**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
  - 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

**SYSTEM**

**METHODS**

**SUBTOPIC CANDIDATE 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

   - 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

**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. TEMAZ2012).