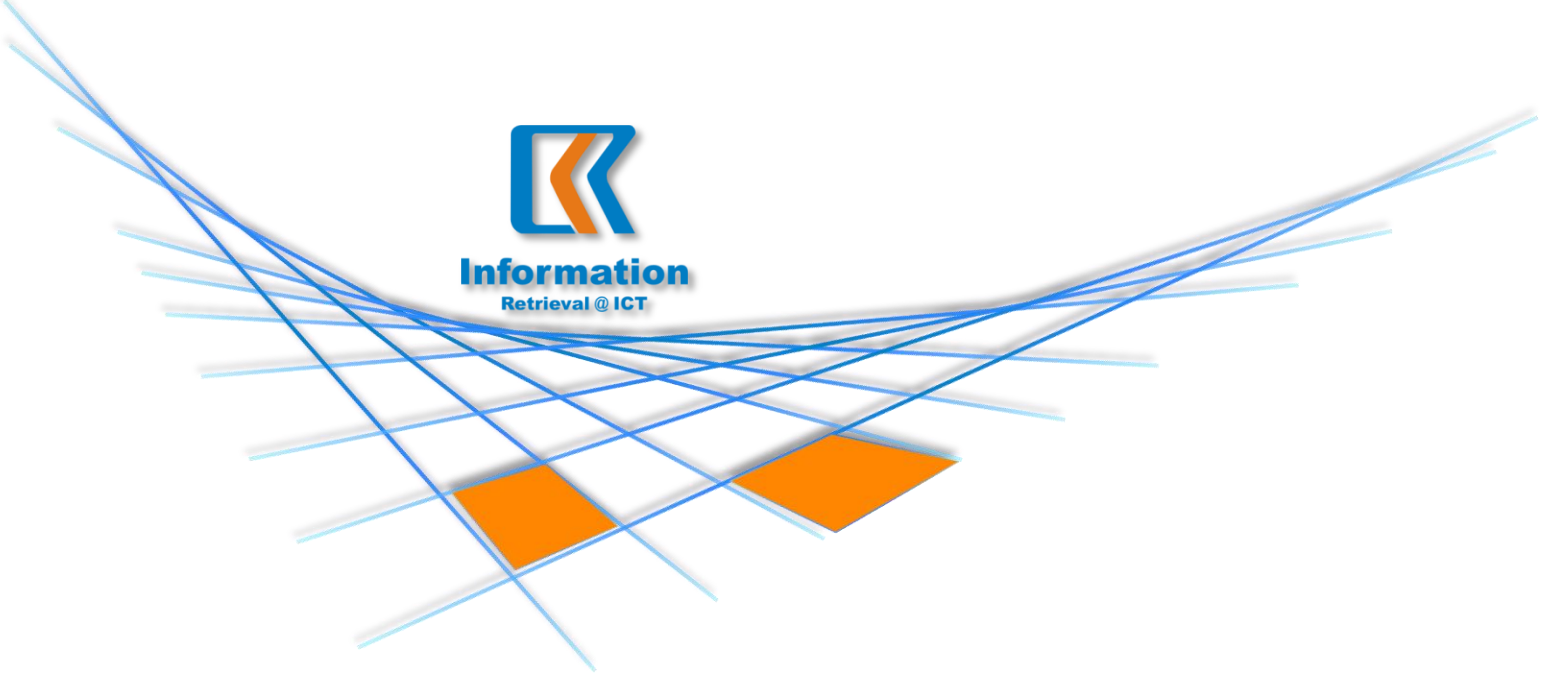


ICTIR Subtopic Mining System at NTCIR-9 INTENT Task

Shuai Zhang, Kai Lu, Bin Wang

Information Retrieval Group

Institute of Computing Technology

Chinese Academy of Science

中科院计算所信息检索组

Outline

- Introduction
- ICTIR subtopic mining system
  - Architecture
  - Dataset
  - Preprocessing
  - Clustering method
  - Subtopic ranking
- Evaluation
- Official result
- Conclusion

Outline

- Introduction
- ICTIR subtopic mining system
  - Architecture
  - Dataset
  - Preprocessing
  - Clustering method
  - Subtopic ranking
- Evaluation
- Official result
- Conclusion

Introduction

- **Intent task**
  - Many web queries are short and vague. By submitting one query, users may have different intents.
  - For an ambiguous query, users may seek for different interpretations.

