Understanding the Query: THCIB and THUIS at NTCIR-10 Intent Task

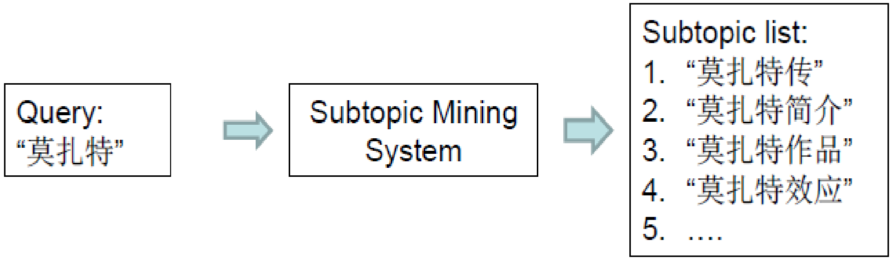
Junjun Wang
2013/4/22

Outline

- Introduction
- Related Word
- System Overview
- Subtopic Candidate Mining
- Subtopic Ranking
- Results and Discussion
- Conclusion

Introduction

- NTCIR Intent Task
  - Why?
    - Many web queries are short and vague.
    - By submitting one query, users may have different Intents.
  - What?
    - Subtopic Mining
    - Document Ranking
- Subtopic Mining
  - THCIB for English
  - THUIS for Chinese



Query:
“莫扎特”

Subtopic Mining System

Subtopic list:
1. “莫扎特传”
2. “莫扎特簡介”
3. “莫扎特作品”
4. “莫扎特效应”
5. ....

Related Word

- Intent Mining
  - Systems using multiple resources tend to be advantageous
    - THUIR (Y. Xue, F. Chen, et al, 2011)
    - ICTIR (R. Song, M. Zhang, et al, 2011)
    - HITCSIR (W. Song, Y. Zhang,et al, 2011)
  - Clustering on subtopic candidates is helpful to find intents
    - ICTIR (R. Song, M. Zhang, et al, 2011)
    - HITCSIR (W. Song, Y. Zhang,et al, 2011)
- Intent Ranking
  - Most NTCIR-9 intent systems rank intents and subtopic based merely on relevance score
    - THUIR (Y. Xue, F. Chen, et al, 2011)
  - MMR model
    - MSINT (J. Han, Q. Wang, et al, 2011)

