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

J. Wang[1], G. Tang[1], Y. Xia[1], Q. Zhou[1], F. Zheng[1], Q. Hu[2], S. Na[2], Y. Huang[2]
[1]Tsinghua National Laboratory for Information Science and Technology, Tsinghua University
[2]Canon Information Technology (Beijing) Co. Ltd.

SUMMARY

- **TEAM**
  - THUIS team comprises of researchers from Intelligent Search group, Center for Speech and Language Technology, Tsinghua University
  - THCIB is a joint team between THUIS and Canon Information Technology (Beijing) Co. Ltd..

- **TASK**
  - SUBTOPIC MINING

  Systems are required to return a ranked list of ***subtopic strings*** in response to a given topic query while the top N subtopic strings should be ***both relevant and diversified*** as much as possible.

- **NOVELTY**
  - ***Concept***-based query analysis: converting the query into a set of concepts, which are extracted from the knowledge in Wikipedia

    "battles in the civil war"
    ➔ "battle", "civil war"

  - ***Sense***-based text clustering to discover intents underlying the subtopic candidates, which are extracted from multiple resources.

    Synonym and polysemy

  - ***Unified*** subtopic ranking model combining relevance, source importance and diversity

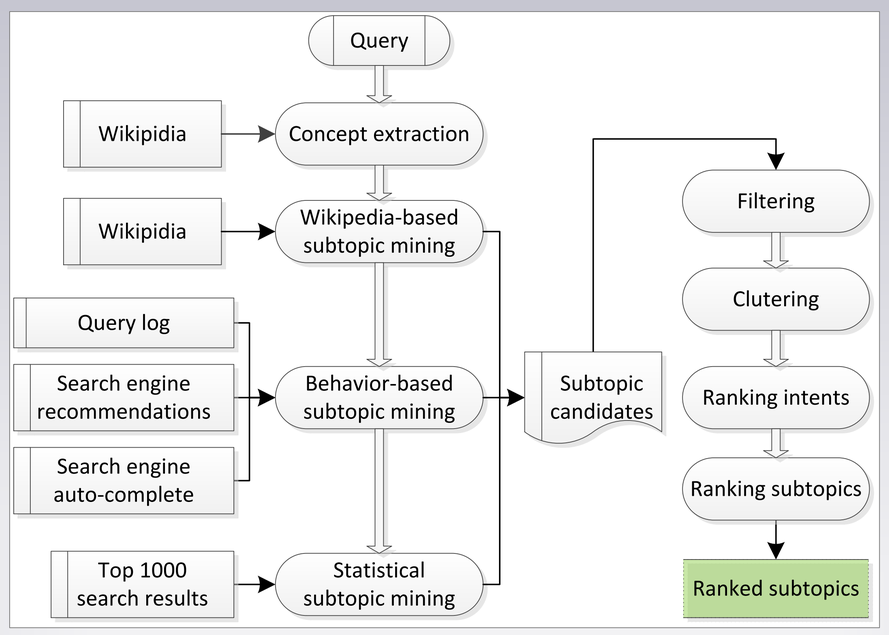
MOTIVATION

- **ISSUE #1**: Query is usually very short
- **SOLUTION #1**: Appling BIGGER CONTEXT in query understanding
  - General knowledge base: Wikipedia
  - User behavior data: Query log, search engine auto-completions and suggestions
  - Search results: Title and snippet

- **ISSUE #2**: Subtopic surface strings are many while redundancy is huge
- **SOLUTION #2**: Discover the implicit intents by clustering the subtopic surface strings
  - A sense-based clustering algorithm

- **ISSUE #3**: Relevance is no longer effective for intent ranking
- **SOLUTION #3**: Ranking intents with a unified model combining relevance and diversity

SYSTEM



Architecture of THCIB/THUIS intent mining system

METHODS

SUBTOPIC CANDIATE MINING (SCM)

1. Extracting Wikipedia concept(s) from Query
- Pre-processing: word segmentation, stemming and tokenization
- Wikipedia concept (entry) matching

2. Extending the Query
- Wikipedia synonyms (redirects and disambiguation pages)
- Intent schemas: manipulating concepts in the query, prepositions, and wildcard(s)
- Concept repositioning