    Eg. “house windows”, “microsoft windows”
  - For a query on a broad topic, users may be interested in different subtopics.

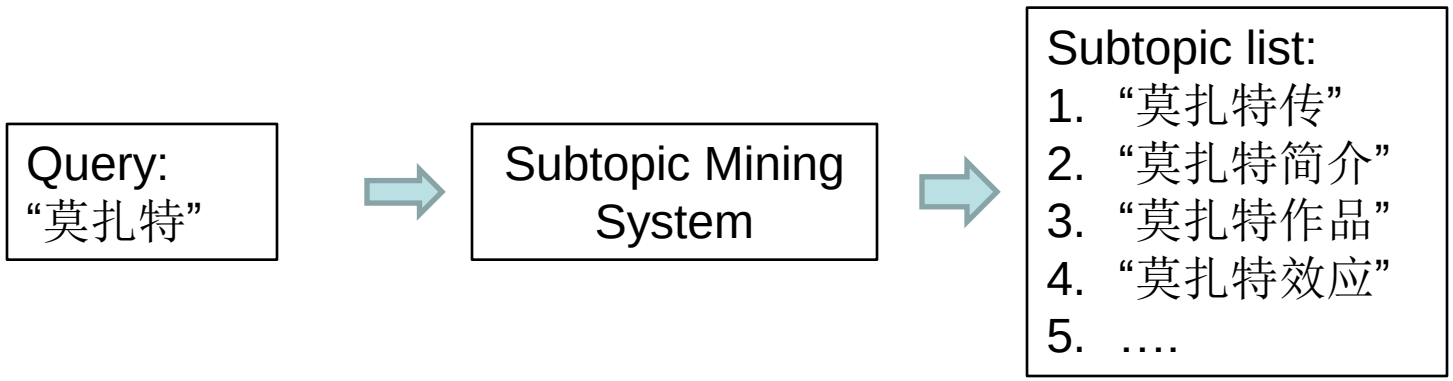
    Eg. “windows update”, “windows phone”

Introduction

- **Subtopic mining**
  - A subtopic could be an interpretation of an ambiguous query or an aspect of a faceted query