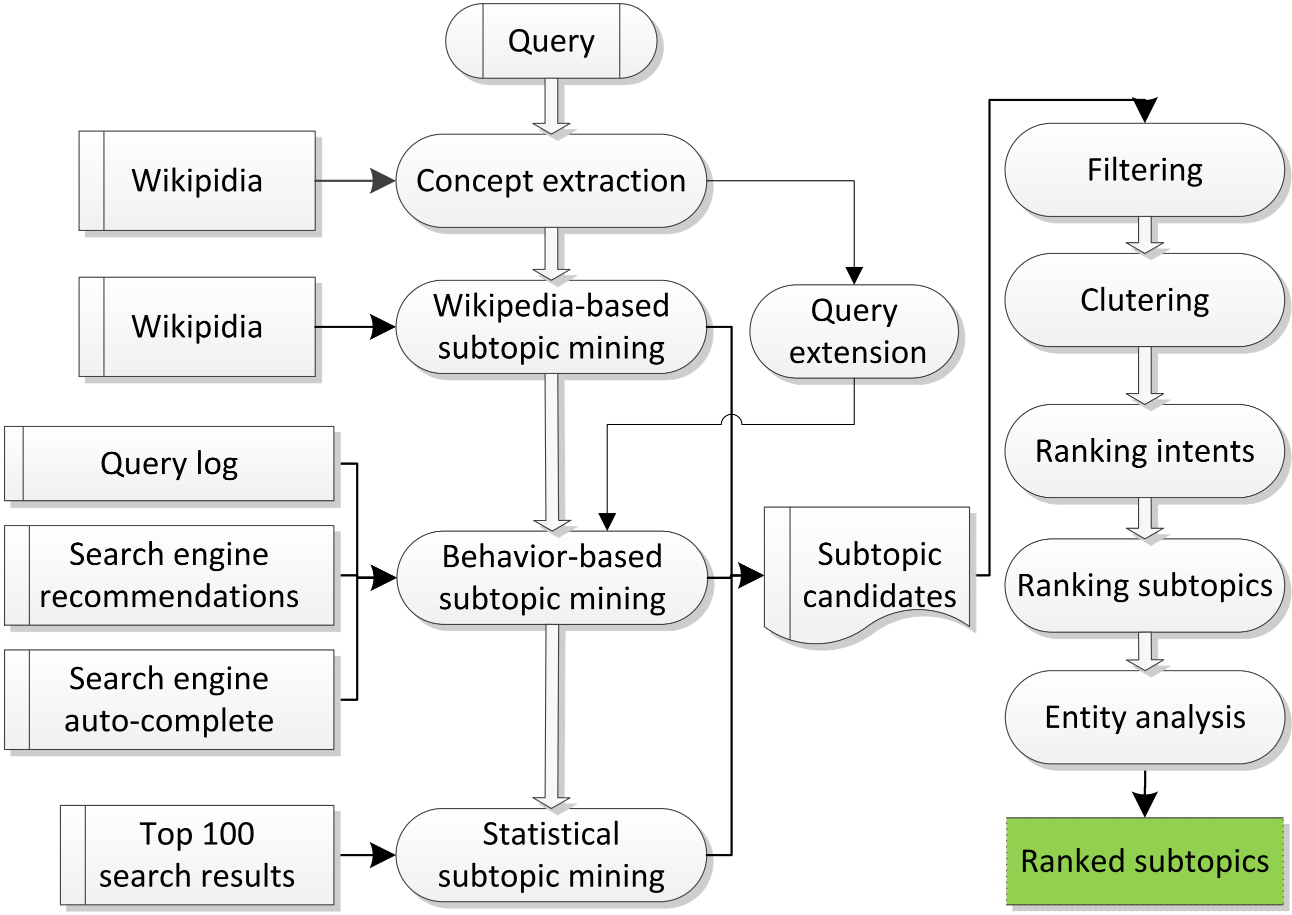
System Overview

- Understand a specific query with context:
  - query
  - knowledge base
  - search results
  - user behavior statistics

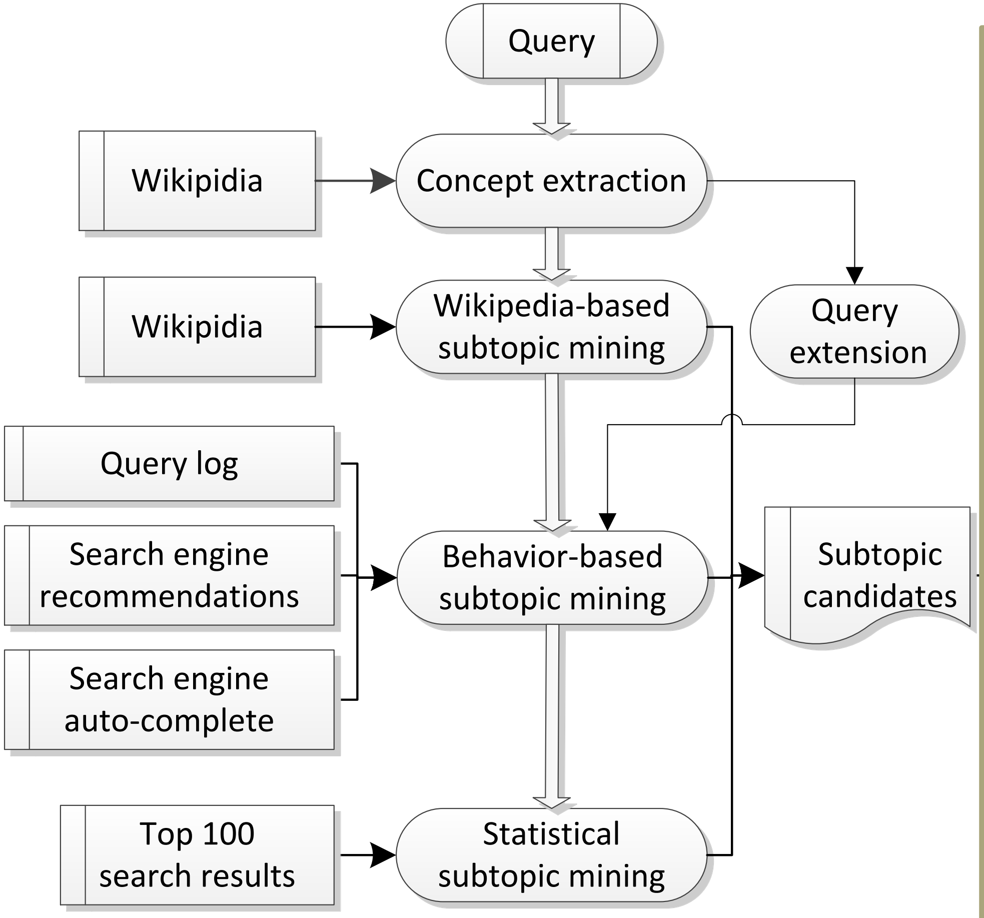
- Discover intents by clustering the subtopic candidates

