THUIR at NTCIR-10 INTENT-2 Task

IR Group of Tsinghua University

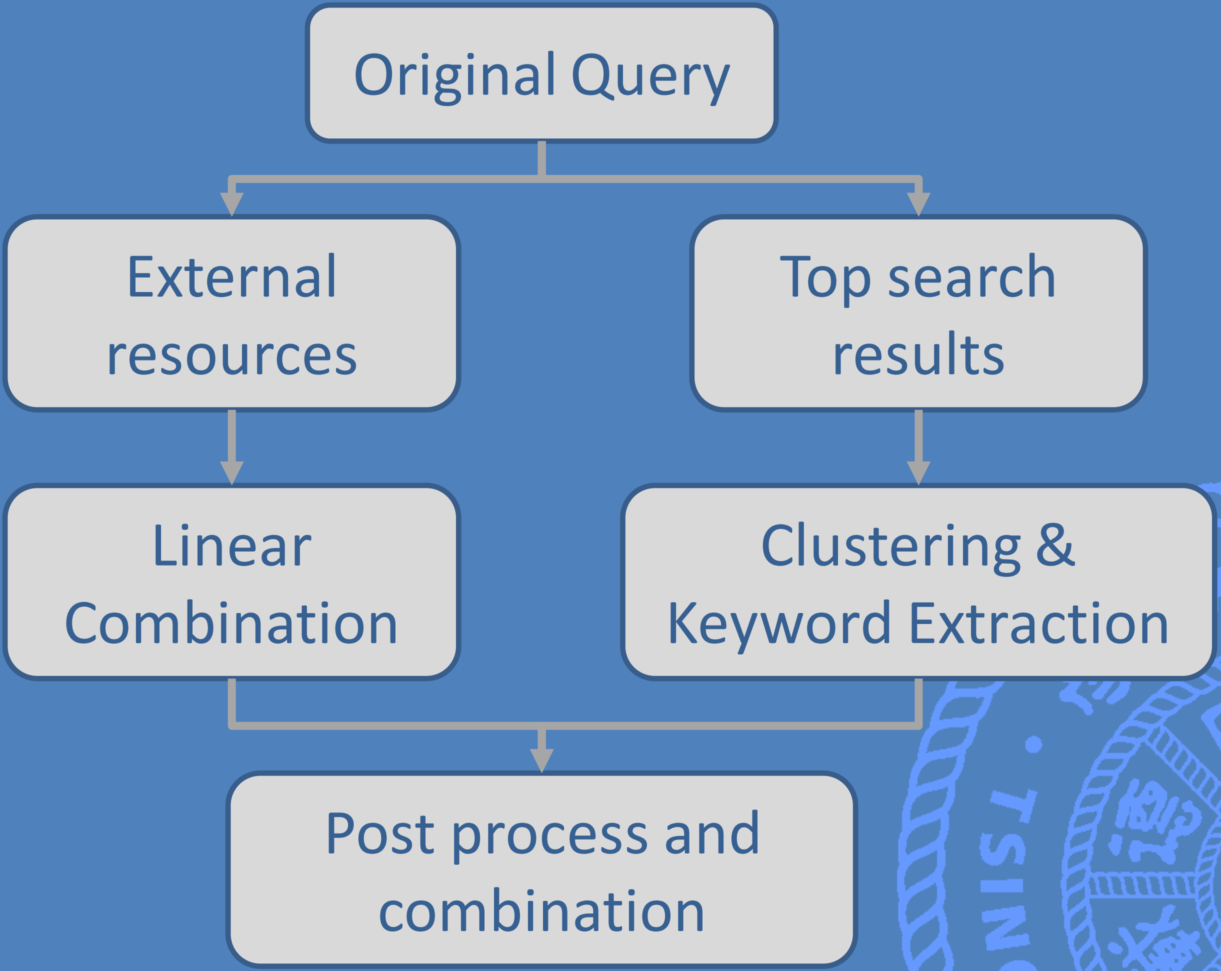
Yufei Xue, *Fei Chen*, Aymeric Damien, Cheng Luo,
Shuai Huo, Min Zhang, Yiqun Liu, Shaoping Ma

Overview

- THUIR@INTENT2: three subtasks
  - – English Subtopic Mining
  - – Chinese Subtopic Mining
  - – Document Ranking