Input: a query, Eg. “莫扎特” *Mozart*

Output: a ranked list of subtopic strings

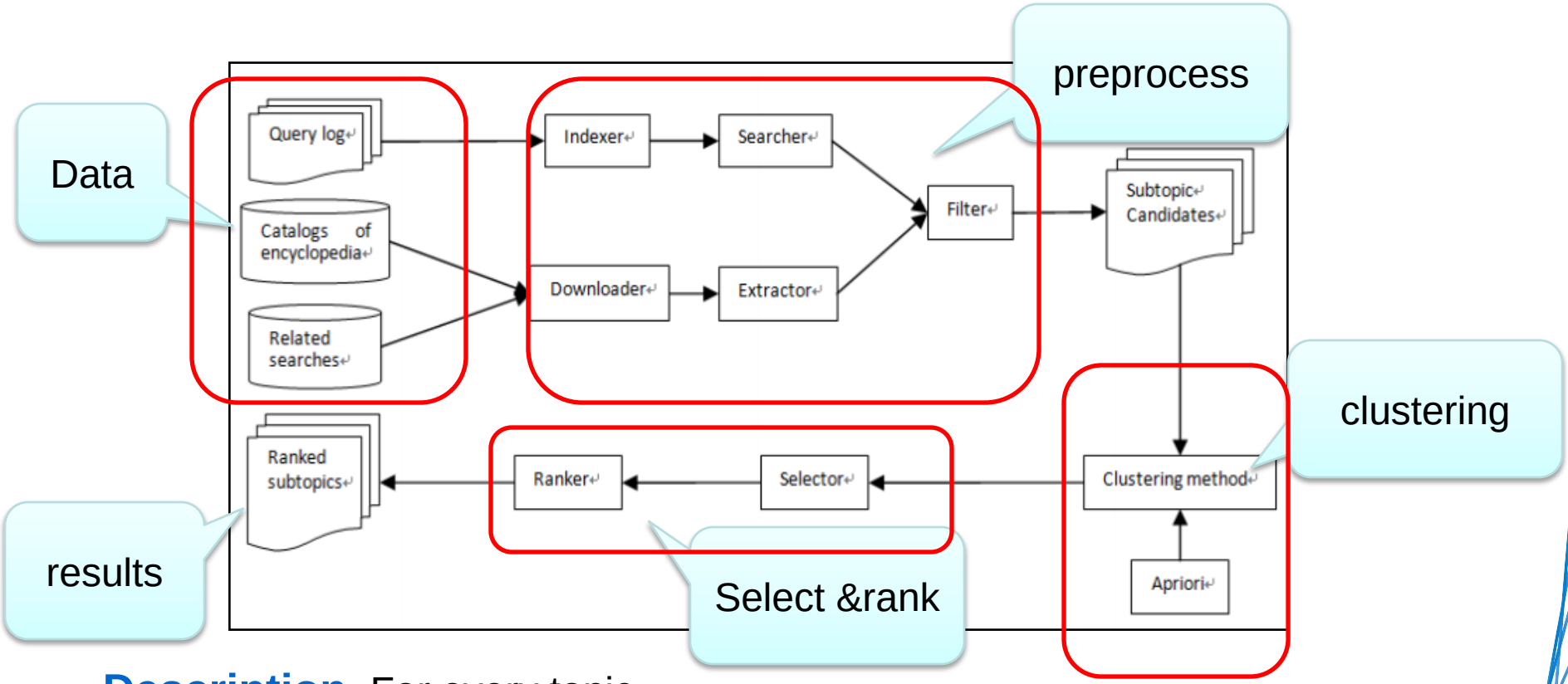


Basic idea: **query clustering**

Outline

- Introduction
- ICTIR subtopic mining system
  - Architecture
  - Dataset
  - Preprocessing
  - Clustering method
  - Subtopic ranking
- Evaluation
- Official result
- Conclusion

System Architecture



Description: For every topic

1, Collect subtopic candidates from query logs, encyclopedia catalogs and related searches .

2, The frequent term-set based clustering algorithms are conducted.

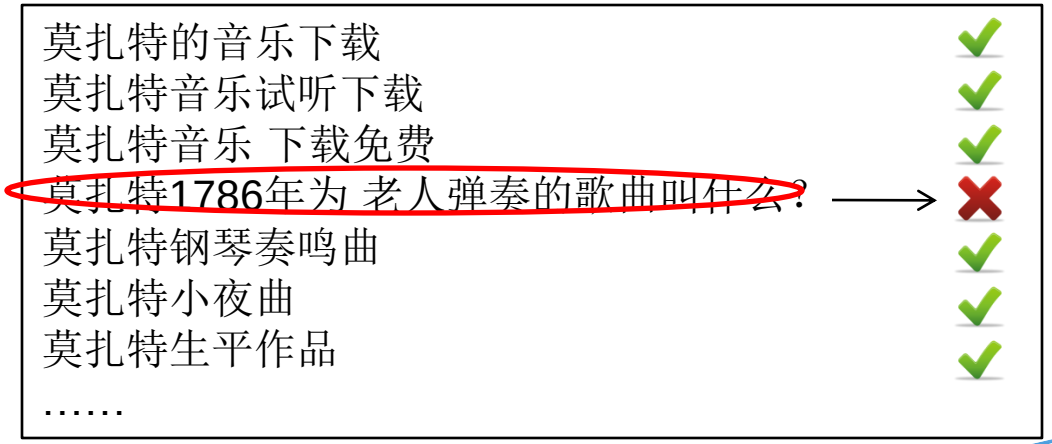
3, The centroids of clusters are selected to represent the subtopic.

Data we used

- Query log
  - SogouQ: query logs in June 2008
  - Sina iAsk : query logs from September to October, 2006
- Online encyclopedia
  - Wikipedia (Chinese)
  - Hudong
- Related searches from search engines
  - Commercial search engine: Baidu, Sogou, Soso