- A unified model to rank
  - Relevance
  - diversity

System Overview



System Overview


Query
Wikipidia
Concept extraction
Wikipidia
Wikipedia-based subtopic mining
Query extension
Query log
Search engine recommendations
Search engine auto-complete
Behavior-based subtopic mining
Subtopic candidates
Top 100 search results
Statistical subtopic mining

SCM (Subtopic Candidate Mining):

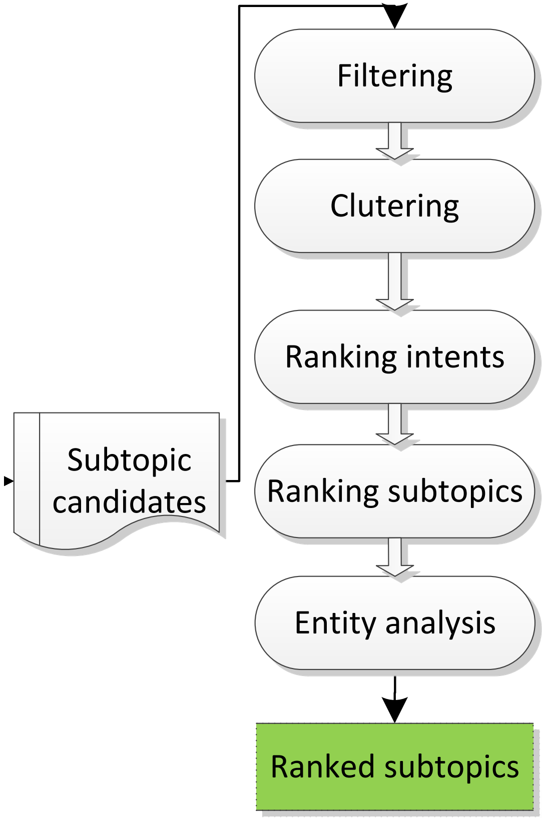
Mine subtopic candidates as many as possible

System Overview

SCR

(Subtopic Candidate Ranking):

Rank relevant subtopic candidates according to both relevance and diversity



Subtopic Candidate Mining

- Extracting Concept(s) from Query
  - "battles in the civil war" : {battle, civil war}

  - Concept dictionary: Wikipedia
    - English version for English Task
    - Chinese version for Chinese Task

  - Procedure:
    - Stemming & tokenizing for English (TreeTagger)

      Word segmentation for Chinese (ICTCLAS)
    - Search the Wikipedia
    - Extract Wikipedia entries (i.e. concepts) using Bi-direction Matching Method

Extending the Query

- The same intent can be expressed with different queries. We adopt two manners to extend the query.

- Involving conceptually identical word/phrases to get synonymous queries
  - Wikipedia redirect and disambiguation pages

- Constructing intent schemas
  - Schema = concepts + prepositions + wildcards

    "hobby store"
    - " * of hobby store "
    - " hobby stores in * "
  - Adjusting order of the concepts in the query

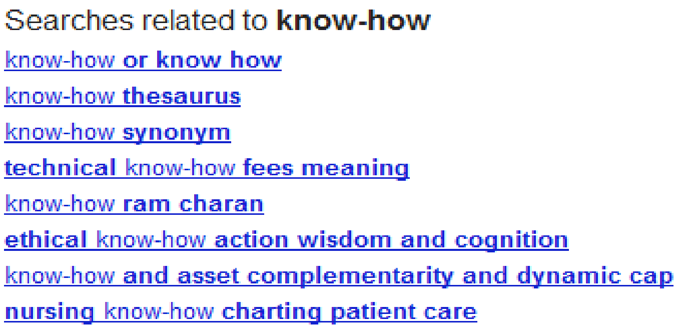
Mining Subtopics in Wikipedia

- Disambiguation
  - “battle”
    - surname, military confliction, music, film, and so on
- Redirects
  - “shortest path” -> “shortest path problem”
- Catalogs
  - “rock art”
    - rock art background, rock art type, rock art studies, etc.
- Related entries which contain the query as a substring
  - “rock”
    - “rock music”, “rock band”, and so on

Mining Subtopics in User Statistics

- Query log
  - SogouQ for Chinese
  - Anchor Text Query Log of ClueWeb09 for English
  - Index and search query logs with Lucene

