels to represent short texts. A novel semantic representation for query subtopics is introduced which including continuous distributed representation and query category distributional representation. Both of these representations are learned based on global information. They carry different levels of semantics. Continuous distributed representations are learned from large scale texts in an unsupervised way and have fine-grained syntactic and co-occurrence information. In contrast, query category distributional representation captures high level topic information. Since these representations describe query subtopics from different views, it is useful to consider all information together. Therefore, we combine these semantics in a vector space model and compute a joint similarity for clustering query reformulations. We conducted experiments on the NTCIR subtopic mining Task collections. The experiments show that our approach benefits short text clustering and improves the performance of query subtopic mining.

To summarize, the major contributions of our work are described as follows:

1) Incorporate novel semantic representations of query subtopics: continuous distributed representation and category distributional representation. These representations carry global semantic relatedness at different levels and complement traditional representations based on hard word match and local co-occurrence.

2) Conducted experiments on NTCIR dataset for comparing the effectiveness of different semantic representations and their combinations. We find that the continuous distributed representation, which is got via semantic composition method, is the best single solution for query subtopic mining. We combine different semantic representations in a vector space model for computing joint similarity between texts. And find that the combination of continuous distributed representation and category distributional representation achieves the best performance for query subtopic mining and outperforms a state of the art method based on local information.

The remaining parts of this paper are organized as follows. In Section 2, we introduce and compare the related work. In Section 3, we introduce the proposed method in detail. The experimental setup and the evaluation of our approach are presented in Section 4 and Section 5. We conclude this work in Section 6.

2. Related Work

This Section will briefly introduce the related work on query subtopic mining and short text processing.

2.1. Query Subtopic Mining

It has long been recognized that queries often have multiple senses or facets, and that only considering relevance cannot provide satisfactory information retrieval. Hence, researchers have tackled various problems related to recognizing the intents behind the query. Those problems also can be called subtopic mining. The existing research about subtopic mining can be generally divided into the following categories:

Search result analysis based methods. Search engines (SE) return search results according to the keywords giving by the user. The result pages contain two parts (titles and snippets), that bring us very relevant information of the initial query and some important keywords. With this information, one can extract the subtopic implied in the search result. Clustering algorithms with this information presented a series of clusters that represent subtopics of the initial query. Cluster based method has been applied several times in previous studies [1, 5, 10-11]. Zamir *et al.* (1998, 1999) use the suffix tree clustering method to find implicated subtopics [15]. Hearst *et al.* (1995) applied decentralized collection techniques in order to organize and browse documents interactively [16]. Chen *et al.* (2000) conducted the clustering algorithm using predefined categories and extract phrase representative subtopics[17]. As an extension of this method, Wang *et al.* (2007) selected the result pages not only considering the original query but also related ones.[18]. More recently, Guo *et al.*(2011) employed a regularized topic model to automatically learn the potential intents by using both the snippets and the regularization from query co-clicks, then applied the technique to query recommendation [19]. Damien *et al.* (2013) generated a subtopic list with fusion resources covering both information seeker and information provider aspects, and an efficient Bayesian optimization approach was proposed to improve the performances of the resources selection[3].Head-Modifier relation, score calculations and word co-occurrence information has also been used in other works [2,6].

Query logs analysis based methods. Search query log data records user behaviors in a search session at a certain period of time, contain the information like search keywords, and click-through data. The rich information retained in the search query log provides an opportunity to discover subtopics embedded in a search query. Previous studies mined subtopics for the initial query by analysis diffident aspects of the query log data, such as the click-through data, query session data, search keywords, and so on. Beeferman (2000)viewed the query log as a query-URL bipartite graph, then conducted clustering algorithms based on this bipartite graph, each obtained cluster coveting multiple queries is a subtopic[11]. Li (2008) investigated a completely orthogonal approach based on regularized click-through graphs. They inferred the class memberships of unlabeled queries from those labeled ones according to their proximities in a click graph. Moreover, they regularized the learning with click graphs by content-based classification to avoid propagating erroneous labels [20]. Luo *et al.* (2014) continued to expand those methods, they analyzed query log data to establish bipartite graphs, then generated large-scale candidate sets by a random walk method. After that, they combined abstract click information to fully consider the comprehensiveness of the subtopic set and machine learning methods were used to eliminate noise points in order to improve the accuracy of the subtopic excavation[21]. Strohmaier *et al.* (2009) filtered the noise phrase by click-through data, and then proposed a preliminary algorithm based on the historic query log data to measure the accuracy of the score for subtopic ranking [22]. Hu *et al.* (2012) showed two interesting phenomena of user behavior ("one subtopic per search" and "subtopic clarification by keyword"). For effectively leveraging the two phenomena, a clustering algorithm aimed to automatically mine the major subtopics was proposed [23]. Meng *et al.* presents a new algorithm of web queries clustering using user click information in the query logs and applies it into query expansion [24]. Wang *et al.* collected the information from related URLs in the query log to generate the subtopic [25].