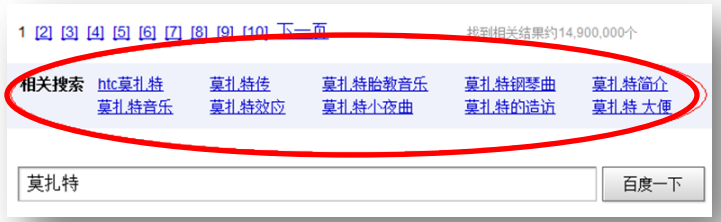
Preprocessing

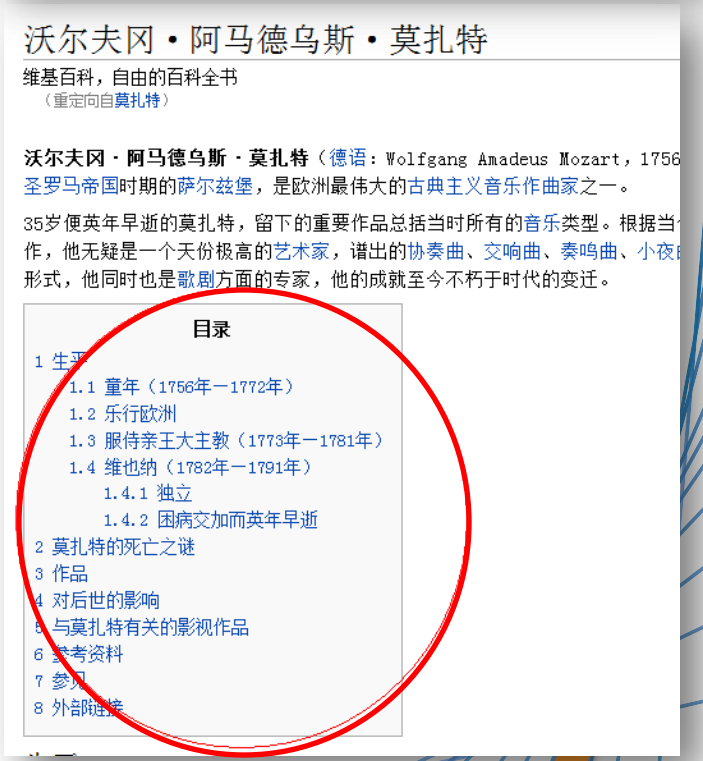
- query logs
  - Index query logs by single words, using Lucene. Given a query, search all the relevant query logs.
  - We utilize some heuristic method to filter noises. Features such as the length of a query and its Edit Distance to the topic is utilized.
- Eg. “莫扎特” *Mozart*



Preprocessing

- Encyclopedia & Search engine
  - Download the WebPages
  - Extract the information we need
- Collect into candidates set
  - After preprocessing, we treat the strings we get from three kinds of data as subtopic candidates, and put them into a subtopic candidates set.





Frequent term-set based clustering

- Motivation：
  - After obtaining subtopic candidates, we introduced a frequent term-set based clustering method to mine subtopics.
  - Candidates string in a cluster have the same pattern. That is they all contains the same term-set.
  - Eg. “莫扎特作品” *Mozart works*

Term-set: {莫扎特，作品}:

Cluster :

