THUSAM@NTCIR-IMine

Cheng Luo, Xin Li, Alisher Khodzhaev, Fei Chen, Keyang Xu, Yujie Cao, Yiqun Liu, Min Zhang, Shaoping Ma

Tsinghua University

Dec 11th, 2014

Query Specificity Based Taxonomy

- Specificity based Taxonomy (Song et al., 2008)
  - Ambiguous: a query that has more than one meaning
  - Broad: a query that covers a variety of subtopics, and a user might look for one of the subtopics by issuing another query
  - Clear: a query that has a specific meaning and covers a narrow topic
  - **According to their results, 16% of queries in a real search log are ambiguous.**

曖昧

（ambiguous）

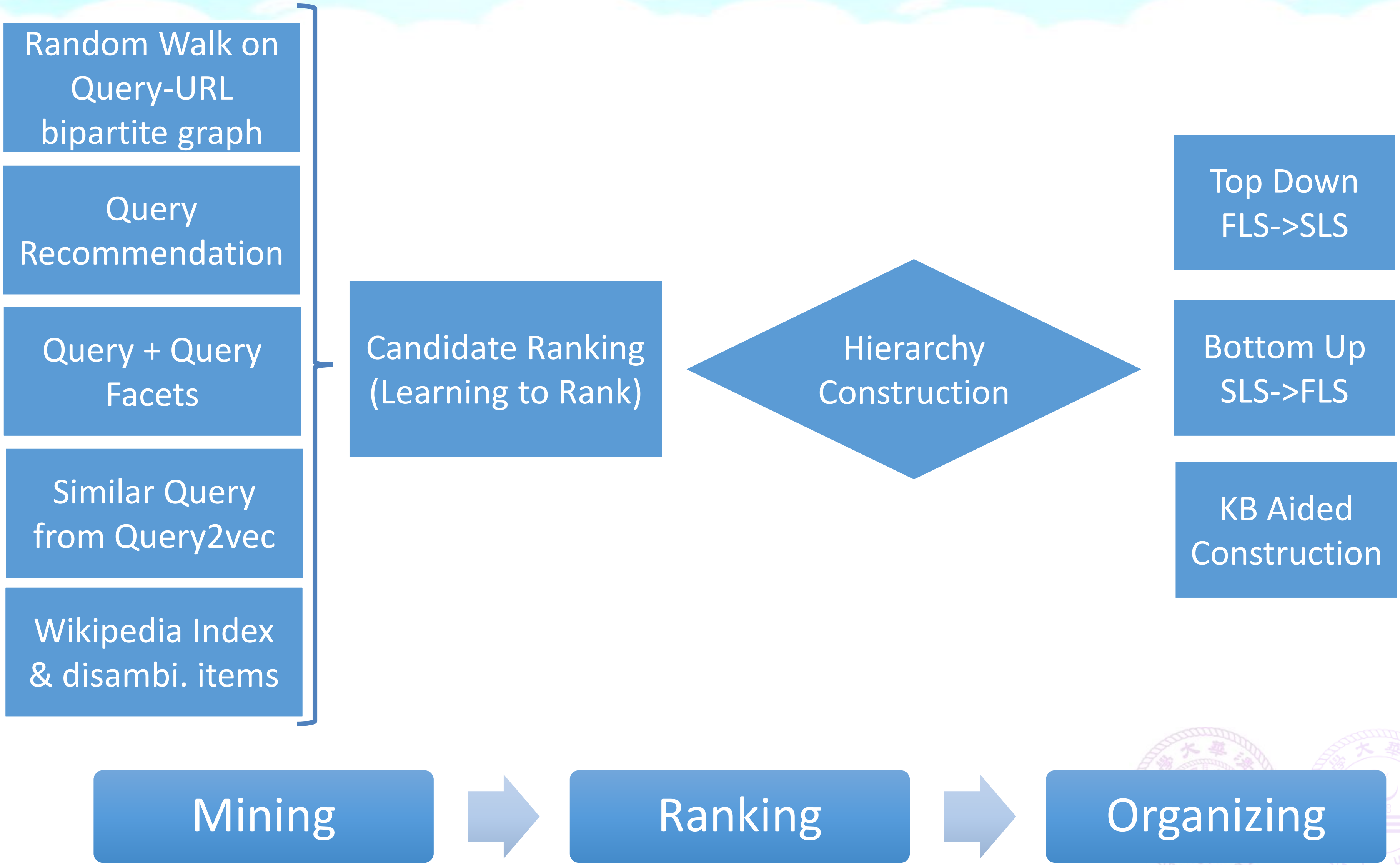
Subtopics for Diversified Search

- For an ambiguous query issued by user, if the search engine do not know about the user's profile and search context, the best it can do is to provide a diversified result list.
- Two diversified strategies:
  - Implicit: Diversified by the differences between documents
    - Result -> Diversified Results