Knowledge base based methods. External knowledge sources such as Wikipedia, Word Net are also used in the subtopic mining method. Hu *et al.* (2009) understanding the intent behind the query by mapping the query into the Wikipedia intents representation space [13]. Zheng *et al.* (2013) integrated use of the structured and unstructured data to extract valid subtopics. To complete the subtopics, combinations of diffident data were used [26]. In Wang's category-based method, the external resources were used to recognize the subtopics for the given query [25]. And Ren *et al.* (2014) introduced a heterogeneous graph to mine subtopics and enhance the subtopic's quality by taking advantage of a combination with Wikipedia concepts [27].

2.2. Short Text Processing

This research focuses on short text clustering based on rich semantic representations. We try to incorporate more novel semantic representations and combine them together for query subtopic mining. The existing work could be easily integrated into our framework by providing query subtopic candidates from different resources. For a given short text, one can model it more precisely by deriving latent topics from existing large corpus with topic models such as latent Dirichl *et al.,* location (LDA). Likely, Chen *et al.,* moved forward along this direction and proposed a method to leverage topics at multiple granularities [28]. In other ways, a search and vote strategy with search results was used in labeling the query candidates in Sun's paper [29]. Meng *et al.,* presents an effective algorithm for semantic similarity metric of word pairs [30] and a new similar queries metric for query suggestions, then measure similar queries from the level of semantic level [31]. Song *et al.,* constructed a proximity matrix based on word similarity and then used it to convert the raw text into vectors. The results of text clustering experiments show that their method improves the performance of short text processing [5]. Compared with these short text processing methods, our novel semantic representations have the same effect and have improved upon clustering performance.

3. Proposed Method

3.1. Framework

As showing in Figure 1, the framework of the proposed method is divided into three parts, Aspect Phrase Extraction, Semantic Representations and Clustering & Subtopic Mining. In the first part, the related queries of the topic (original query) are extracted from the query log and denote the query with multi-word phrase. Then, novel semantic representations and combinations are used to represent the query aspect phrases for distinguishing the semantics of words, such as, the synonymous with special-shapes or words with different meanings. Finally, we adopt the clustering approach to generate the subtopics and each cluster denotes one subtopic of the initial query.

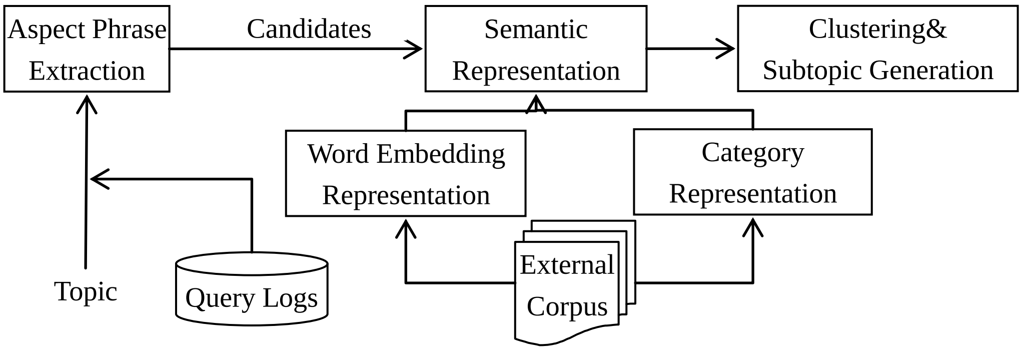

Figure 1. The Flowchart of Proposed Approach

Terminology notations. Queries tend to be composed of three or four keywords [21]. In the study of Baeza-Yates *et al.,* they found that a query can be expressed with a combination of Query and Aspect Phrase [9]. Here we only consider the form of the string beginning with or ending with the Query. The candidates we extracted from the querylogs, will be transformed into those reformulations. Means of Query and related names are introduced as follows:

- Query: the user issued keywords which represent (vague) information users needed, noted as q, *e.g*. "canon camera".
- Query reformulation: a short text span which contains the query as a substring. For a query, we would extract a set of query reformulations $R = \{r\}$, where each $r \in R$ is a reformation of query q. For simplicity, we only consider the text spans that begin with or end with the query q, *e.g*. "canon camera 5D review" and "canon camera price" are two reformulations.
- Aspect Phrase: the text surrounding the original query in a query reformulation form an aspect phrase. By this definition, a query reformulation *r* could be represented as $r = q + a$ or $r = a + q$,where a is a query aspect phrase, such as the reformulation "canon camera 5D review", the "5D review" is an aspect phrase.

3.2. Aspect Phrase Extraction

Aspect phrase extraction is the base component of our framework, the quality of the phrase directly affects the subtopic's quality. Dealing with the candidates extracted from query logs, a multi-word reformulation is adopted to express aspect phrase, and the words behind or after the query are considered separately, the candidates should be separated by the query when the query appears neither in the beginning nor the end of the candidate string. The co-occurrence frequency of the words is an important indicator to construct the reformulations, and then the following approach based on Term Frequency (TF) was chosen to calculate the affinity of words:

$$Affinity(w_i, w_j) = \left|\frac{tf(w_i,w_j)}{tf(w_i)} - \frac{tf(w_i,w_j)}{tf(w_j)}\right| \tag{1}$$

Where $tf(w_1, w_2)$denotes the co-occurrence frequency of the word w_1 and w_2 in all documents we use in this paper. And $tf(w_i)$ is the frequency of the word w_i. The words with $Affinity(w_i, w_j) < \xi$ can be joined as a multi-word representation, the others are removed as noise, the ξ is an experimentally defined threshold, in this paper $\xi = 0.2$.

Obviously, in the generated multi-words reformulation, each word in this phrase plays a different role in understanding the query (*e.g*. in the query "canon camera 5D review", the "5D" has the important information than "review"). In this scenario, a weight model was assigned to detect the critical degree of one word on understanding the query, which is calculated as follows:

$$weight(w) = (0.5 + 0.5\ tf(w, q)/m(w)) * log\ N/tf(w, q) \tag{2}$$

Where $tf(w, q)$ denotes the frequency of the word w co-occurred with the query q in all aspect phrases, $m(w)$ is the frequency of the word w in the whole collection and N denotes the size of the collection.

3.3. Aspect Phrase Representation

Representing short texts plays an important role for measuring relatedness between them. Different representations capture different assumptions behind. Two simple representations will be introduced first: word space representation and document space representation, which are based on vector space model and co-occurrence respectively.

Term space representation (TSR).This is the most commonly used representation for representing documents. The dimension of this representation is the size of the vocabulary. The weights in this vector are the frequencies of terms occurring in the aspect phrase. This representation actually reflects the term match information. If two aspect phrases have more common terms, they have a higher similarity.

Document space representation (DSR).The co-occurrence information in top ranked search results are also widely used for measuring relatedness between short texts. Here, we represent each aspect phrase by recording which documents it occurs within. The dimension of this representation is the number of top retrieved documents. If one document contains all words of the aspect phrase, the corresponding field is set to 1. If two aspect phrases co-occur in more documents, they are more similar to each other.

In addition to the above representations, we introduce two novel word representations as follows:

Phrase embedding representation (PER). This representation is come from deep learning. An important product of deep learning is to represent the words by distributed continuous vectors, noted as embedding. One of the advantages of this method is that the continuous vector representations could be used to measure the semantic relatedness between words which vary in lexical surface strings, such as “teacher” and “professor”. In order to overcome the term-match problems, when processing short texts, we borrow the idea of using continuous vector representations and attempt to represent query subtopics in this way. Aspect phrases are usually multi-word expressions; we compute the semantic vectors of phrases by using a linear semantic composition method based on individual word vectors. The general form is shown in Equation 3.

$$phrVec(w_1, \dots, w_n) = \sum_{i=0}^{n} \alpha_i vec(w_i) \quad (3)$$

Where $vec(w)$ is used to represent the vector of word w, and $phrVec(w_1, \dots, w_n)$ is used to represent the vector representation of a phrase with n words, and α is a weighting parameter for the *i*-th word in the phrases, subject to $\sum_{i=0}^{n} \alpha_i = 1$.In this way, a phrase is represented using a vector which has the same dimensions with a word vector. This is convenient to compute similarities between phrases with different lengths.