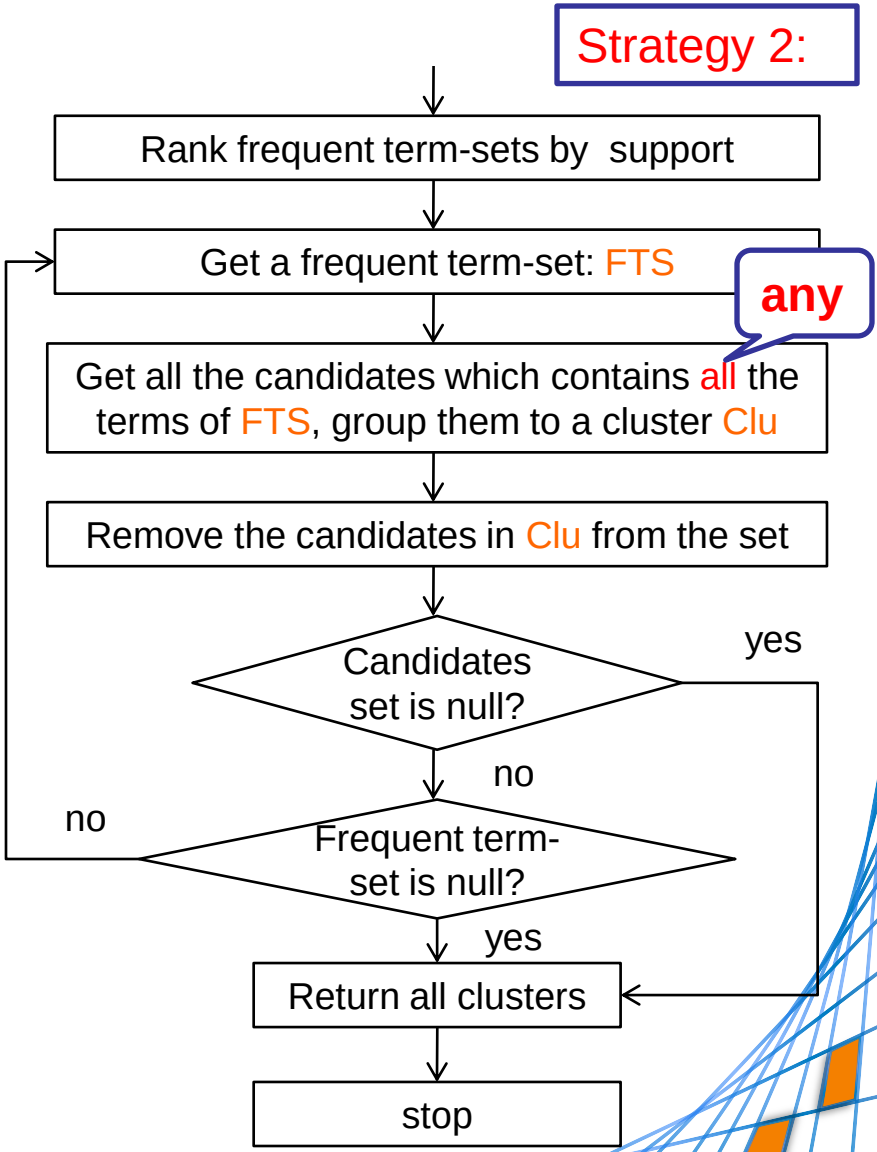
莫扎特生平作品
莫扎特的作品
所有莫扎特的作品
莫扎特作品风格
莫扎特作品资料
莫扎特生平作品简介
莫扎特代表作品
莫扎特作品比赛
莫扎特作品阿里路亚
莫扎特作品目录
莫扎特作品介绍
莫扎特 作品
莫扎特 作品 统计
莫扎特作品年表
莫扎特的所有作品

Frequent term-set 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.
3. Partition the subtopic candidates set into clusters based on the frequent term-sets.

flow chart on the left



Mining parameter

- Apriori parameter: min_support, it is a threshold
- If the frequency of a term-set is larger then the min_support, we consider it frequent
- Affect the number of subtopics, the granularity of clustering