English Subtopic Mining

- External resources v.s. Top search results



English Subtopic Mining

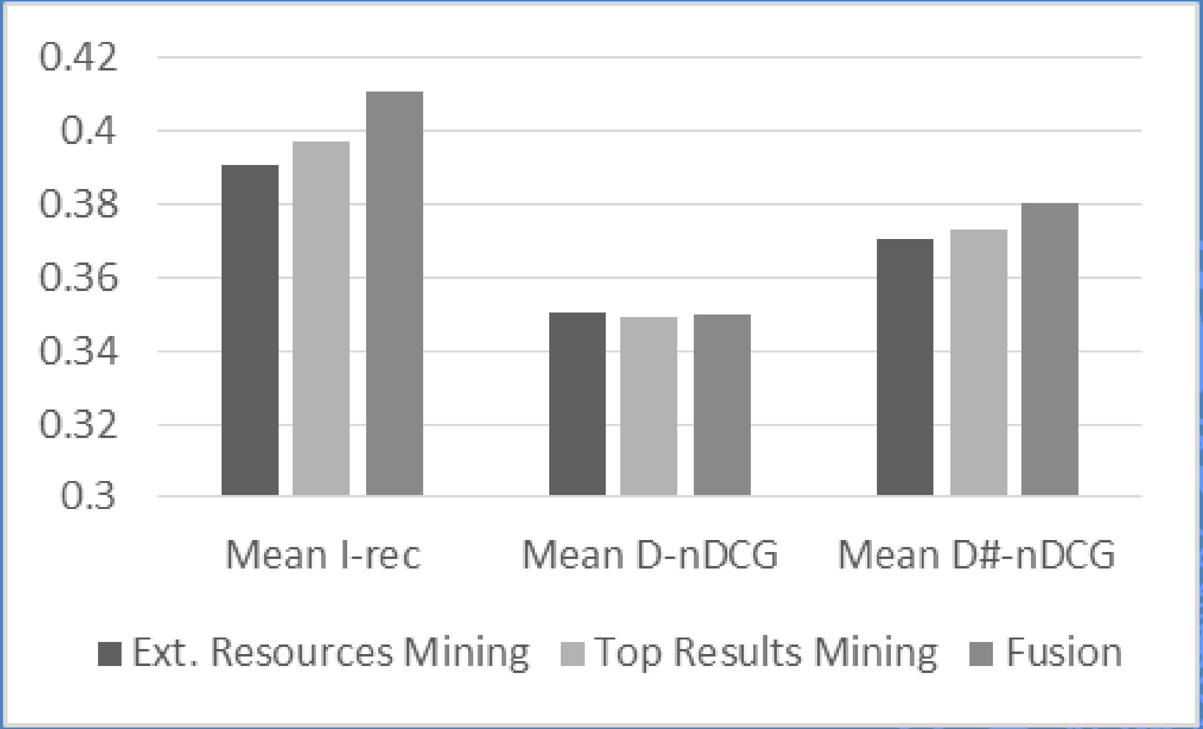
- External Resource Based Subtopic Mining
  - Subtopic candidate generation
    - Query Completion (Google, Bing, Yahoo)
    - Query Suggestion (Google, Bing, Yahoo)
    - Google Insights / Google Keywords Generator
    - Wikipedia (Disambiguation items)
  - Post process: Remove candidates without any query keywords
  - Linear combination
    - Google Insights: 0.15; Google Keywords Generator: 0.75; Query Suggestion / Completion: 0.05

English Subtopic Mining

- Top Results Based Subtopic Mining
  - Result document description
    - Search result snippets
    - Important fields of result documents ("h1", anchor, …)
    - BM25 scores are calculated for each word
  - Result clustering
    - PAM (Partitioning Around Medoids) algorithm
    - Without assigning the number of clusters
  - Keyword extraction for each cluster
    - Select the most frequent word and extend it to an n-gram.
  - Rank keywords by their clusters.

English Subtopic Mining

- Combination of subtopics
  - Linear combination
  - Duplication removing with WordNet
  - Normalization and re-ranking.

