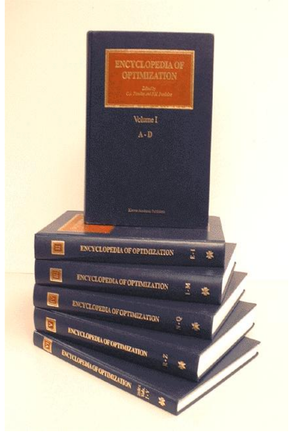

ICTIR Subtopic Mining System at NTCIR-9 INTENT Task

Shuai Zhang(张帅), Kai Lu(鲁凯), Bin Wang(王斌)
IR Group, Institute of Computing Technology, Chinese Academy of Sciences
{zhangshuai01, lukai, wangbin}@ict.ac.cn
http://ir.ict.ac.cn

Introduction

The main idea of our system is to first collect candidate query strings from different resources including query logs, online encyclopedias and commercial search engines, then mine frequent term patterns using Apriori algorithm, finally cluster the remaining candidates into different subtopics, which are represented by clusters.

System Architecture

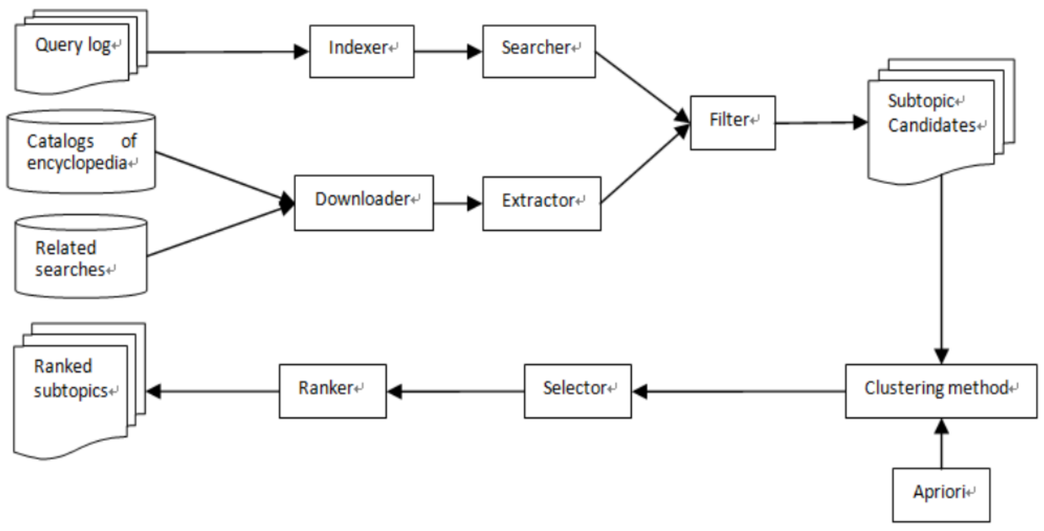

Figure 1. The architecture of our subtopic mining system

Resources

1. Query logs

SogouQ: query logs in June 2008
Sina iAsk : query logs from September to October, 2006

2. Online encyclopedia

Wikipedia (Chinese) ; Hudong

3. Related searches from search engines

Commercial search engine: Baidu, Sogou, Soso

Preprocessing

- Index query logs by single words, using Lucene.
  Given a query, search all the relevant logs.
- Filter the query logs using some heuristic rules.
  The length of query string, its distance to the topic and some other features are considered. Edit Distance is used as the distance measure.
- Download the webpages from the search engines and two online encyclopedias, then extract the related searches and catalogs.

Pattern-based clustering

Clustering Process

1. Segment all the subtopic candidates from text to a set of terms, using ICTCLAS analyzer.
2. Mining frequent term-sets. using Apriori algorithm. Parameter: min_support
3. Partition the candidates into clusters based on the patterns (frequent term-sets).

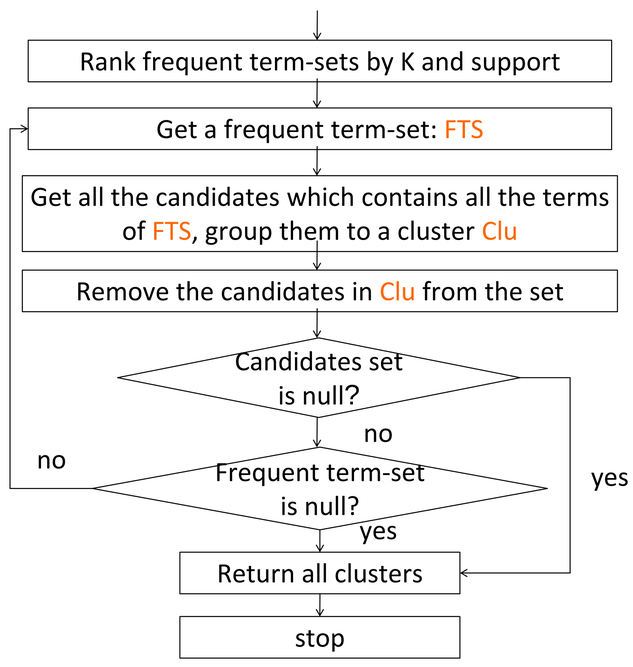

Figure 2. Frequent term-set based clustering

Subtopic selection & ranking

- Central candidates of clusters are chosen as subtopics. Edit Distance is used to compute the distance between strings.
- Subtopics are ranked by the size of their corresponding clusters.