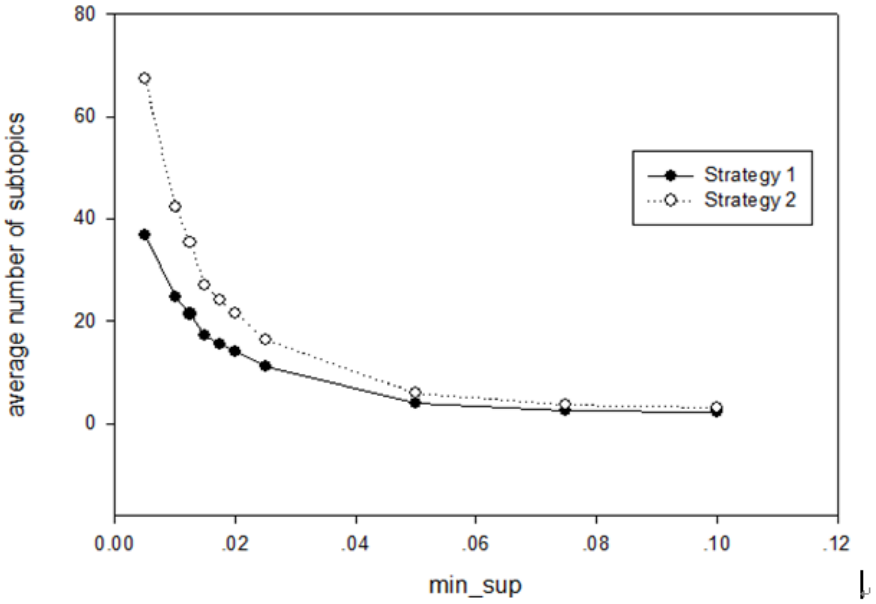


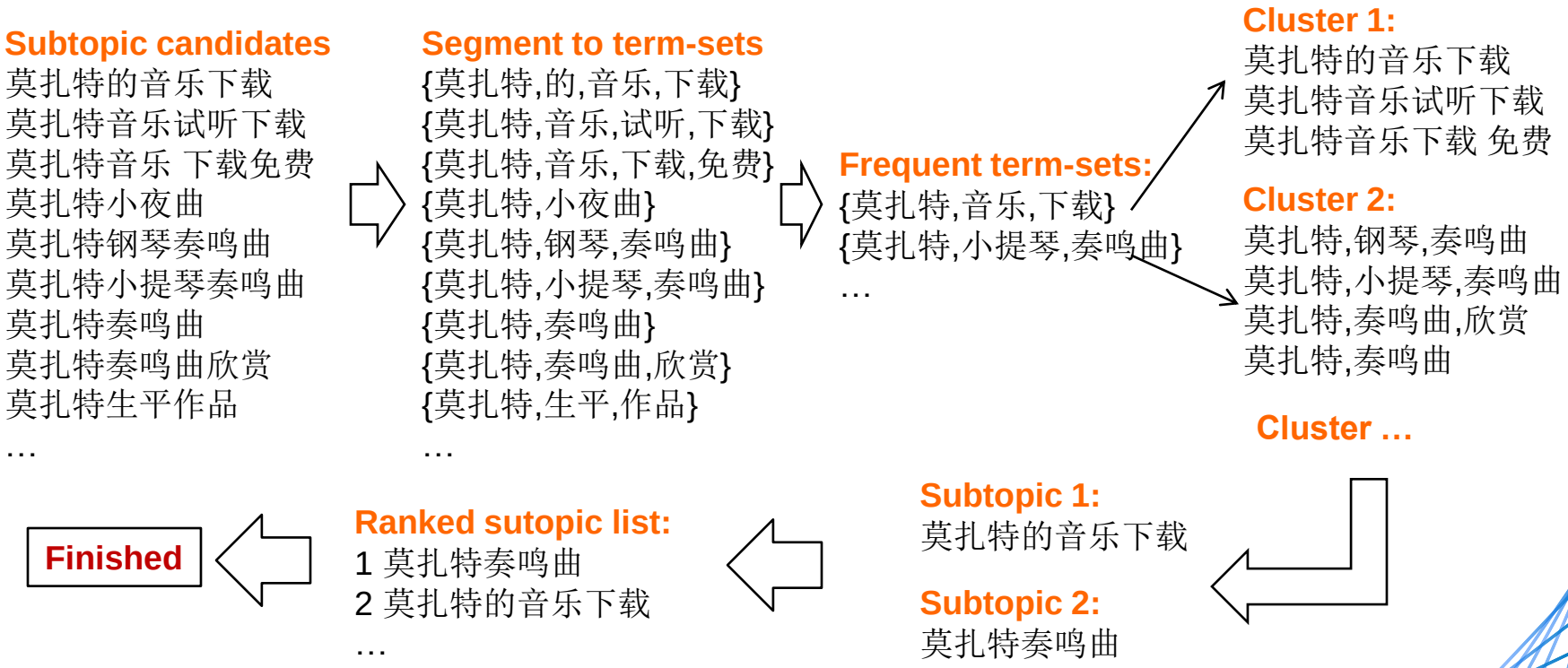
Figure 3. Relationship between min_sup and the average number of subtopics per topic

Subtopic selection & ranking

- Centroids of clusters are chosen as subtopics
  - Use Edit Distance to compute the distance between strings.
  - Central point is the point which has the shortest average distance to others in the cluster.
- Rank subtopic
  - Subtopics are ranked simply based on the size of the its cluster. (The number of subtopic candidates the cluster contains)

Example from subtopic candidates set

- Example
  - Query: “莫扎特” Mozart



Outline

- Introduction
- ICTIR subtopic mining system
  - Architecture
  - Dataset
  - Preprocessing
  - Clustering method
  - Subtopic ranking
- Evaluation
- Official result
- Conclusion

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).