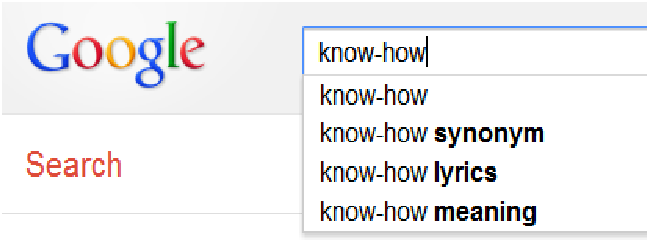
- Search recommendations



Searches related to **know-how**

know-how **or know how**
know-how **thesaurus**
know-how **synonym**
technical know-how **fees meaning**
know-how **ram charan**
ethical know-how **action wisdom and cognition**
know-how **and asset complementarity and dynamic cap**
nursing know-how **charting patient care**

- Auto-completions



know-how

know-how
know-how **synonym**
know-how **lyrics**
know-how **meaning**

Mining Subtopics in Search Results

- Search results
  - Titles and snippets of top 100 results and sentences that contain the query completely or partially in the original result page
  - Google for English
  - Sogou for Chinese

- Word sense induction (WSI) framework
  - LDA
  - Unigram & bigram
  - The top word of each topic together with the query

Subtopic Candidate Filtering

- For some queries, we obtained more than 1000 subtopic candidates. It is necessary to exclude the less likely subtopic candidates before the next step.

- Filtering Rule
  - Rule #1: Candidates that are contained in the query are excluded.
  - Rule #2: Candidates that do not contain all concepts (or corresponding synonymous concepts) of the query are excluded.

- 28.8 percent of subtopic candidates are deleted after filtering.

Subtopic Ranking

- Three factors affect rank of a subtopic
  - Relevance of the subtopic: $\boldsymbol{w_{ST}(t)}$
  - Importance of the subtopic source: $\boldsymbol{w_{SC}(t)}$
  - Significance of the intent that the subtopic belongs to: $\boldsymbol{w_{IN}(t)}$

- $\boldsymbol{w_{ST}(t)}$
  - Pseudo relevance feedback
  - Voted by the words in top 100 search results
  - Normalized by the length of the subtopic candidate

- $\boldsymbol{w_{SC}(t)}$
  - Assigning empirical weights to the involved sources

Subtopic Ranking

Resource and their weights for English

Source name	Weight
Bing Completion	1
Bing Suggestion	1
Google Completion	1
Yahoo Completion	1
Query extension	0.9
Query Log	0.2
SRC	0.4
Search Result Title	0.2
Wiki Concept Definition	0.8
Wiki Disambiguation & Redirects	1
Wiki Related Entries	0.8

Resource and their weights for Chinese

Source name	Weight
Bing Suggestion	1
Baidu Suggestion	1
Sogou Suggestion	1
Google Suggestion	1
Query extension	0.9
Query Log	0.6
SRC	0.4
Search Result Title	0.2
Wiki Concept Definition	0.8
Wiki Disambiguation & Redirects	1
Wiki Related Entries	0.8

Subtopic Clustering

- Organize the subtopic candidates into clusters (i.e. intents)

- Affinity propagation clustering algorithm
    - (Frey, Dueck, et al, 2009)
    - Exemplars are identified among data points, and clusters of data points are formed around these exemplars.
    - It operates by originally considering all data point as potential exemplars and then exchanging messages between data points until a good set of exemplars and clusters emerges.
    - Two initial inputs:
        - A similarity matrix M, where M_{ij} represents how points i prefers point j to be the exemplar
        - A preference list P, where p_i represents how likely point i should be an exemplar.

Subtopic Clustering

- M: sense-based similarity
  1. Applying Word Sense Induction to get different senses of each subtopic candidate.
  2. Choosing the most similar senses between each pairs of subtopic candidates and calculate their cosine similarity.

- P: two versions
  - **Standard AP algorithm**
    - the mean value of the similarity matrix
  - **Revised AP algorithm**
    - $p_t = w_{ST}(t) + w_{SC}(t)$

Ranking the Intent

- Considering certain intent containing subtopic candidates

$$\{t_i\}(i = 1 \ldots N)$$

significance score of the intent is calculated as follows.

$$w_{IN} = \sum_{i=1}^{N} [(w_{ST}(t_i) + w_{SC}(t_i)]$$

- Considering relevance and diversity at the same time, we propose to consider the intents that include more than 5 subtopic candidates.