In general, words in a phrase are not equally important to indicate its meaning. Therefore, we aim to assign more weights to words in order to keep the composite vector close to its real position in vector space. We use a simple metric to measure the importance of a word in a phrase. We assume that the more times a word *w* co-occur with the query q, the more important it is. Formally, the weight α_i for word w is defined as $\alpha_i = co(w_i, q)/\sum co(w_j, q)$, where $co(*,*)$ is a function representing the co-occurrence times of two strings in query logs.

Category distributional representation (CDR). The category information provides evidence for assigning aspect phrases into subtopics. Therefore, a novel representation is introduced. Each aspect phrase is represented by a distributional category vector which indicates its probability over a given topic taxonomy. We achieve this by training a query classifier.

There are various ways to classify a query. In Song’s work, the query were classified into three categories depend on the query ambiguity [32]. Pre-defined categories also are effectiveness for query classifier [33]. Here, we learn a query classifier via query log data and public available web directories (The detail could be found in Section 4).This classifier could assign a distributional category vector to any keyword queries.

In our search, a trained classifier is used to classify the concatenation of each aspect phrase and the original query, since the aspect phrase alone cannot determine its category. Therefore, the category vector of aspect $a$actually is the category distribution of $q + a$.

Integrated representation. Now, there are multiple representations of aspect phrases in the framework we proposed, which reflect the characteristics of the aspect phrases from different views. It is necessary to combine them together to jointly represent the aspect phrases. Here, we adopt a simple method by concatenating multiple representations into an integrated vector. In this way, we could manipulate all semantic representations in the vector space model. Of course, we can use semantic fusion and metric learning approaches to learn a better joint representation. We leave it as future work.

3.4. Subtopic Generation

For the aspect phrases extracted and represented with the representations introduced in previous Section, the K-Means clustering algorithm was used to process the aspect phrases. Clustering algorithms have been used in other works and perform perfectly. The distance function we used in the clustering algorithm was defined as follow:

$$dis(q_i, q_j) = 1 - \frac{q_i \cdot q_j}{\|q_i\| \|q_j\|} \tag{4}$$

Where q_i denote the vectors of the query i represented by the model discuss in Section 3.3, and $\|\cdot\|$denotes the norm of the vector. We conduct clustering on the phrases that were extracted from the previous Section and are represented with multiple semantic representations. For each topic, clustering method is performed on a single representation or combination of a few representations. Each cluster generated by the clustering algorithm indicated one subtopic of the initial query.

4. Experimental Setup

4.1. Dataset

To evaluate the performance of the proposed approach and the multiple representations, our experiments were carried out on the public Chinese data collection offered by the NTCIR-9 [5] intent task which contains 100 queries/topics. All queries are median-frequency which has been sampled from both Sogou and Bing search logs. And the number of three types of queries (ambiguous, broad and clear) in this collection is approximately equal. In this collection, topics have been artificially annotated with the subtopics and the importance by the task assessors. And in order to complete the candidates for each query in this data collection, the SogouQ dataset as an additional resource was used to extract related query strings, which contains the information in a search session such as search keywords, rank score and the click-through data. Combinations of the related strings and the data collection constitute the candidate set for a given topic in our experiment. For each original query, we collect the top 200 results from Bing search engine as the search result document.

4.2. Evaluating Metrics

We evaluate our method and the baseline performance with the following industry standard evaluation metrics: intent recall (also called “I-rec”), D-nDCG and D#-nDCG [34]. The intent recall metric is a measure of subtopic diversity, and the D-nDCG measures the overall relevance across subtopic sets. D#-nDCG is a linear combination of previous metrics which consider both relevant performance and the subtopic coverage. The document cutoff for those metrics is set to 10 while the reliability is validated elsewhere [5]. With those metrics, we aim to validate the effectiveness of semantic representations and combinations.

4.3. Query Classification

At the first, a set of categorized websites is collected from ODP for Chinese[1], YahooCN[2] and Baidu sites[3]. Then the Section and level categories of these Chinese were assigned directories into the taxonomy of KDD cup 2005 manually. Some categories in KDD cup 2005 taxonomy was merged in this experiment, since these categories are considered to be the same in Chinese directories. Finally, taxonomies with 59 categories and about 56000 distinct categorized websites were obtained. For each URL, its category was determined by looking up the used taxonomy. If there is no matched entry, prune the postfix of the URL and look up again. This process is repeated until a match is found or a miss is detected. The queries in the query log were used in this experiment when its URL could be classified into only one category. About 23.1 million categorized queries were got, and were used to train a query classifier by the lib-linear toolkit [35]. The performance of the query classifier on a dataset of 2000 queries was evaluated and the accuracy is 91% [36].