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

  - In the official experiment:

    measurement depths: l=10, 20, 30

    γ=0.5, simple average

Outline

- Introduction
- ICTIR subtopic mining system
  - Architecture
  - Data we used
  - Data preprocessing
  - Clustering method
  - Subtopic ranking
- Evaluation
- Official result
- Conclusion

Outline

- Introduction
- ICTIR subtopic mining system
  - Architecture
  - Dataset
  - Preprocessing
  - Clustering method
  - Subtopic ranking
- Evaluation
- Official result
- Conclusion

Official Results

- We submitted 5 runs for evaluation.
- ICTIR-S-C-1 achieves the highest I-rec values.
- ICTIR-S-C-1 and ICTIR-S-C-2 show good performance among all runs.

Run id	Description		D#nDCG		
	strategy	Min_sup	**@10**	**@20**	**@30**
ICTIR-S-C-1	1	0.005	**0.5797**	**0.6579**	0.6261
ICTIR-S-C-2	2	0.01	0.5701	0.6452	**0.6482**
ICTIR-S-C-3	2	0.02	0.5669	0.5881	0.5464
ICTIR-S-C-4	1	0.015	0.5726	0.5893	0.539
ICTIR-S-C-5	1	0.01	0.5273	0.5615	0.5165

Conclusion

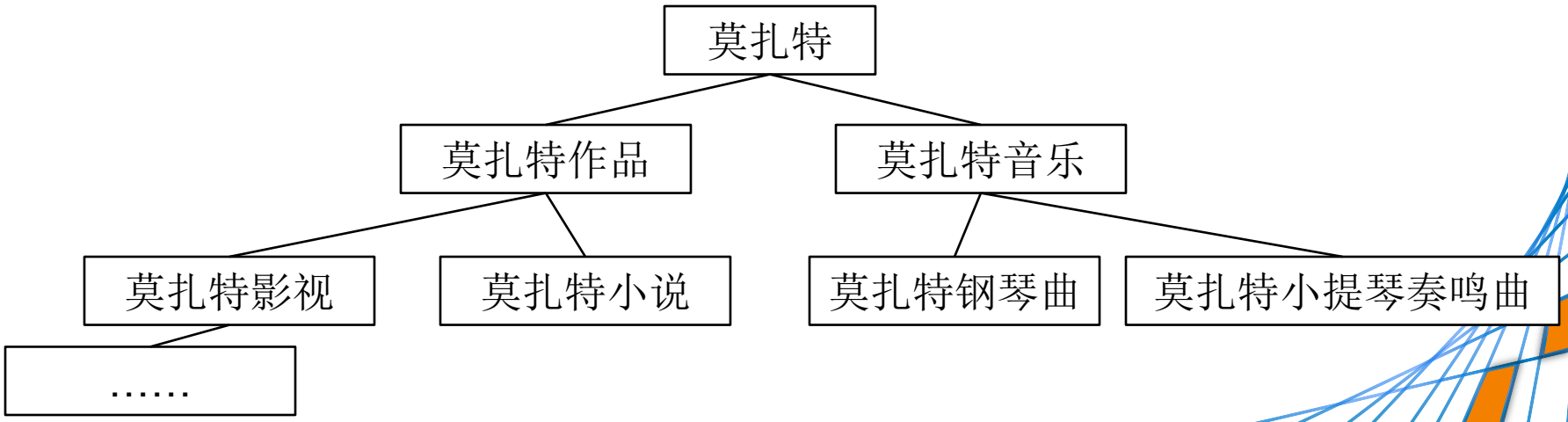
- **Summary**

1. We utilize multiple resources in a unified method, which can provide more information and achieve better results. As the results show, ICTIR-S-C-5 is not as good as others.
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. So we can get better subtopic candidates.
3. The clustering method is based on frequent pattern mining which is very intuitive. We group the strings in a cluster because they share the same pattern. The results show that the approach is very effective.

Conclusion

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 clusters for each topic like k-means algorithm.

   Actually, it's hard to decide the number of subtopics. Subtopics also have subtopics. It's a tree structure. Eg. "莫扎特" *Mozart*



Feature work

- Need to improve

1. Try and compare other clustering method (Eg. Hierarchical clustering)

2. Try other distance measure(Eg. Longest common sequence) for preprocessing.

3. Improve the subtopic ranking algorithm. Utilize more features.

Thanks for your attention!

Question & Answer