Example

For topic: "莫扎特" Mozart

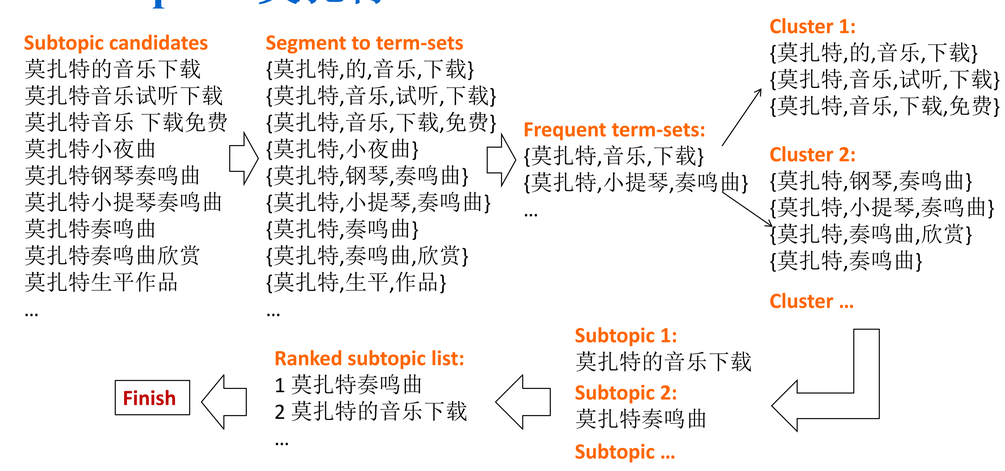

Figure 3. An example for topic *Mozart*

Evaluation

Primary evaluation metric

D#-nDCG: a linear combination of *intent recall* (or "I-rec", which measures diversity) and *D-nDCG (which* measures overall relevance across intents.

$$D\sharp\text{-}measure@l = \gamma I\text{-}rec@l + (1-\gamma) D\text{-}measure@l$$

Experiments

We submitted five runs. ICTIR-S-C-5 uses query log only while the others use external resources. ICTIR-S-C-1 and 4 use the same clustering strategy and 2, 3 use another. The min_support of ICTIR-S-C-1, 2, 3 ,4 are 0.005, 0.01, 0.02 and 0.015. As a result, 1, 2 get more subtopics than 3, 4.

Table 1. The official subtopic mining results for D#nDCG

Run id	D#nDCG@10	D#nDCG@20	D#nDCG@30
ICTIR-S-C-1	**0.5797**	**0.6579**	0.6261
ICTIR-S-C-2	0.5701	0.6452	**0.6482**
ICTIR-S-C-3	0.5669	0.5881	0.5464
ICTIR-S-C-4	0.5726	0.5893	0.539
ICTIR-S-C-5	0.5273	0.5615	0.5165

Table 2. The official subtopic mining results for I-rec

Run id	I-rec@10	I-rec@20	I-rec@30
ICTIR-S-C-1	**0.5161**	**0.6997**	**0.7224**
ICTIR-S-C-2	0.4826	0.6444	0.707
ICTIR-S-C-3	0.4808	0.5849	0.6062
ICTIR-S-C-4	0.5035	0.6206	0.634
ICTIR-S-C-5	0.4714	0.5803	0.5924

Table 3. The official subtopic mining results for D-nDCG

Run id	D-nDCG@10	D-nDCG@20	D-nDCG@30
ICTIR-S-C-1	0.6434	0.6162	0.5299
ICTIR-S-C-2	**0.6576**	**0.646**	**0.5895**
ICTIR-S-C-3	0.653	0.5913	0.4867
ICTIR-S-C-4	0.6417	0.5579	0.4441
ICTIR-S-C-5	0.5832	0.5427	0.4407

Conclusion

1. We utilize multiple resources in a unified method, which can provide more information and achieve better results.
2. Some heuristic methods are applied in the data preprocessing. Features such as the length of query and its distance to topic are employed to filter noises.
3. The clustering method is based on frequent pattern mining which is intuitive and explainable. We group the strings in a cluster because they share the same pattern. The results show that the approach is very effective.
4. The system has a universal parameter min_support, which controls the granularity of clustering. So we don't need to specify the number of subtopics for each topic like k-means algorithm.