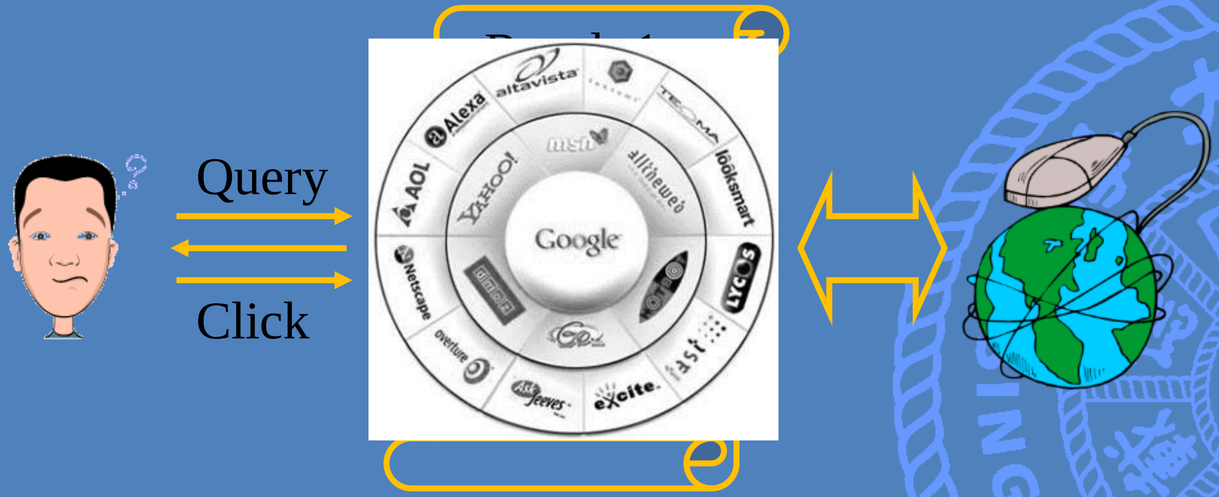


Chinese Subtopic Mining

- Candidate subtopic generation
  - Query suggestions collected from Google, Sogou, Baidu and Bing
  - Disambiguation items
  - collected from Hudong.com and Wikipedia
  - Keywords extracted from LDA topics generated on clicked snippets
- Candidate ranking
  - Credibility of external resources (e.g. Wikipedia = 2, Google = 2, Hudong = 1, …)
  - Number of common words
  - Length of the subtopic
  - Number of words in clicked snippets

Chinese Subtopic Mining

- Clicked snippets & user intent
  - User clicks a result => user is interested in the snippets of the results
  - Click-through information: SogouQ
- LDA on clicked snippets
  - 10 implicit topics for each query.



