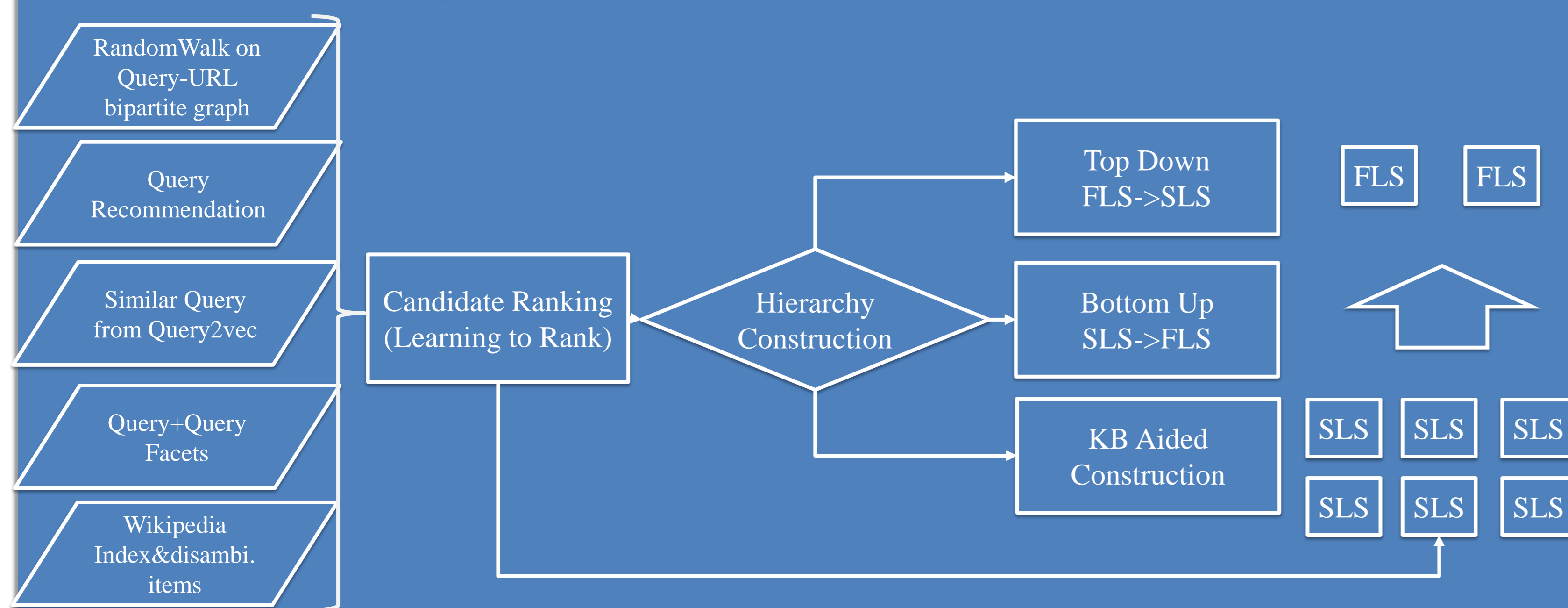


Cheng Luo, Xin Li, Alisher Khodzhaev, Fei Chen, Keyang Xu,
Yujie Cao, Yiqun Liu, Min Zhang, Shaoping Ma
Department of Computer Science and Technology,
Tsinghua University, Beijing, China
c-luo12@mails.tsinghua.edu.cn

