KECIR at the NTCIR-10 INTENT Task

Cheng Guo, Yu Bai, Jianxi Zheng, Dongfeng Cai
Research Center for Knowledge Engineering,
Shenyang Aerospace University, Shenyang 110136, China

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.

Experimental Results

RunID	Description	D# -nDCG
KECIR-S-C-1B	The baseline method, use FSM Algorithm to get subtopics.	0.3570
KECIR-S-C-2B	Base on KECIR-S-C-1B, use the first method in section 2.2	0.3854
KECIR-S-C-3B	Base on KECIR-S-C-1B, employ the second method in section 2.2	0.3116
KECIR-S-C-4B	Base on KECIR-S-C-1B, use the third method in section 2.2	0.3001

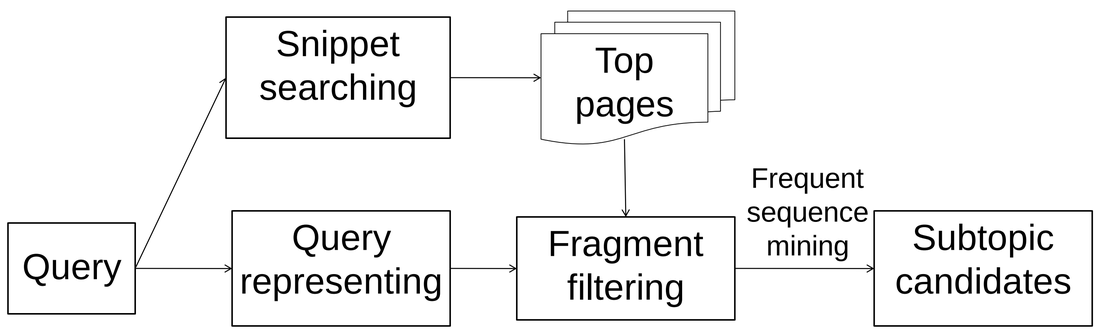
Subtopic Mining

Assumption

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

Modeling

➢ *Candidates Subtopics Mining*



➢ *Subtopics Clustering and Ranking*

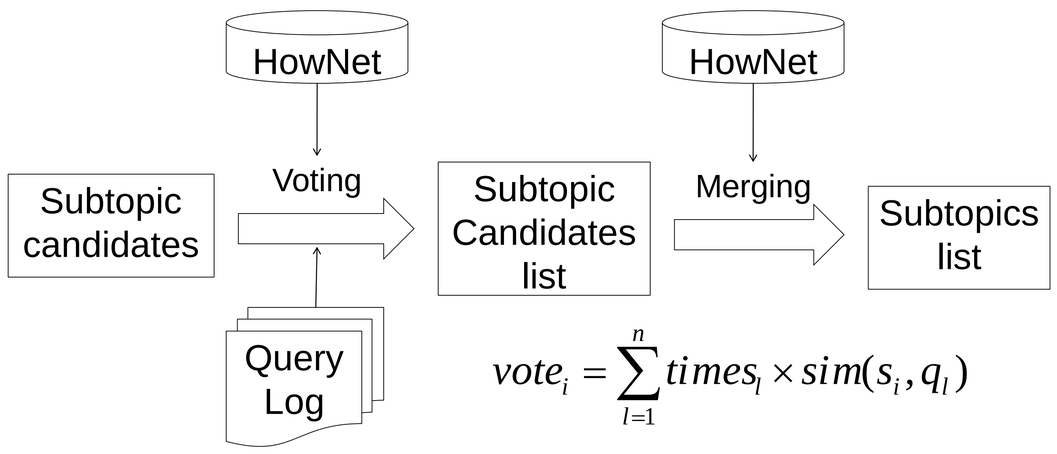
✓ **Semantic similarity**

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

✓ **Merging based on the DEF**

In this method , we cluster and rank the subtopics based on finding the same DEF in HowNet of the representative words .

✓ **Voting**



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

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.

Document Ranking

➢Score Documents Directly

✓***Just considering the coverage of subtopics in documents.***

✓***Considering the coverage and the position of subtopics***

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

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

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.

➢Map Back to Snippets

✓***If one document snippet contains a subtopic, the document will get a vote.***

✓***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.

Experimental Results

RunID	Description	D# -nDCG
KECIR-D-C-1B	Based on the baseline result and appearances of subtopics in the snippets.	0.5005
KECIR-D-C-2B	Based on the similarity result and appearances of subtopics in the htmls.	0.3938
KECIR-D-C-3B	Based on the similarity result and appearances of subtopics in the snippets.	0.5182
KECIR-D-C-4B	Based on the query log result and appearances of subtopics in the snippets.	0.5005
KECIR-D-C-5B	Based on the query log and HowNet results, also cumulative gain of subtopics in the snippets.	0.4942