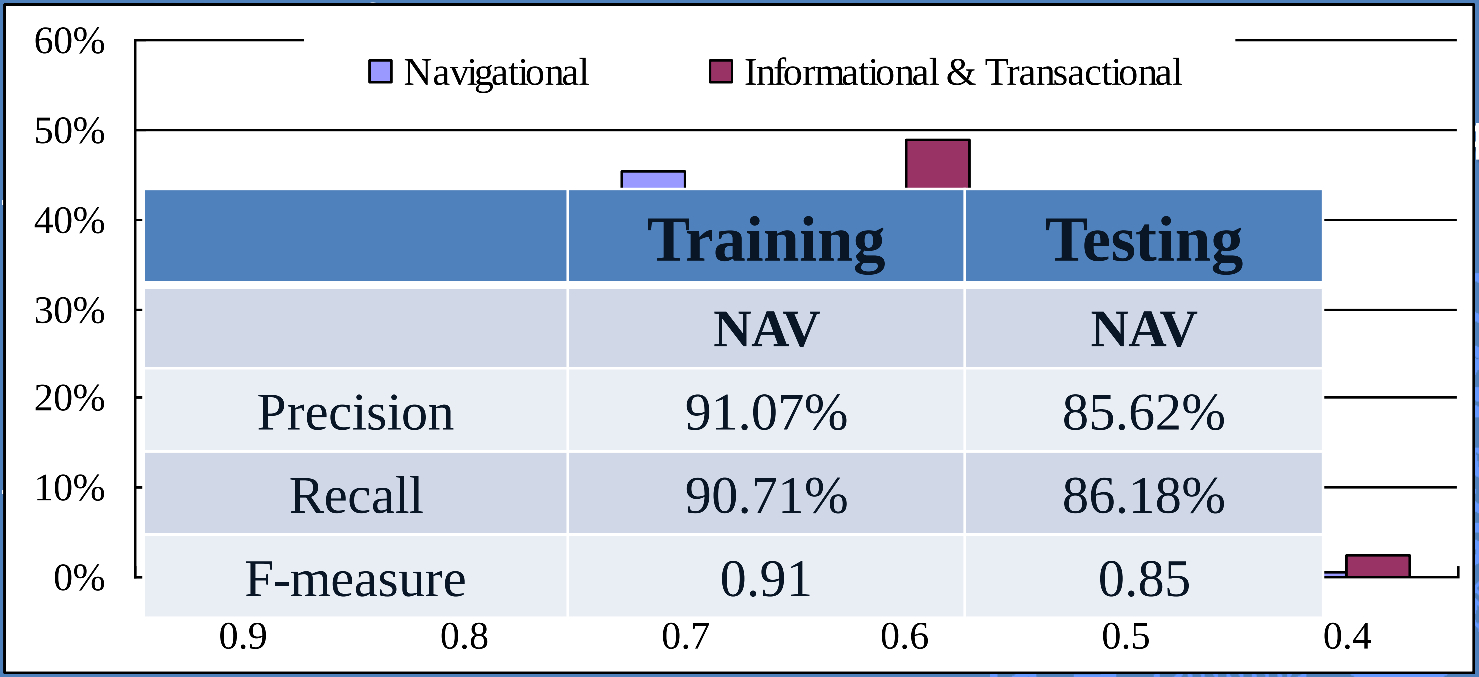
Chinese Subtopic Mining

- Result comparisons
  - Snippet click-through information helps improve candidate ranking
  - Candidates generated by LDA on snippets are not so effective

		I-rec@10	D-nDCG@10	D#-nDCG@10
1	Query suggestion	0.3792	0.4739	0.4266
2	1 + Snippet	0.3786	***0.5028***	***0.4407***
3	2+LDA	***0.3839***	0.4843	0.4341

Document Ranking

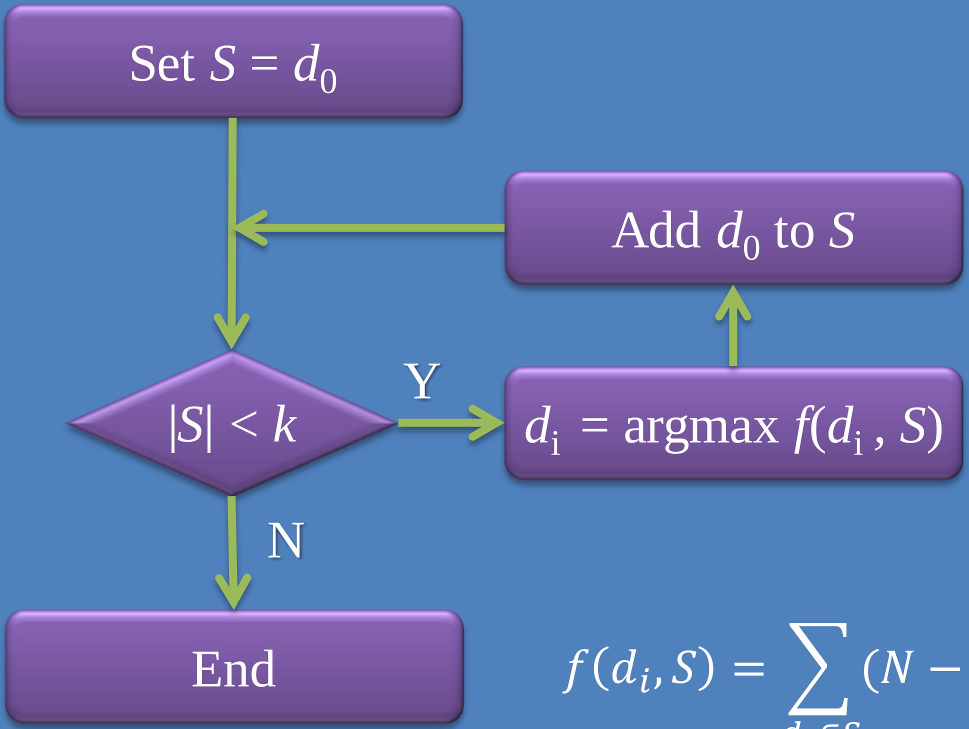
- Selective Diversification
  - Informational query:
    - IA-Select according to the D#-nDCG value of a document
- Query Type Identification
  - nCS(q):



	Training	Testing
	NAV	NAV
Precision	91.07%	85.62%
Recall	90.71%	86.18%
F-measure	0.91	0.85

Document Ranking

- Diversify Results Based on Novelty



$$f(d_i, S) = \sum_{d_j \in S} (N - r_j) + \lambda \times \cos(w_i, w_j)$$

Document Ranking

- Experimental Results

	I-rec@10	D-nDCG@10	D#-nDCG@10	DIN-nDCG@10	P+Q
Baseline	0.7247	***0.4207***	0.5727	0.2858	0.2653
Selectively diversification	0.6731	0.3587	0.5159	0.2611	0.2203
Novelty based diversification	***0.7258***	0.4201	***0.5729***	***0.2865***	***0.2663***

Thank you

THUIR