Subtopic Mining

Framework

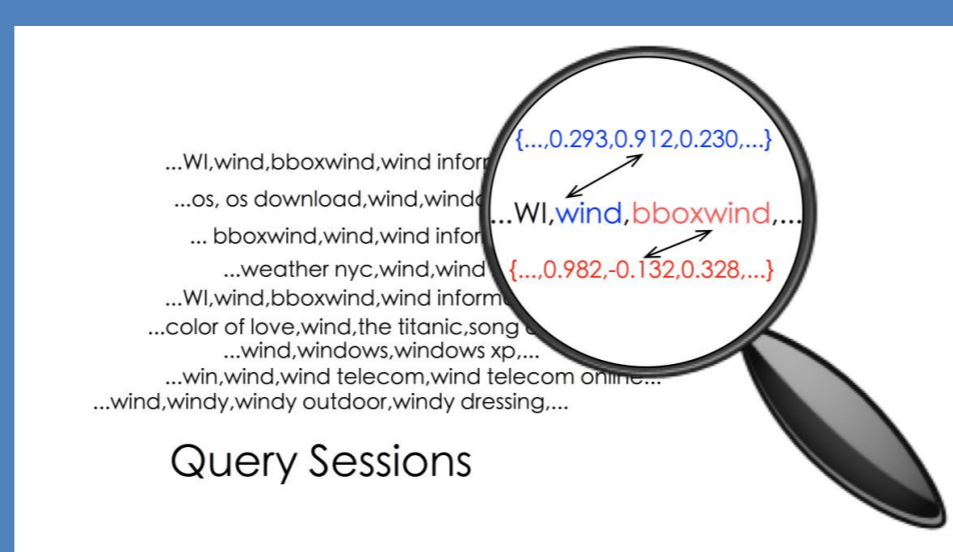
We propose a 3-step framework in Subtopic Mining Subtask: Candidate Mining, Candidate Ranking and Hierarchy Construction.



- Notion:
FLS: First Level Subtopic
SLS: Second Level Subtopic
KB Aided: Knowledge-Based Aided

Candidate Mining From Various Resources

- Similar Queries from *Query Recommendation*, *Random Walk on Query-URL Bipartite Graph* and *Query2vec*.
- Query + Query Aspect from *Query Facets*, *Wikipedia Indexes and Disambiguation Items*.
- *Query2vec*
- Query \leftarrow words
- Session \leftarrow Sentence
- Each query can be represented as a vector.
- Find Similar Queries with cosine similarity.



Candidate Ranking with LTR Algorithms

- Goal: Find the high quality subtopic candidates
- Rank candidates using Learning To Rank algorithm
- Training Set: ranked subtopics from NTCIR Intent-2 data
- Feature: Similarity between query and candidate
Text similarity: Length difference, Jaccard similarity, Edit Distance
Search Result Similarity: number of shared results...
- Metric to optimize: NDCG@50

Method	NDCG@50 training	NDCG@50 testing
MART	0.8012	0.6951
RankNet	0.683	0.6675
RankBoost	0.743	0.7303
AdaRank	0.7049	0.7034
Coordinate Ascent	0.7037	0.676
LambdaMART	0.8274	0.688
ListNet	0.6912	0.6959
Random Forests	0.7896	0.6981

	波斯猫	云轩	遮天
1	波斯猫歌词	云轩阁 阳神	4399太古遮天
2	波斯猫歌曲	云轩阁小说	遮天txt下载
3	波斯猫眼睛	云轩阁 盘龙	遮天快眼看书
4	波斯猫 歌词	云轩阁小说网	太古遮天官网
5	波斯猫的眼睛	云轩阁txt下载	百度遮天官网
...
-5	孟买	德州	书库
-4	蜃	嘉兴	题材
-3	浓度	海口	媒体
-2	洛威拿	邢台	类型
-1	眼大	金华	唐砖

Hierarchy Construction in Three Ways

- Top-Down Hierarchy Construction
Find the FLSs first and classify other candidates into FLS categories
A heuristic method to pick out FLSs.
$$Score = a * Novelty - b * \frac{candidate\ length}{query\ length} + c * Relvance + d * Frequency$$
- Bottom-Up Hierarchy Construction
Cluster all the candidates, for each cluster, choose the best one as FLS
N-gram ranked by Learning to Rank Algorithms/Metric to optimize:P@5
- Knowledge Base Aided Construction
Use the Wikipedia Indexes and Disambiguation Items as FLSs
Classify all the Candidates into FLS categories
- Clustering: Using TF-IDF vectors extracted from snippets/titles on SERP
- Classifying: Linear Regression Classifier learnt from INTENT-2 results.

