KECIR at the NTCIR-10 INTENT Task

Cheng Guo
KERC, Shenyang Aerospace University, Shenyang, 110136, China
guocheng1987@163.com

Yu Bai
KERC, Shenyang Aerospace University, Shenyang, 110136, China
baiyu@sau.edu.cn

Jianxi Zheng
KERC, Shenyang Aerospace University, Shenyang, 110136, China
zhengjxkercir@163.com

Dongfeng Cai
KERC, Shenyang Aerospace University, Shenyang, 110136, China
caidf@vip.163.com

ABSTRACT

This paper describes the approaches and results of our system for the NTCIR-10 INTENT task. We present some methods for Subtopic Mining subtask and Document Ranking subtask. In the Subtopic Mining subtask, we employ a voting method to rank candidate subtopics and semantic resource HowNet was used to merge those candidate subtopics which may impact diversity. In the Document Ranking Subtask, we also employ a voting method based on the mined subtopics. In the Chinese subtopic mining, our best values of $I-rec@10$, $D-nDCG@10$ and $D\#-nDCG@10$ were separately 0.3743, 0.3965 and 0.3854. In the Document Ranking subtask, they were separately 0.6366, 0.3998 and 0.5182.

Team Name

KECIR

Subtasks

Chinese Subtopic Mining and Document Ranking

Keywords

INTENT, Subtopics Mining, Diversity, Document Ranking

1. INTRODUCTION

Currently, commercial search engines are still traditional keyword-based retrieval, so, the users have to convert the search questions into short query strings. It may lead to lose the users' intents. Today, mining users' underlying intents of a query is an interesting topic for both IR communities and commercial search engines [1]. However, how can we identify the user real query intent? NTCIR-10 INTENT task was conducted to explore the problem. The task consists of two subtasks: Subtopic Mining and Document Ranking.

A subtopic could be an interpretation of an ambiguous query or an aspect of a faceted query. For example, "Chocolate（巧克力）" may refer to Dove chocolate candies（德芙巧克力糖果）or LG chocolate mobile phones (LG巧克力系列手机). The aim of subtopic mining is to return a ranked and diversified list of possible "subtopic strings" that correspond to a query. In this paper, we present some approaches for subtopic mining, and to further improve result diversity, we employ a voting method to rank and merge candidate subtopics.

The Document Ranking Subtask further explores systems to diversify search results based on mined subtopics[2]. In this subtask, we rank the documents based on the mined subtopics.

2. RELATED WORK

Diversified IR tasks such as subtopic retrieval [3] have been discussed earlier. A closely related problem is investigated in the interactive track of TREC-10(2001), where the purpose is to study how an interactive retrieval system can best support user gathering information about the different aspects of a topic [4]. The NTCIR9(2011) evaluation workshop also launched a new task called INTENT to study the problem[1].

A state-of-the-art diversification approach is based on query suggestions from Web search engines (WSEs) [5]. And in previous work, many approaches were proposed by researchers to mine the wealth of information hidden in the query log [6]. Zhang and Lu [7] first find out the related queries from query logs, then group them into different clusters using a frequent term-set based clustering algorithm. Finally, the central query of each cluster is used to represent the subtopic of this cluster. Radlinski [8] proposed an approach for inferring query intents from reformulations and clicks. For an input query, the click and reformulation information are combined to identify a set of possibly related queries to construct an undirected graph. An edge is introduced between two queries if they were often clicked for the same documents. Finally, the random walk similarity is used to find intent cluster.

Diversity document ranking is also a hot topic in recent years. Researchers proposed kinds of approaches to satisfy the requirement of diversity. In [9, 10], the investigators used Scatter/Gather algorithm to cluster the top documents retrieved from a traditional information retrieval system. As studied in [11], supervised learning algorithms were used to extract meaningful phrases from the search result snippets, and these phrases are then used to group search results.

3. SUBTOPIC MINING

3.1 Candidate Subtopics Mining

For each topic, we first collect snippets from top1000 pages searched from SogouT which was the document collection for Chinese topics in NTCIR-10. Then, we extract text fragments containing all query words from snippets. Finally, the frequent sequence mining algorithms are conducted to search for candidate subtopics.

