THUIR at NTCIR-12 IMine Task

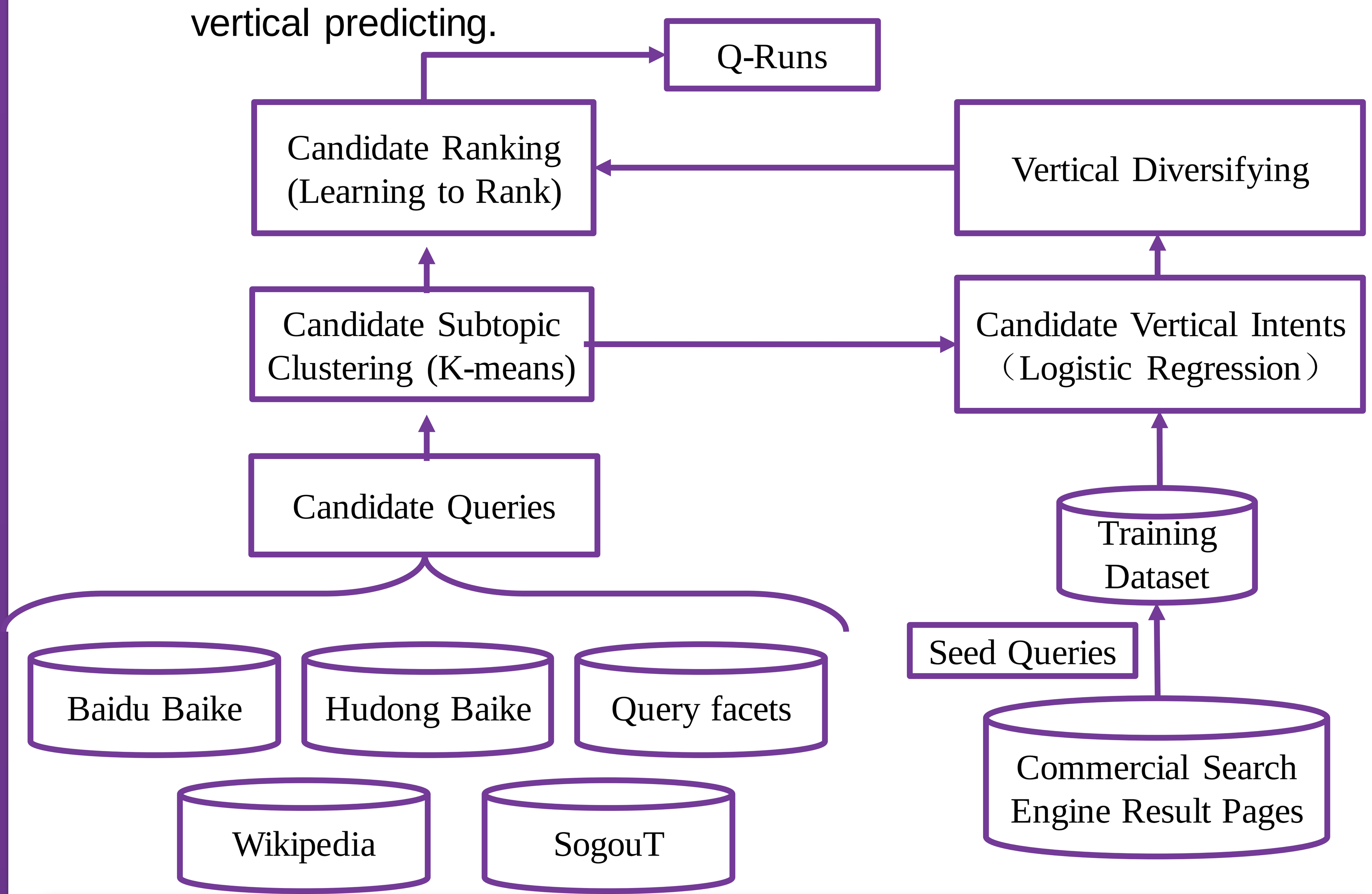
Zeyang Liu, *Ye Chen*, Rongjie Cai, Jiaxin Mao, Chao Wang, Cheng Luo, Xin Li, Yiqun Liu , Min Zhang, Huanbo Luan, Shaoping Ma

Information Retrieval Group, Department of Computer Sci. & Tech., Tsinghua University

chenye617@gmail.com

❖ Framework Overview

- **4-step framework in query understanding subtask**
  - Candidate mining, candidate clustering, candidate ranking and vertical predicting.

