

ISCAS at Subtopic Mining Task in NTCIR9

Xue Jiang, Xianpei Han, Le Sun
Storage & Information Retrieval Laboratory
Institute of Software, Chinese Academy of Sciences, China

Introduction

● Benefits of mining subtopics

✓ Specify user's query intent

Users often submit a general query which can not express his intent, subtopic describe user's intent more particularly.

✓ Diversifying query suggestions and retrieved documents

Diversifying query suggestions and retrieved documents by clustering or classification according to these subtopics.

● Our Approach

✓ Find the related queries in the corpus, such as query log, searching results.

✓ Cluster these subtopics and rank them with consideration of the relevance.

Modeling

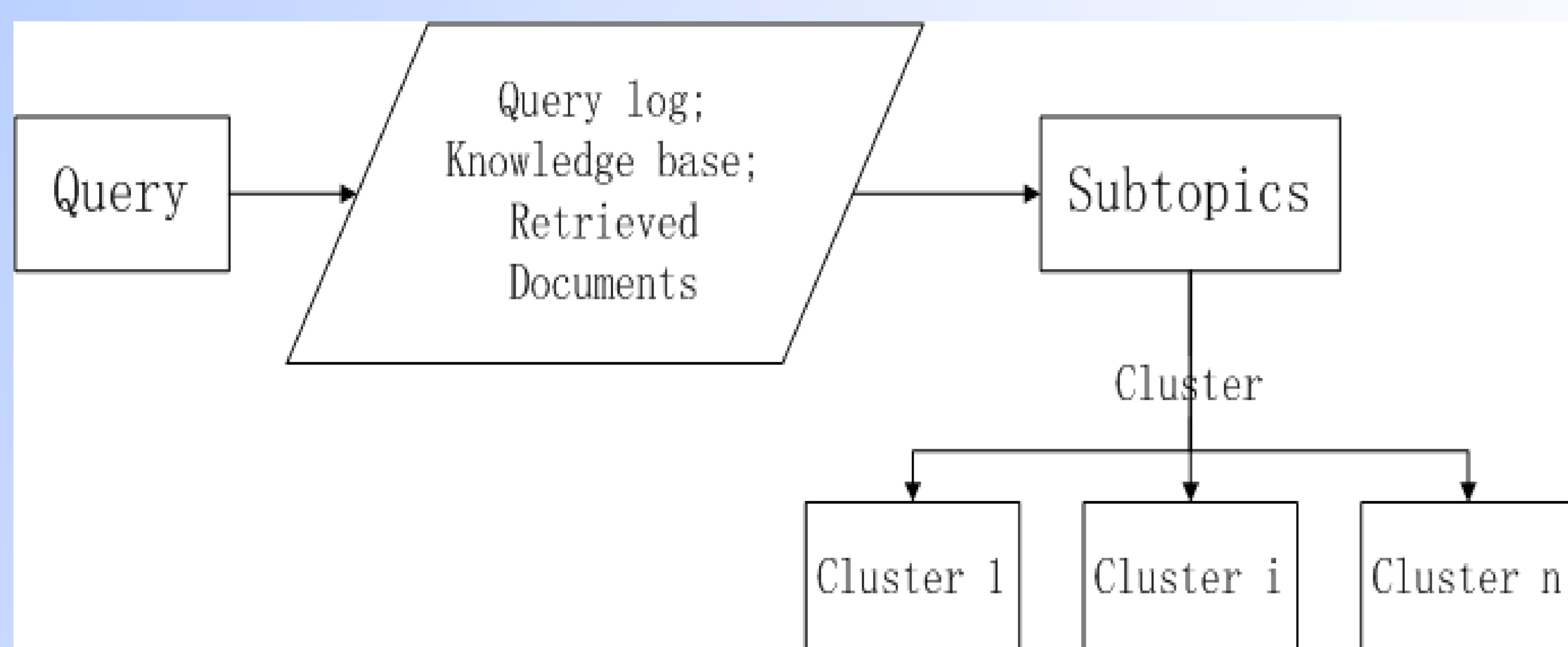
● Assumption

✓ Subtopics are specifications of original queries, they may exist in the form of the modifications or specifications of original queries.

✓ Most queries are noun or noun phrase, we can find corresponding entries in online encyclopedia.

✓ Knowledge bases contain rich information about the subtopics or aspects of certain object.

● Model



● Data

✓ SogouQ

History queries issued by users

✓ SogouT

Titles of documents retrieved by Indri

✓ Catalogue of corresponding entry in Baidu Baike

✓ Titles of retrieved documents by Baidu, Google

Mining Subtopics

● Calculate the similarity between original query and possible subtopics.

$$Sim(Q, S) = \frac{\sum_{w \in Q \cap S} c_Q(w)c_S(w)}{|Q| \cdot |S|} f(c(s))$$

✓ Q: query, S: possible subtopic, $c_Q(w)$: counts of word w occurs in Q, $f(c(s))$: function of $c(s)$, $c(s)$: counts of s occurs in the dataset.

Clustering and Ranking

● K-means with 5 or 10 clusters respectively

● Score of each Cluster G

$$Score(G) = \sum_{i=1}^n Sim(Q, S_i)$$

● Rank clusters first, then iteratively select the top subtopics in each cluster.

Experiments

Runs	I-rec	D-nDCG	D#-nDCG	Runs	I-rec	D-nDCG	D#-nDCG
ISCAS-S-C-1	0.5022*#	0.6336*#	0.5679*#	ISCAS-S-C-1	0.6406*#	0.6387*#	0.6397*#
ISCAS-S-C-2	0.3019	0.4491	0.3755	ISCAS-S-C-2	0.3922	0.4434	0.4178
ISCAS-S-C-3	0.491*#	0.6386*#	0.5648*#	ISCAS-S-C-3	0.6478*#	0.637*#	0.6424*#
ISCAS-S-C-4	0.3062	0.481	0.3936*	ISCAS-S-C-4	0.4053	0.4626	0.434

Table 1. Top 10 results

Table 2. Top 20 results

Runs	I-rec	D-nDCG	D#-nDCG	Runs	Data	Num of clusters
ISCAS-S-C-1	0.6861*#	0.5783*#	0.6322*#	ISCAS-S-C-1	SogouQ, Baidu, Google, Baidu Baike	10
ISCAS-S-C-2	0.432	0.4059	0.4189	ISCAS-S-C-2	SogouQ	10
ISCAS-S-C-3	0.6884*#	0.5419*#	0.6152*#	ISCAS-S-C-3	SogouQ, Baidu, Google, Baidu Baike	5
ISCAS-S-C-4	0.4394	0.4066	0.423	ISCAS-S-C-4	SogouQ	5

Table 3. Top 30 results

Table 4. Description of each run

Future Work

● Word mismatch

✓ Traditional problems in NLP, we may introduce semantic resources to solve this problem

● Word overmatch

✓ some query S may be similar to query Q, but not the subtopic of Q

✓ query “汶川地震原因” and “汶川地震校舍倒塌原因” are lexically similar, but they are of different topics, the former is to find the reason of earthquake, the later is to find the reason of collapse of the building

✓ user clicks recorded by the query log, or some semantic resources could be used to solve this problem.