We use the key words vector to represent an original query, and suppose that:

1) A subtopic should be the most frequent sequence which contains the key words vector of original query.

2) The more a frequent sequence contains others the less likely it is to be selected as a subtopic. And on the contrary, the more a frequent sequence is contained in others the more likely it is to be selected as a subtopic.

A brief description of frequent sequence mining(FSM) algorithm is shown as follows:

---

for each fragment f_i **do**
  for each fragment $f_j (j > i)$ **do**
    Frequent sequence set $\leftarrow LCS(f_i, f_j)$;
  end for
end for
for each frequent sequence s_k **do**
  $Score(s_k)$;
end for
Rank frequent sequence by its score and get top 100 frequent sequences as candidate subtopics.

In every fragment filtered from top pages, we got a frequent sequence set by using LCS Algorithm. Then, we rank frequent sequences according to

$$Score(s_k) = rfs(s_k) * \log(N/(fs(s_k) + 1)),$$

where s_k refers to a frequent sequence, $rfs(s_k)$ is the number of frequent sequences which contained s_k and $fs(s_k)$ is the number of frequent sequences that s_k contained.

3.2 Subtopics Clustering and Ranking

To further improve result diversity, we cluster candidate subtopics so that candidates whose intents are similar in a cluster. In the rest of this chapter, we will discuss the methods we employed in the subtopic mining subtask.

The first method: the idea of this method is quite clear. we assume that the similarity among diversity subtopics should be minimized. So semantic resource HowNet[12] is used to compute semantic similarity. The API of HowNet provides an interface to compute the semantic similarity between two concepts. The similarity operations allowed in our system are:

Segment all the subtopic candidates from strings to words, and compute the similarity $Sim(phrase_a, phrase_i)$ between subtopic a and other subtopics. Then rank subtopics based on the similarity values in ascending order.

We define the semantic similarity between two phrases as:

$$Sim(phrase_a, phrase_b) = \frac{1}{n_a \times n_b} \sum_{i=1}^{n_a} \sum_{j=1}^{n_b} Similarity(c_i^a, c_j^b),$$

where n_a and n_b are the number of semantic concepts of $phrase_a$ and $phrase_b$ in HowNet respectively. c_k^x is the k^th concept of $phrase_x$.

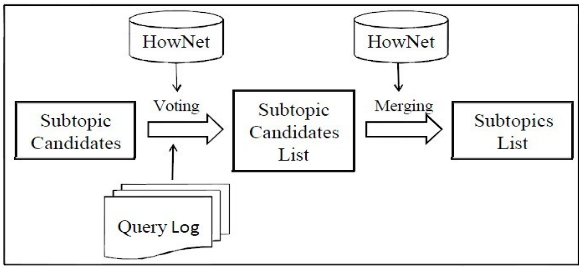

Figure 1: The Processing Mechanism of Subtopics Clustering and Ranking

The second method: considering that the frequent word in query log should suggest the most interesting aspect of a topic, another data resource we used was SogouQ log[13], which was provided by NTCIR organisers. This method can be described as three steps:

Step 1: segment words, for one subtopic, vote the words in the subtopic by query logs to find the greatest influence word. We reputed the word can represent the subtopic;

Step 2: find the DEF of the representative word in the HowNet, and extract the first sememe;

Step 3: merge the subtopics with the same sememe into one cluster, and rank the clusters by votes;

The third method: we use both query logs and HowNet. Experimental process of the method is in [5]. The processing mechanism of subtopics clustering and ranking is shown in Fig. 1.

At stage of voting, we calculate query log vote for each candidate subtopic according to

$$vote_i = \sum_{l=1}^{n} times_l \times sim(s_i, q_l),$$

where $times_l$ refers to the frequency of $query_l$ in log file; $sim(s_i, q_l)$ is the similarity between $subtopic_i$ and $query_l$, here we use semantic resource HowNet to compute semantic similarity.