Entity Analysis

- Exclusive Entities in subtopic candidates
  - person, organization, location and so on

    furniture for small spaces City

    furniture for small spaces City
  - exclusive entities sometimes lead to intent fission

- Ontology-based Entity Analysis
  - recognize the entities and their ontology type with Freebase
  - generalize subtopic candidates to associate named entities with the same ontology type to ontological clusters

- No such module in THUIS system

Subtopic Ranking

- Ranking Strategy

  1.Rank the subtopic candidates in the declining order of

  $$w_{ST}(t) + w_{SC}(t)$$

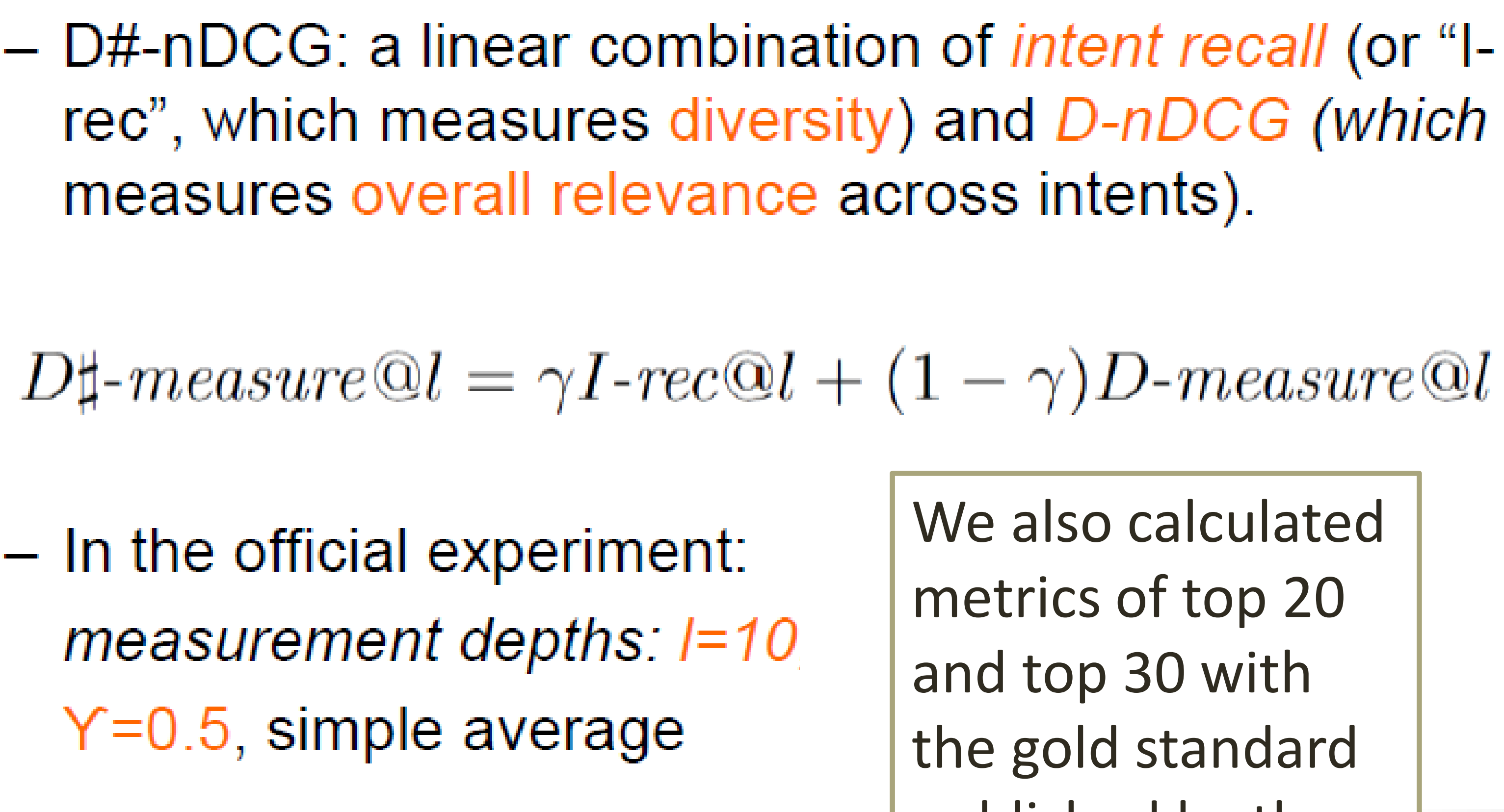
  2.Ranking the subtopic candidates inter and intra intents

  - Calculating the $w_{IN}(t)$ of each cluster after clustering and ranking the intents in declining order
  - Sorting the subtopic candidates in each cluster in descending order of $w_{ST}(t) + w_{SC}(t)$
  - Iteratively getting the top subtopic candidate in each cluster until all subtopic candidates are returned

  3.Rerank the subtopic candidates using entity analysis

  - Enlarge the distance of homogenous subtopics to enhance diversity

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).