4.4. Baselines

To evaluate the performance of the proposed multiple semantic representations, we compare them with some classical representations and clustered them based on subtopic mining approach introduced as follows:

- Term Space Representations (TSR). TSR is the standard text representation based on the terms’ occurrence. The semantic similarity based on this representation depends on exact matches between terms.
- Document Space Representations (DSR). The aspect phrases are represented based on the occurrences in search result documents. The semantic similarity based on this representation depends on the co-occurrence of aspect phrases in search results.
- Clustering Query Aspects Reformulations(CAR) [36]. An approach generates aspect reformulations from anchor text and Microsoft Web N-Gram Services, and using a clustering algorithm to find the subtopics for the given query. Here we compare this approach with our final system.

5. Experiment Results Analysis

5.1. Comparison of Semantic Representations

In this Section, the performance of different semantic representations and their combinations are demonstrated. We aim to answer the following research questions: 1) whether the new introduced semantic representations of aspect phrases outperform

[1] http://www.dmoz.org/World/Chinese_Simplified/
[2] http://site.yahoo.com.cn/
[3] http://site.baidu.com/

traditional representations on query reformulation clustering. 2) whether the combination of multiple semantics benefit query subtopic mining.

Table 1. Performance Comparison of Diffident Representations on Official Evaluation Metric (*K*=15)

	I-rec@10	D-nDCG@10	D#-nDCG@10
TSR	0.4127	0.6881	0.5868
DSR	0.4054	0.6385	0.5678
CDR	0.4701	0.6021	0.5613
PER	**0.5027**	**0.7044**	**0.6062**

Table 1 Shows the effectiveness of individual semantic representations when the number of clusters is set to 15. When k is set to other numbers, the performances have similar trends. CDR represents the category distributional representation, while PER represents the phrase embedding representation. We can see that both CDR and PER outperform TSR and DSR largely, especially on I-rec@10. This indicates that the new introduced representations could effectively find more diverse query subtopics. It is easy to understand why this is the case. The global information benefits from connecting aspect phrases with similar semantics but with different lexicons. However, CDR performs worse on D-nDCG@10. The reason is that the category information might be too coarse to represent fine-grained subtopics. Some fine-grained subtopics are clustered together. PER achieves best performance on both diversity and relevance.

The performance of PER with diffident number of clusters is shown in the Figure 2. In our experiment, four diffident values from 10 to 30 are chosen. The best performance arises when the value of *k* is 10, which is very close to the average number of subtopics put forward by the report on the NTCIR-9 [5].

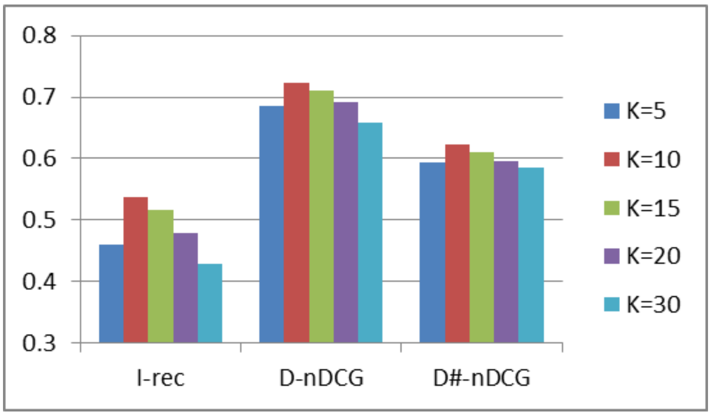

Figure 2. Performance of PER with Diffident Clusters (*k*=5, 10, 15, 20 and 30)

5.2. Comparison of Semantic Composition Methods

Here, further comparison on different semantic composition methods for PER is demonstrated. In this work, a query-related weighting method (QR) is proposed to improve the performance of the semantic composition, and then compare with commonly used average weighting method (AVE) and IDF weighting method. Figure 3 shows the performance using diffident term weight models. It can be clearly observed that the combination with average weight model has the worst performance, and the IDF weight model has a small degree improvement over the average one. The QR method is outstanding among those approaches and has the best performance on all evaluation metrics.