A brief description of merging algorithm is shown as follows:

---

for each topic **do**
  K = 0.9;
  N = 30;
  Merging(subtopic_list[ ], K, N)
  for each subtopic $s_i \epsilon$subtopic_list[ ] **do**
    for each subtopic $s_j (j > i) \epsilon$subtopic_list[ ] **do**
      if $Sim(s_i, s_j) > K$ **then**
        subtopic_list[ ]← subtopic_list[ ]−subtopic s_j;
      end if
    end for
  end for
  while size of subtopic_list[ ] > N **do**
    Merging(subtopic_list[ ], K−0.1, N);
  end while
  Return subtopic_list[ ];
end for

---

4. DOCUMENT RANKING

In this section, we describe the experiment setup used for the document ranking subtask. For one document, consider that the coverage of subtopics largely reflect the diversity of the document, the approaches we use are all based on the mined subtopics.

4.1 Score Documents Directly

For this method, we use two strategies to accomplish the experiment.

1. Vote directly document by subtopics.We just consider the coverage of subtopics in documents.If a subtopic appears in the document, it will vote for this document.

2. This approach not only contains the subtopic coverage, also relates to the location information of the subtopics in the ranking list. The former the subtopic locates in the list, the higher score the document obtains. The sum of the subtopic scores are the final score of the document. We define the document score as:

$$Score(document) = \sum_{i=1}^{n} Score(subtopic_i)$$

$$Score(subtopic_i) = \frac{\sum subtopic_T}{Pos(subtopic_i)},$$

where $subtopic_i$ is the subtopic the document contains, and $Score(subtopic_i)$ is the score the $subtopic_i$ gets. $\sum subtopic_T$ is the sum of subtopics that topic T owns, and $Pos(subtopic_i)$ is the position of subtopic i in the topic T ranking list.

Finally, we rerank the documents by the scores.

4.2 Map Back to Snippets

As the subtopics are extracted from the document snippets, we map the subtopics back to snippets. We also use two methods to accomplish the experiment.

1. If one document snippet contains a subtopic, the document will get a vote. One subtopic only vote once for the same document.

2. Based on the above method, we add the location information of the subtopics. The $Score(snippet-document)$ is defined as:

$$Score(snippet-document) = \sum_{i=1}^{n} \frac{2^{score(subtopic_i)} - 1}{\log_2(1+i)},$$

where n refers the subtopic numbers of the snippet contains, and $score(subtopic_i)$ is the $subtopic_i$ original score in the ranking list.

Finally, we rerank the documents by the scores.

5. RESULTS AND ANALYSIS

The official effectiveness performance measures for the intent subtopic mining and document ranking tasks are: $I-rec$, which measures the proportion of intents covered by the documents in the search results list; $D-nDCG$, which uses a global gain to measure how relevant each document is to an intent, weighted by the importance of each intent; $D\#-nDCG$, which is a linear combination of $I-rec$ and $D-nDCG$ [14]. The measurement depths(i.e. number of top ranked items to be evaluated) were set to l = 10. According to the official report, run rankings and significance test results based on l = 30 are not so reliable, at least when the pool depth is 20 [15][16]. D#-nDCG was chosen as the primary evaluation measure by the task organisers.

5.1 Subtopic Mining Results

We submitted 4 runs for the subtopic mining subtask, and each of them was described in Table 1. For each method, we got top 30 strings in the list as the subtopics output.

Table 1: Description of comparative experiments

RunID	Description
KECIR-S-C-1B	The baseline method, use FSM Algorithm to get subtopics.
KECIR-S-C-2B	Base on KECIR-S-C-1B, use the first method in section 2.2
KECIR-S-C-3B	Base on KECIR-S-C-1B, employ the second method in section 2.2
KECIR-S-C-4B	Base on KECIR-S-C-1B, use the third method in section 2.2

Table 2 illustrates the revised experiment results of our runs for top 10 over 98 topics. From the results, our best run is KECIR-S-C-2B, the value of $D\# - nDCG$ 0.3854. Prove that semantic resources is useful for dealing with the subtopic clustering and ranking.