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

- In the official experiment: *measurement depths:* *l=10* Y=0.5, simple average

We also calculated metrics of top 20 and top 30 with the gold standard published by the organiser

Submitted Runs

Module		Module No
SCM	Extracting Concept(s) from Query	M.1
	Extending the Query	M.2
	Mining Subtopics in Wikipedia	M.3
	Mining Subtopics in User Statistics	
	Mining Subtopics in Search Result	
SCR	Ranking Strategy 1:	R.1
	Ranking Strategy 2: Standard AP	R.2.1
	Ranking Strategy 2: Revised AP	R.2.2
	Ranking Strategy 3: Entity Analysis	R.3

Submitted Runs

- 5 runs for English

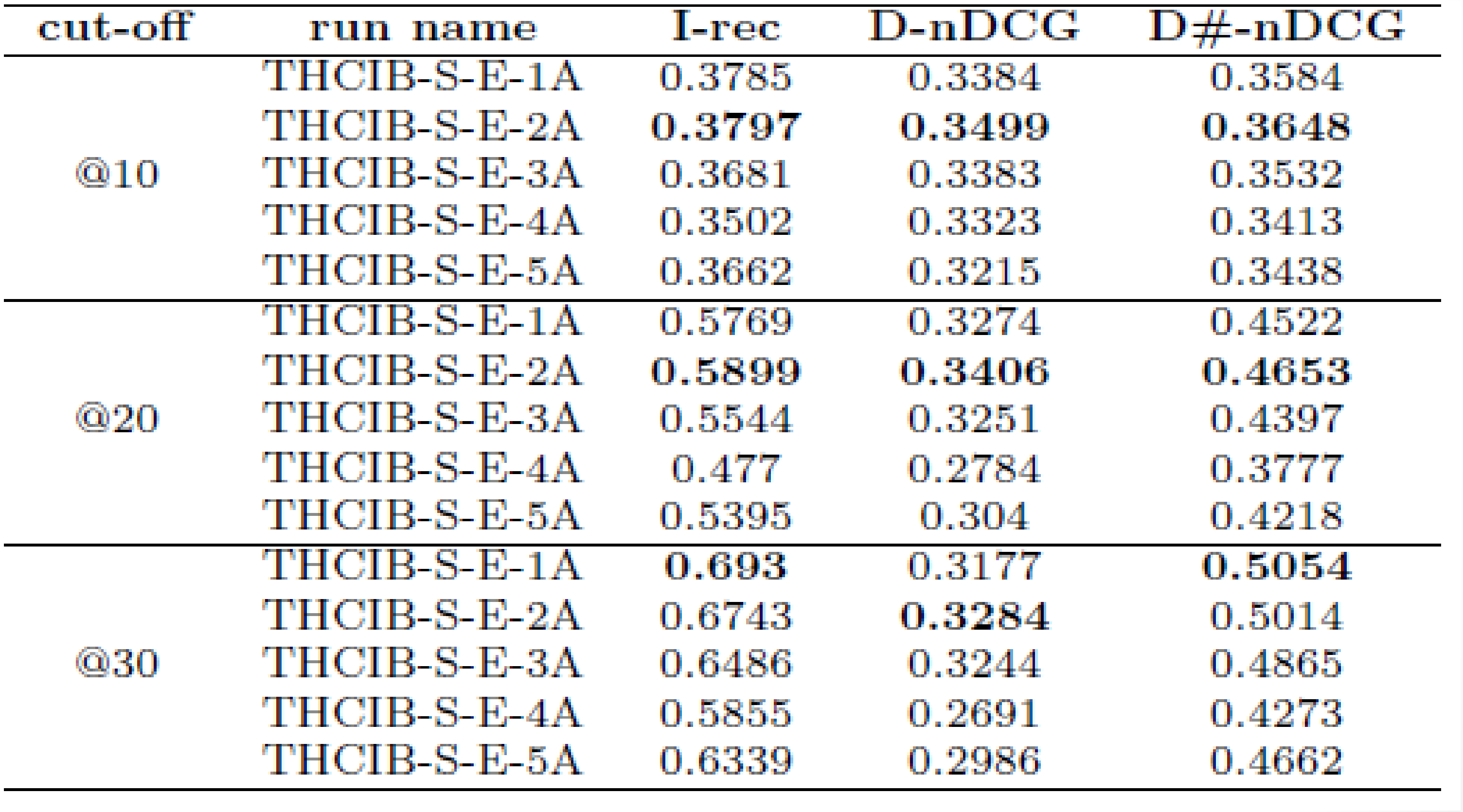
THCIB-S-E-1A	**M1 + M3 + R1**
THCIB-S-E-2A	M1 + M2 + M3 + R1
THCIB-S-E-3A	M1 + M2 + M3 + R1 + R3
THCIB-S-E-4A	M1 + M2 + M3 + R2.1 + R3
THCIB-S-E-5A	M1 + M2 + M3 + R2.2 + R3