3. Mining Subtopics in Wikipedia
- Wikipedia ambiguous entry: related concepts
- Wikipedia redirects: synonyms
- Wikipedia concept definition

4. Mining Subtopics in User Behavior Data
- Mining co-occurring relevant queries in Query log
- Search engine tools based on user behavior data: auto-completion, recommendations

5. Mining Subtopics in Search Results
- A word sense induction (WSI) framework (LDA)
- Extracting keywords from each topic as extensions of the query

SUBTOPIC CANDIDATE RANKING (SCR)

0. Subtopic Filtering

1. Assigning source importance score

2. Calculating relevance score

3. Finding intents in the subtopics
- A sense-based **subtopic** similarity measure (title and snippet of every top 20 search results using the subtopic as query).
- Affinity Propagation (AP) clustering algorithm

4. Entity analysis
- Exclusive entities

  "furniture for small spaces New York",
  "furniture for small spaces Los Angeles".
- Using Freebase to recognize the exclusive homogeneous entities

5. Calculating intent importance score

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

- relevance score of the subtopic
- importance score of the source

6. Subtopic selecting
- Iteratively get the top subtopic candidate in each cluster
- Assign penalty to the exclusive homogeneous entities

EVALUATION

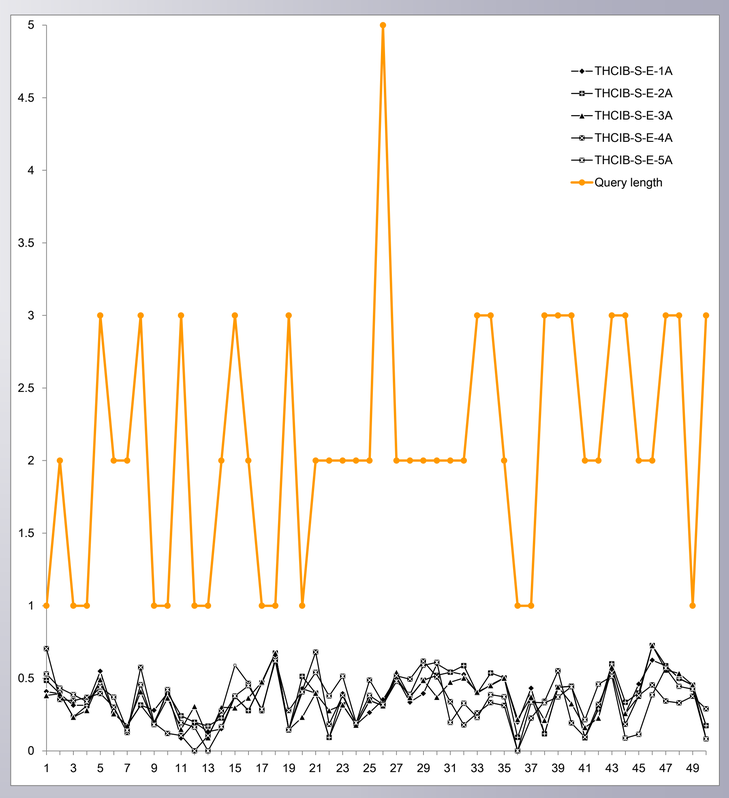
- Submitted RUNS
  - *THCIB-S-E-1A*: SCM (1.Concept extraction + 3.Wikipedia + 4.Querylog + 5.Search results + 6.Filtering) + SCR (source importance+ relevance)
  - *THCIB-S-E-2A*: THCIB-S-E-1A + SCM (2.Query expansion)
  - *THCIB-S-E-3A*: THCIB-S-E-2A +SCR(4.Entity analysis)
  - *THCIB-S-E-4A*: THCIB-S-E-3A +SCR(3.Intent mining with standard AP + 5.Intent ranking + 6. Subtopic selecting)
  - *THCIB-S-E-5A*: THCIB-S-E-4A + SCR (Revised AP)

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

Evaluation results of English Subtopic Mining runs

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

Evaluation results of Chinese Subtopic Mining runs



System performance upon English topics and query length

CONCLUSION

- Incorporating concepts and word senses in subtopic mining and ranking brings marginal performance gain.
- The unified intent ranking model is promising in producing satisfactory results. Further tuning is planned as the future work.

This work is supported by Canon Inc. (No. TEMA2012).