5.3. Combination of Semantic Representations

We further examine the combinations of different semantic representations. Table 2 shows the performance when we add other representations and combine them with the best single representation PER. We can see that combining two semantic representations always improves the performance compared with using single representations. The combination of CDR and PER achieves the best performance among all combinations. The reason is that CDR carries high level semantic information so that it is useful for dealing with ambiguous queries. Because PER represents more fine-grained information, combining these two semantics provide evidence from different semantic levels and benefits all types of queries. We guess the reason might be that we simply concentrate all representations together but don't assign different weights according to their importance. Therefore, incorporating less effective representations decrease the performance. It is necessary to use more sophisticated techniques to combine these representations better.

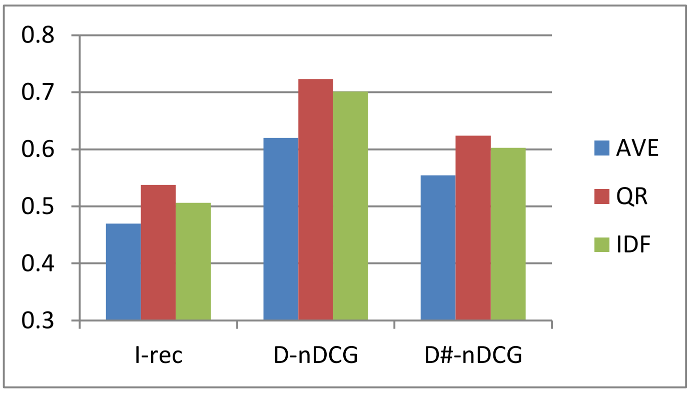

Figure 3. Performance of PER with Diffident Weight Models

5.3. Combination of Semantic Representations

We further examine the combinations of different semantic representations. Table 2 shows the performance when we add other representations and combine them with the best single representation PER. We can see that combining two semantic representations always improves the performance compared with using single representations. The combination of CDR and PER achieves the best performance among all combinations. The reason is that CDR carries high level semantic information so that it is useful for dealing with ambiguous queries. Because PER represents more fine-grained information, combining these two semantics provide evidence from different semantic levels and benefits all types of queries. We guess the reason might be that we simply concentrate all representations together but don't assign different weights according to their importance. Therefore, incorporating less effective representations decrease the performance. It is necessary to use more sophisticated techniques to combine these representations better.

Table 2. Performance of Representation Combinations on Official Evaluation Metric (*K*=15)

TSR+PER	0.5231	0.7155	0.6117
DSR+PER	0.5173	0.7124	0.6093
CVR+PER	**0.5363**	**0.7230**	**0.6234**
ALL	0.4424	0.6765	0.5808
CAR	0.5146	0.6835	0.5975

6. Conclusions and Future Work

This paper presents a query subtopic mining approach by exploiting multiple semantic representations. Two novel representations are introduced in this research: the phrase embedding representation (PER) and category distributional representation (CDR). The two representations bring in global semantic information in different levels. The experimental results show that PER is the best single representation for clustering based query subtopic mining. The combination of PER and CDR further improves the performance and outperforms a state-of-the-art approach.

There are several issues we want to further explore to enhance our current work. An improved method for joint representation learning will be our next target, as it plays an important role in combining multiple semantic representations. We also will attempt to apply our proposed subtopics mine framework in the search result of diversification applications.

Acknowledgment

This work was supported in part by National Science Foundation of China under Grants No. 61303105 and 61402304; the Humanity & Social Science general project of Ministry of Education under Grants No.14YJAZH046; the Beijing Natural Science Foundation under Grants No. 4154065; the Beijing Educational Committee Science and Technology Development Planned under Grants No.KM201410028017 and Beijing Key Disciplines of Computer Application Technology.

References

[1] H. T. Yu and F. Ren, "Subtopic Mining via Modifier Graph Clustering", in: V. Tseng, T. Ho, Z.-H. Zhou, A.P. Chen, H.-Y. Kao (Eds.), Advances in Knowledge Discovery and Data Mining, Springer International Publishing, **(2014)**, pp. 337-347.

[2] S. J. Kim and J. H. Lee, "Method of Mining Subtopics Using Dependency Structure and Anchor Texts", in: L. Calderón-Benavides, C. González-Caro, E. Chávez, N. Ziviani (Eds.), String Processing and Information Retrieval, Springer Berlin Heidelberg, **(2012)**, pp. 277-283.