- 4 runs for Chinese

THUIS-S-C-1A	**M1 + M3 + R1**
THUIS-S-C-2A	M1 + M2 + M3 + R1
THUIS-S-C-3A	M1 + M2 + M3 + R2.1
THUIS-S-C-4A	M1 + M2 + M3 + R2.2

Results

D#-nDCG of English Subtopic Mining runs at various cut-off levels



cut-off	run name	I-rec	D-nDCG	D#-nDCG
@10	THCIB-S-E-1A	0.3785	0.3384	0.3584
	THCIB-S-E-2A	**0.3797**	**0.3499**	**0.3648**
	THCIB-S-E-3A	0.3681	0.3383	0.3532
	THCIB-S-E-4A	0.3502	0.3323	0.3413
	THCIB-S-E-5A	0.3662	0.3215	0.3438
@20	THCIB-S-E-1A	0.5769	0.3274	0.4522
	THCIB-S-E-2A	**0.5899**	**0.3406**	**0.4653**
	THCIB-S-E-3A	0.5544	0.3251	0.4397
	THCIB-S-E-4A	0.477	0.2784	0.3777
	THCIB-S-E-5A	0.5395	0.304	0.4218
@30	THCIB-S-E-1A	**0.693**	0.3177	**0.5054**
	THCIB-S-E-2A	0.6743	**0.3284**	0.5014
	THCIB-S-E-3A	0.6486	0.3244	0.4865
	THCIB-S-E-4A	0.5855	0.2691	0.4273
	THCIB-S-E-5A	0.6339	0.2986	0.4662

Discusscion

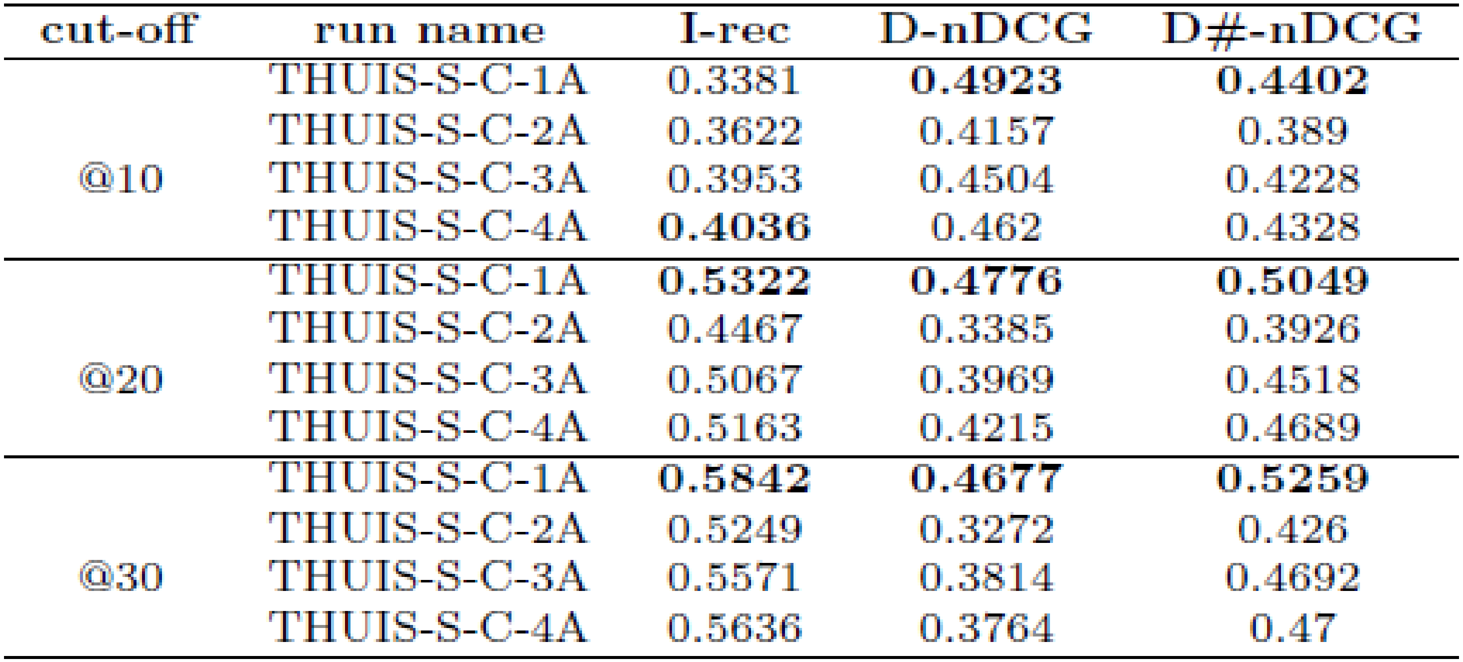
- THCIB-S-E-2A and THCIBS-E-1A
- Concept-based query expansion helps to recall more relevant subtopics