Table 2: The performance of 4 runs measured using top 10 results

RunID	$I-rec$	$D-nDCG$	$D\#-nDCG$
KECIR-S-C-1B	0.3341	0.3799	0.3570
KECIR-S-C-2B	0.3743	0.3965	0.3854
KECIR-S-C-3B	0.3001	0.3231	0.3116
KECIR-S-C-4B	0.2917	0.3085	0.3001

However, compared with the best results, our results are not satisfactory. Official query suggestion data which is provided by NTCIR organisers is only used for supplement when the ranking list contains less than 30 subtopics. Analyzing the suggestion data set, we find the data contains many more answers. However, according to our four results, a very small amount of suggestion is used for supplement. Furthermore, the measurement depths were set to 10. So, it is may be an important cause for the unsatisfactory results. Moreover, As mentioned in [15], organisers intentionally include 23 navigational(nav) topics in the topic sets. A navigational query should require one answer or one website(different from informational(inf) queries), and therefore may not require diversification. But we do not discriminate whether the subject is navigational or not. The mean $D\#-nDCG$ results about the nav-topics and inf-topics can be shown in Fig. 2. It can be observed that nav-topics values reduce the all-topics mean $D\#-nDCG$ results. Maybe it is another cause for the results.

Analyzing the per-topic $I-rec$, $D-nDCG$ and $D\#nDCG$ performances, the maximum values are all from navigational

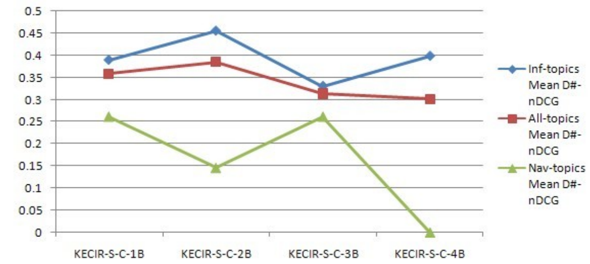

Figure 2: Inf-topics and Nav-topics Mean $D\# - nDCG$ Results

topics with only one intent. Recall that such topics illustrated in [15], $D - nDCG$ reduces to $nDCG$. As long as the nav-query term(s) appears at the first position in the result, the metric values of this result will be 1.0. In addition, As [15] mentioned, INTENT-2 organizers decided to select up to 9 intents per topic based on the votes, some intents in our results such as "金鱼手提包", "金鱼女包" are not contained in the "金鱼" topic answers. This led directly to the official metric values of topic "金鱼"(topic ID, 0202) in KECIR-S-C-1B are 0.

5.2 Document Ranking Results

We submit five runs in subtopic mining subtask of intent task. All of them are shown in Table 3. We get all the results based on the organisers' baseline non-diversified document ranking run.

Table 3: Description of comparative experiments

RunID	Description
KECIR-D-C-1B	Base on KECIR-S-C-1B result, use the first method in section 3.2.
KECIR-D-C-2B	Base on KECIR-S-C-2B result, use the method in section 3.1
KECIR-D-C-3B	Base on KECIR-S-C-2B result, employ the first method in section 3.2
KECIR-D-C-4B	Base on KECIR-S-C-3B result, use the first method in section 3.2
KECIR-D-C-5B	Base on KECIR-S-C-4B result, use the second method in section 3.2

The official mean $I - rec$, $D - nDCG$, $D\# - nDCG$ performances of our runs for top 10 over 97 topics are shown in Table 4. In addition, the mean performances according to the intent type-sensitive metrics $DIN - nDCG$ and $P + Q$ are also shown in Table 5.

Table 4: The performance of 5 runs measured using top 10 results

RunID	$I - rec$	$D - nDCG$	$D\# - nDCG$
KECIR-D-C-1B	0.6095	0.3914	0.5005
KECIR-D-C-2B	0.5204	0.2672	0.3938
KECIR-D-C-3B	0.6366	0.3998	0.5182
KECIR-D-C-4B	0.6095	0.3914	0.5005
KECIR-D-C-5B	0.6313	0.3571	0.4942

From the Table 4, we can see KECIR-D-C-3B is better than other runs. However, compare to the best official value of $D\# - nDCG$ 0.5753, the evaluated values are too low.