[3] A. Damien, M. Zhang, Y. Liu and S. Ma, "Improve Web Search Diversification with Intent Subtopic Mining", in: G. Zhou, J. Li, D. Zhao, Y. Feng (Eds.), Natural Language Processing and Chinese Computing, Springer Berlin Heidelberg, **(2013)**, pp. 322-333.

[4] T. Sakai and R. Song, "Diversified search evaluation: lessons from the NTCIR-9 INTENT task", Information Retrieval, vol. 16, **(2013)**, pp. 504-529.

[5] R. Song, M. Zhang, T. Sakai, M. P. Kato, Y. Liu, M. Sugimoto, Q. Wang and N. Orii, "Overview of the NTCIR-9 INTENT Task", Proceedings of NTCIR-9, **(2011)**, pp. 82-105.

[6] S. J. Kim and J. H. Lee, "Subtopic Mining Based on Head-Modifier Relation and Co-occurrence of Intents Using Web Documents", in: P. Forner, H. Müller, R. Paredes, P. Rosso, B. Stein (Eds.), Information Access Evaluation. Multilinguality, Multimodality, and Visualization, Springer Berlin Heidelberg, **(2013)**, pp. 179-191.

[7] R. L. T. Santos, C. Macdonald and I. Ounis, "Exploiting query reformulations for web search result diversification", Proceedings of the 19th international conference on World wide web, ACM, Raleigh, North Carolina, USA, **(2010)**, pp. 881-890.

[8] C. K. Huang, L. F. Chien and Y. J. Oyang, "Relevant term suggestion in interactive web search based on

contextual information in query session logs", Journal of the American Society for Information Science and Technology, vol. 54, **(2003)**, pp. 638-649.

[9] R. Baeza-Yates, C. Hurtado and M. Mendoza, "Query recommendation using query logs in search engines", Current Trends in Database Technology-EDBT 2004 Workshops, Springer, **(2005)**, pp. 588-596.

[10] H. J. Zeng, Q. C. He, Z. Chen, W. Y. Ma and J. Ma, "Learning to cluster web search results", Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, **(2004)**, pp. 210-217.

[11] D. Beeferman and A. Berger, "Agglomerative clustering of a search engine query log", Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, Boston, Massachusetts, USA, **(2000)**, pp. 407-416.

[12] J. Xu and W. B. Croft, "Query expansion using local and global document analysis", Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, **(1996)**, pp. 4-11.

[13] J. Hu, G. Wang, F. Lochovsky, J.-t. Sun and Z. Chen, "Understanding user's query intent with wikipedia", Proceedings of the 18th international conference on World wide web, ACM, **(2009)**, pp. 471-480.

[14] J. Kamps, R. Kaptein and M. Koolen, "Using anchor text, spam filtering and wikipedia for web search and entity ranking", **(2010)**.

[15] O. Zamir and O. Etzioni, "Web document clustering: A feasibility demonstration", Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, ACM, **(1998)**, pp. 46-54.

[16] M. A. Hearst and J. O. Pedersen, "Reexamining the cluster hypothesis: scatter/gather on retrieval results", Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, **(1996)**, pp. 76-84.

[17] H. Chen and S. Dumais, "Bringing order to the web: Automatically categorizing search results", Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, **(2000)**, pp. 145-152.

[18] X. Wang and C. Zhai, "Learn from web search logs to organize search results", Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, **(2007)**, pp. 87-94.

[19] J. Guo, X. Cheng, G. Xu and X. Zhu, "Intent-aware query similarity", Proceedings of the 20th ACM international conference on Information and knowledge management, ACM, Glasgow, Scotland, UK, **(2011)**, pp. 259-268.

[20] X. Li, Y. Y. Wang and A. Acero, "Learning query intent from regularized click graphs", Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, ACM, Singapore, Singapore, **(2008)**, pp. 339-346.

[21] C. Luo and Y. Liu, "Query recommendation research based on the user intent recognition", Journal of Chinese Information Processing, **(2014)**, pp. 64-72.

[22] M. Strohmaier, M. Kr, #246, ll and C. K, rner, "Intentional query suggestion: making user goals more explicit during search", Proceedings of the 2009 workshop on Web Search Click Data, ACM, Barcelona, Spain, **(2009)**, pp. 68-74.