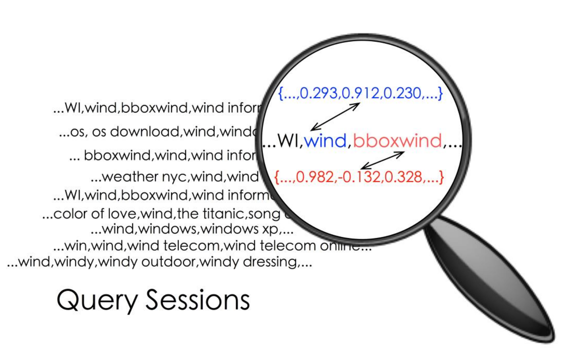


❖ Candidate Mining From Various Resources

- **Disambiguation Items from Wikipedia, Baidu Baike and Hudong Baike.**
- **Query Facets: Query Completions and Query Suggestions**
- **Query Reformulations extracted from SogouT**
- **Query Recommendations from Sogou Search Engine**

❖ Candidate Clustering with K-Means

- **Goal: Find diversified candidates**
- **Cluster candidates with K-means algorithm**
- **Query vector representation**
  - Word embedding trained based on SogouT dataset
  - Long query candidate: average word embedding of the words
- **Cluster candidates into n clusters**
  - n = 5 or 10
- **Candidate evaluation**
  - Clusters



$c_1, c_2, c_3, \cdots, c_n$

  - Inner distance

$$S_{inner}(q) = \frac{\sum_{q_k \epsilon c_i,\, q_k \neq q} dist(q, q_k)}{|c_i| - 1}, q \epsilon c_i$$

  - Outer distance

$$S_{outer}(q) = \min\{\frac{\sum_{q_k \epsilon c_i} dist(q, q_k)}{|c_i|}\}(j = 1,2,\cdots,n, j \neq i), q \epsilon c_i$$

  - Candidate score

$$S(q) = \frac{S_{outer}(q) - S_{inner}(q)}{\max\{S_{outer}(q), S_{inner}(q)\}}$$

❖ Candidate Ranking with Learning to Rank

- **Goal: Find high quality subtopic candidates**
- **Rank candidates with Learning To Rank algorithm (RankBoost)**
- **Features:**

Text similarity: length difference, Jaccard similarity, edit distance…

Word embedding: average, medium, top 3 average of cosine similarities

Search Result Similarity: number of shared results…

- **Metric to optimize: NDCG@10**
- **Training set: Ranked Subtopics from NTCIR-11 Imine**

❖ Vertical Predicting

- **Training query**
  - Seed Urls generated from an Open Directory Project
  - Random walk on a click-through bipartite graph
- **Vertical Distribution**
  - Vertical information collected from search result pages

$$\mathrm{P-score} = \frac{1}{\log(1 + R_i)}$$

- **Model Construction**
  - Logistic Regression
  - Word embedding query representation
  - Six prediction models for each type of verticals
- **Vertical Diversification**
  - Empirical rules
  - Replace the top 3 Web intent with corresponding vertical intent
  - The top vertical result is chosen to be the vertical intent if there are more than two verticals in the top 3 positions

❖ Experimental Results

RUNNAME	SYSTEM DESC.	D#-nDCG	V-score	QU-score
THUIR-QU-1A	Cluster all subtopic candidates into 10 clusters and select the candidate with the highest S(q) from each cluster.	0.5204	0.5579	0.5392
THUIR-QU-2A	Cluster all subtopic candidates into 5 clusters and select two candidates with the highest two S(q) from each cluster.	0.5550	0.5506	0.5528
THUIR-QU-1B	Rerank the 10 subtopics generated by THUIR-QU-1A with learning to rank algorithm.	0.5368	0.5763	0.5565
THUIR-QU-2B	Rerank the 10 subtopics generated by THUIR-QU-2A with learning to rank algorithm.	0.5436	0.5686	0.5561
THUIR-QU-3A	Cluster all subtopic candidates into 5 clusters and select the candidate with the highest ten S(q).	0.4973	0.5942	0.5458

❖ Retrieval Models

- **Probabilistic model based on BM25 and our previous proposed word pair model.**
- **Relevance score for subtopic**
  - $R(q, D) = W_{BM25} + \alpha \cdot W_{wp}$
  - $W_{BM25} = \sum_{i=1}^{m} log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})}$
  - $W_{wp} = \sum_{i=1}^{m} log \frac{N - n(q_i q_{i+1}) + 0.5}{n(q_i q_{i+1}) + 0.5} \cdot \frac{f(q_i q_{i+1}, D) \cdot (k_1 + 1)}{f(q_i q_{i+1}, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})}$
- **Relevance score for query**
  - $R(Q, D) = \sum_{i=1}^{10} R(q_i, D) \times S(q_i)$
- **Vertical importance based on subtopic candidate score**
  - $I(v) = \alpha \cdot S - score(v)$
- **Combination of $R(Q, D)$ and $I(v)$**

❖ Experimental Results

RUNNAME	D#-nDCG (unclear topics)	nDCG (clear topics)	D#-nDCG+nDCG (all topics)
THUIR-QU-1A	0.6677	0.5756	0.6594
THUIR-QU-2A	0.6664	0.5652	0.6573
THUIR-QU-1B	0.6594	0.5416	0.6488
THUIR-QU-2B	0.6632	0.5442	0.6525
THUIR-QU-3A	0.6429	0.5506	0.6346