- THCIB-S-E-4A and THCIB-S-E-5A
- The revised AP algorithm outperforms the standard AP algorithms in most evaluation metrics

- THCIB-S-E-2A, THCIB-S-E-3A and THCIB-S-E-5A
- The unified ranking model do not bring performance gain on this dataset
  - the strategy we is relatively simple

Results

D#-nDCG of Chinese Subtopic Mining runs at various cut-off levels



cut-off	run name	I-rec	D-nDCG	D#-nDCG
@10	THUIS-S-C-1A	0.3381	**0.4923**	**0.4402**
	THUIS-S-C-2A	0.3622	0.4157	0.389
	THUIS-S-C-3A	0.3953	0.4504	0.4228
	THUIS-S-C-4A	**0.4036**	0.462	0.4328
@20	THUIS-S-C-1A	**0.5322**	**0.4776**	**0.5049**
	THUIS-S-C-2A	0.4467	0.3385	0.3926
	THUIS-S-C-3A	0.5067	0.3969	0.4518
	THUIS-S-C-4A	0.5163	0.4215	0.4689
@30	THUIS-S-C-1A	**0.5842**	**0.4677**	**0.5259**
	THUIS-S-C-2A	0.5249	0.3272	0.426
	THUIS-S-C-3A	0.5571	0.3814	0.4692
	THUIS-S-C-4A	0.5636	0.3764	0.47

Discusscion

- Similar observations are made on the Chinese task except:

- THUIS-S-C-4A and THUIS-S-C-1A
- Subtopic candidate clustering helps to improve I-recall @10

- THUIS-S-C-4A and THUIS-S-C-1A
- Query extension lead to performance degradation

- The difference in language features and search engines between Chinese and English

References

R. Song, M. Zhang, T. Sakai, M. P. Kato, Y. Liu, M. Sugimoto, Q. Wang, and N. Orii. Overview of the ntcir-9 intent task. In *Proceedings of NTCIR-9 Workshop Meeting*, pages 82–105, 2011.

W. Song, Y. Zhang, H. Gao, T. Liu, and S. Li. Hitscir system in ntcir-9 subtopic mining task. In *Proceedings of NTCIR-9 Workshop Meeting*, 2011.

S. Tetsuya, Z. Dou, T. Yamamoto, Y. Liu, M. Zhang, and R. Song. Overview of the ntcir-10 intent-2 task. In *Proceedings of NTCIR-10 Workshop Meeting*, 2013.

Y. Xue, F. Chen, T. Zhu, C. Wang, Z. Li, Y. Liu, M. Zhang, Y. Jin, and S. Ma. Thuir at ntcir-9 intent task. In *Proceedings of NTCIR-9 Workshop Meeting*, pages 123–128, 2011.

S. Zhang, K. Lu, and B. Wang. Ictir subtopic mining system at ntcir-9 intent task. In *Proceedings of NTCIR-9 Workshop Meeting*, pages 106–110, 2011.

J. Han, Q. Wang, N. Orii, Z. Dou, T. Sakai, and R. Song. Microsoft research asia at the ntcir-9 intent task. In Proceedings of *NTCIR-9 Workshop Meeting*, pages 116–122, 2011.

B. J. Frey and D. Dueck. Clustering by passing messages between data points. science, 315(5814):972–976, 2007.

Thank you!