Table 5: $DIN - nDCG@10$ and $P + Q@10$ experimental results

RunID	$DIN - nDCG$	$P + Q$
KECIR-D-C-1B	0.2741	0.2134
KECIR-D-C-2B	0.2120	0.1331
KECIR-D-C-3B	0.2789	0.2218
KECIR-D-C-4B	0.2741	0.2134
KECIR-D-C-5B	0.2406	0.2298

6. CONCLUSIONS

In this paper, we present our system on NTCIR10 INTENT-2 task. In subtopic mining subtask, frequent sequence mining algorithm for mining subtopics is applied to the SogouT resources. An improvement could be seen after clustering and ranking the candidate subtopics with the HowNet similarity. However, our experimental results are too low. There were some topics we can't mine any subtopics.The reasons of these problems were limited resources and the simple method that generated subtopics and estimated scores of the subtopics. Therefore, more investigations are needed to explore potential of other resources in recalling diverse query intents, filtering irrelevant ones and detecting the probability of query intents. As for document ranking, two kinds of methods are applied. The experiment results show that the method, that is combining subtopic results directly, outperforms the baseline non-diversified document ranking run.

7. ACKNOWLEDGEMENT

Thank to the organizers of NTCIR-10 INTENT track. And, our work was supported by NSFC 61073123, The National Key Technology R&D Program 2012BAH14F00, 973 Program of China 2010CB530401.

8. REFERENCES

[1] R. Song, M. Zhang, T. Sakai, M. Kato, Y. Liu, M. Sugimoto, Q. Wang, and N. Orii. Overview of the ntcir-9 intent task. In Proceedings of NTCIR-9, 2011.
[2] Jialong Han, Qinglei Wang, et al. Microsoft Research Asia at the NTCIR-9 Intent Task. In Proceedings of NTCIR-9 Workshop Meeting, Pages 116-122, 2011, Tokyo, Japan.
[3] Zhai C, Cohen WW, and Lafferty J. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In Proceedings of ACM SIGIR 2003, pp 10–17.
[4] Ellen M. Voorhees and Donna Harman. Overview of TREC 2001. Proceedings of Text REtrieval Conference (TREC10). NIST Special Publications, 2001. http://trec.nist.gov/pubs.html. 2001.
[5] Y. Bai, J. Zheng, C. Guo, and D. Cai. Research on Subtopic Mining for Diversified IR. In Proceedings of CCIR2012, 2012.
[6] H. Yu, F. Ren, and S. Liu. Qualifier Mining for NTCIR-INTENT. In Proceedings of NTCIR-9 Workshop Meeting, Pages 138-142, 2011.
[7] Shuai Zhang, Kai Lu, and Bin Wang. ICTIR Subtopic Mining System at NTCIR-9 INTENT Task. In

Proceedings of NTCIR-9 Workshop Meeting, Pages 106-110, 2011, Tokyo, Japan.
[8] F. Radlinski, M. Szummer, and N. Craswell. Inferring Query Intent from Reformulations and Clicks. Proceedings of the 19th international conference on World Wide Web, North Carolina, USA, 2010.
[9] M. Hearst and J. Pedersen. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In Proceedings of SIGIR1996, 1996.
[10] P. Pirolli, P. Schank, M. Hearst, and C. Diehl. Scatter/gather browsing communicates the topic structure of a very large text collection. In CHI, 1996.
[11] H. Zeng, Q. He, Z. Chen, et al. Learning to cluster web search results. In Proceedings of SIGIR2004, 2004.
[12] Z. Dong and Q. Dong. Hownet And the Computation of Meaning. World Scientific Publishing Co., Inc.River Edge, NJ, USA, 2006.
[13] SougouQ log data. http://www.sogou.com/labs/dl/q-e.html. 2012.
[14] T. Sakai, N. Craswell, R. Song, et al. Simple evaluation metrics for diversified search results. In Proceedings of the 3rd Annual International Workshop on Evaluationg Information Access.
[15] T. Sakai, Y. Liu, Z. Dou, et al. Overview of the NTCIR-10 INTENT-2 Task. In Proceedings of NTCIR-10, 2013.
[16] T. Sakai, Z. Dou, R. Song, and N. Kando. The reusability of a diversified search test collection. In Proceedings of AIRS 2012, pages 26 – 38, 2012.