[23] Y. Hu, Y. Qian, H. Li, D. Jiang, J. Pei and Q. Zheng, "Mining query subtopics from search log data", Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, ACM, Portland, Oregon, USA, **(2012)**, pp. 305-314.

[24] L. Meng, R. Huang and J. Gu, "A new algorithm of web queries clustering using user feedback", International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 6, **(2013)**, pp. 401-410.

[25] C. J. Wang, Y. W. Lin, M. F. Tsai and H.-H. Chen, "Mining subtopics from different aspects for diversifying search results", Information Retrieval, vol. 16, **(2012)**, pp. 452-483.

[26] W. Zheng, H. Fang, C. Yao and M. Wang, "Leveraging integrated information to extract query subtopics for search result diversification", Information Retrieval, vol. 17, **(2014)**, pp. 52-73.

[27] X. Ren, Y. Wang, X. Yu, J. Yan, Z. Chen and J. Han, "Heterogeneous graph-based intent learning with queries", web pages and Wikipedia concepts, **(2014)**, pp. 23-32.

[28] M. Chen, X. Jin and D. Shen, "Short text classification improved by learning multi-granularity topics", Proceedings of the Twenty-Sectionond international joint conference on Artificial Intelligence - Volume Volume Three, AAAI Press, Barcelona, Catalonia, Spain, **(2011)**, pp. 1776-1781.

[29] A. Sun, "Short text classification using very few words", Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, ACM, Portland, Oregon, USA, **(2012)**, pp. 1145-1146.

[30] L. Meng, R. Huang and J. Gu, "An effective algorithm for semantic similarity metric of word pairs", International Journal of Multimedia and Ubiquitous Engineering, vol. 8, **(2013)**, pp. 1-12.

[31] L. Meng, R. Huang and J. Gu, "Query suggestion based on theme and context", International Journal of u- and e- Service, Science and Technology, vol. 7, **(2014)**, pp. 263-276.

[32] R. Song, Z. Luo, J.-Y. Nie, Y. Yu and H.-W. Hon, "Identification of ambiguous queries in web search", Information Processing & Management, vol. 45, **(2009)**, pp. 216-229.

[33] Y. Yue and T. Joachims, "Predicting diverse subsets using structural SVMs", Proceedings of the 25th international conference on Machine learning, ACM, Helsinki, Finland, **(2008)**, pp. 1224-1231.

[34] T. Sakai and R. Song, "Evaluating diversified search results using per-intent graded relevance", Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, ACM, Beijing, China, **(2011)**, pp. 1043-1052.

[35] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang and C.-J. Lin, "LIBLINEAR: A Library for Large Linear Classification", J. Mach. Learn. Res., vol. 9, **(2008)**, pp. 1871-1874.

[36] V. Dang, X. Xue and W. B. Croft, "Inferring query aspects from reformulations using clustering", Proceedings of the 20th ACM international conference on Information and knowledge management, ACM, Glasgow, Scotland, UK, **(2011)**, pp. 2117-2120.

Authors

Prof. Liu Lizhen, received a PhD in Computer Application from the Beijing Institute of Technology, China. She is currently a Professor at the Capital Normal University. Her research interests include text mining, natural language processing, biological data mining, sentiment analysis and knowledge acquisition. She has published in journals and conferences like Knowledge and Information Systems, International Journal of Information & Computational Science, Journal of Software, Journal of Computers, IEEE World Congress on Intelligent Control and Automation, CSCWD.

Xu Wenbin, is a master’s candidate of Technology of Computer Application at Capital Normal University, His research interests including query understanding and search result diversification.

Dr. Song Wei, received a PhD in Computer Application from Harbin Institute of Technology, China. He is currently a lecturer at the Capital Normal University. His research interests include information retrieval, information extraction and natural language processing. His papers appear in SIGIR, WWW and Coling. Email:wsong@cnu.edu.cn

Dr. Wang Hanshi, received a PhD in Computer Application from the Beijing Institute of Technology, China. He is currently a lecturer at the Capital Normal University. His research interests focus on computational linguistics, natural language processing, biological data mining and natural language processing, especially unsupervised methods in the area. He has published his important work on the famous journal of Computational Linguistics (CL), and other international conferences.

Du Chao, received a master's degree in Computer Application from the Capital Normal University, China. He is currently a lecturer at the Capital Normal University. His research interests focus on Data mining, natural language processing, Sentiment Analysis and biological data mining.