Experimental Results

RUNNAME	SYSTEM DESC.	H-Measure
THUSAM-C-1A	[Bottom Up] Cluster SLS candidate, find the highest-frequency n-gram which can match one of the candidate as FLSs.	0.2773
THUSAM-C-2A	[Bottom Up] Cluster SLS candidate, for each cluster, Learning to Rank the n-gram, find the best ones as FLSs.	0.2204
THUSAM-C-3A	[KB Aided] For queries which appears in Encyclopedia, use the disambiguation items (indexes) as FLS and classify other candidates.	0.1400
THUSAM-C-4A	[Top Down] Learning to Rank SLS candidates, use heuristic greedy select algorithm to find FLSs, and classify other candidates.	0.1404
THUSAM-C-5A	[Top Down] Learning to Rank n-grams as FLSs and classify other candidates.	0.2224
THUSAM-E-1A	[Bottom Up] Extraction from multiple resources (all) + tuned bottom-up hierarchical clustering	0.4257
THUSAM-E-2A	[Top Down] Extraction from multiple resources + up-bottom approach	0.1179

Document Ranking

Documents Retrieval Models

Probabilistic model is leveraged for document ranking, which is based on BM25 and combined with our previous proposed word pair model.

$$R(Q, D) = W_{BM25} + \alpha_1 \cdot W_{wp}$$

$$W_{BM25} = \sum_{i=1}^m \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})}$$

$$W_{wp} = \sum_{i=1}^m \log \frac{N - n(q_i q_{i+1}) + 0.5}{n(q_i q_{i+1}) + 0.5} \cdot \frac{f(q_i q_{i+1}, D) \cdot (k_1 + 1)}{f(q_i q_{i+1}, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})}$$

Result re-ranking with HITS

Top m documents sorted by either Authority or Hub Value in the search result are placed up to the front. Its new rank is determined as follows:

$$R_{new} = R_{old} - R_{old} \times (Authority + Hub)$$

Pruned Exhaustive Search

Previous studies have demonstrated that finding the optimal solution for diversified search is NP-hard

THEOREM: Given $k=l+1$, if there exists a document pair d_l and d_k that satisfies:

$$(G_{kl} - G_{kk}) - (G_{ll} - G_{lk}) > 0$$

The document list containing d_l in its l -th slot and d_k in its k -slot cannot be optimal diversified search result.

Notion:

G_{li} denotes the score for doc_k in the l -th slot

Pruned Exhaustive Search based on the THEOREM

ALGORITHM Pruned Exhaustive Search
INPUT all the selected documents D , the required number of documents L
 1 $S \leftarrow \Phi$, $maxG \leftarrow 0$
 2 function **recursion_full_search**($curD$, $leftD$, d_p , $curG$)
 3 if ($leftD$ is Φ or $|curD|=L$) and $curG > maxG$
 4 $maxG \leftarrow curG$
 5 $S \leftarrow curD$
 6 else
 7 $n \leftarrow |curD|$
 8 foreach d_j in $leftD$
 9 if ($G_{in} - G_{i(n+1)}) - (G_{jn} - G_{j(n+1)}) \geq 0$
 10 **recursion_full_search**($curD \cup \{d_j\}$, $leftD / \{d_j\}$, d_j , G_{j1})
 11 end function
 12 foreach d_i in D
 13 **recursion_full_search**($\{d_i\}$, $D / \{d_i\}$, d_i , G_{i1})
 14 return S

Experimental Results

RUNNAME	SYSTEM DESC.	Coarse-grained D#nDCG	Fine-grained D#nDCG
THUSAM-C-1A	Exhaustive search with window size 4. The SM result is from Subtopic N-gram Learning to rank list.	0.6965	0.6127
THUSAM-C-1B	Exhaustive search with window size 5. The SM result is from Subtopic N-gram Learning to rank list.	0.6943	0.6106
THUSAM-C-2A	Exhaustive search with window size 4. The SM result is from heuristic greedy select from subtopics.	0.3502	0.2623
THUSAM-C-2B	Exhaustive search with window size 5. The SM result is from heuristic greedy select from subtopics.	0.3697	0.2711