  - **Explicit: Diversified by subtopics of the query**
    - **Result & Subtopic -> Diversified Results**

NTCIR: From INTENT to IMine

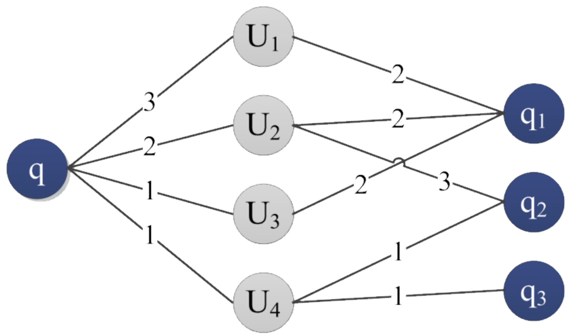
- Goal: explore and evaluate the tech. of satisfying different user intents behind Web search queries.
- Subtasks:
  - **Subtopic Mining(C\E)**: generate a two-level hierarchy of underlying subtopics
  - **Document Reranking(C)**: return a diversified ranked list of no more than 100 results for each query
  - Task Mining: to understand the relationship among tasks for supporting the Web searchers.

Subtopic Mining Framework



Candidate Mining

- Random Walk on Query-URL bipartite graph



$$p_{i,j} = \sum_{k \in U} p_{i,k} \cdot p_{k,j}$$

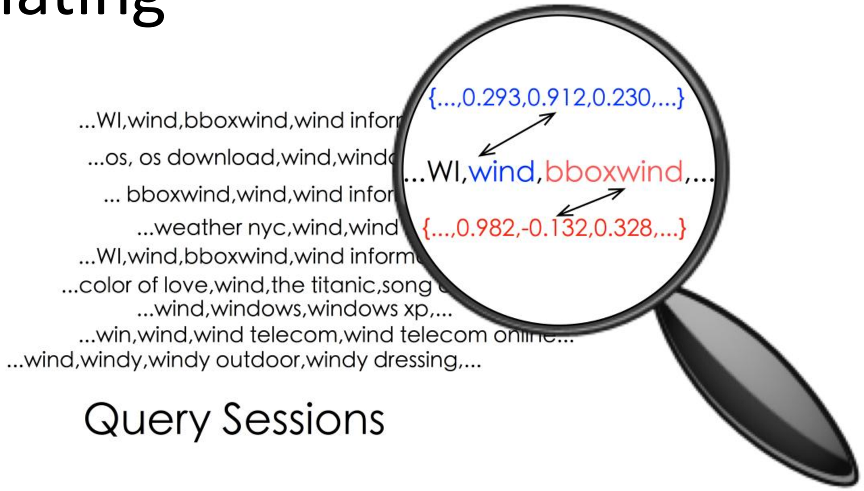
- Query Recommendations from SERP
- Query + Query Facets (Dou et al.,2011)
  - Extracted from lists on top-retrieved search results
  - Clusters of facets are also useful but noisy

Candidate Mining

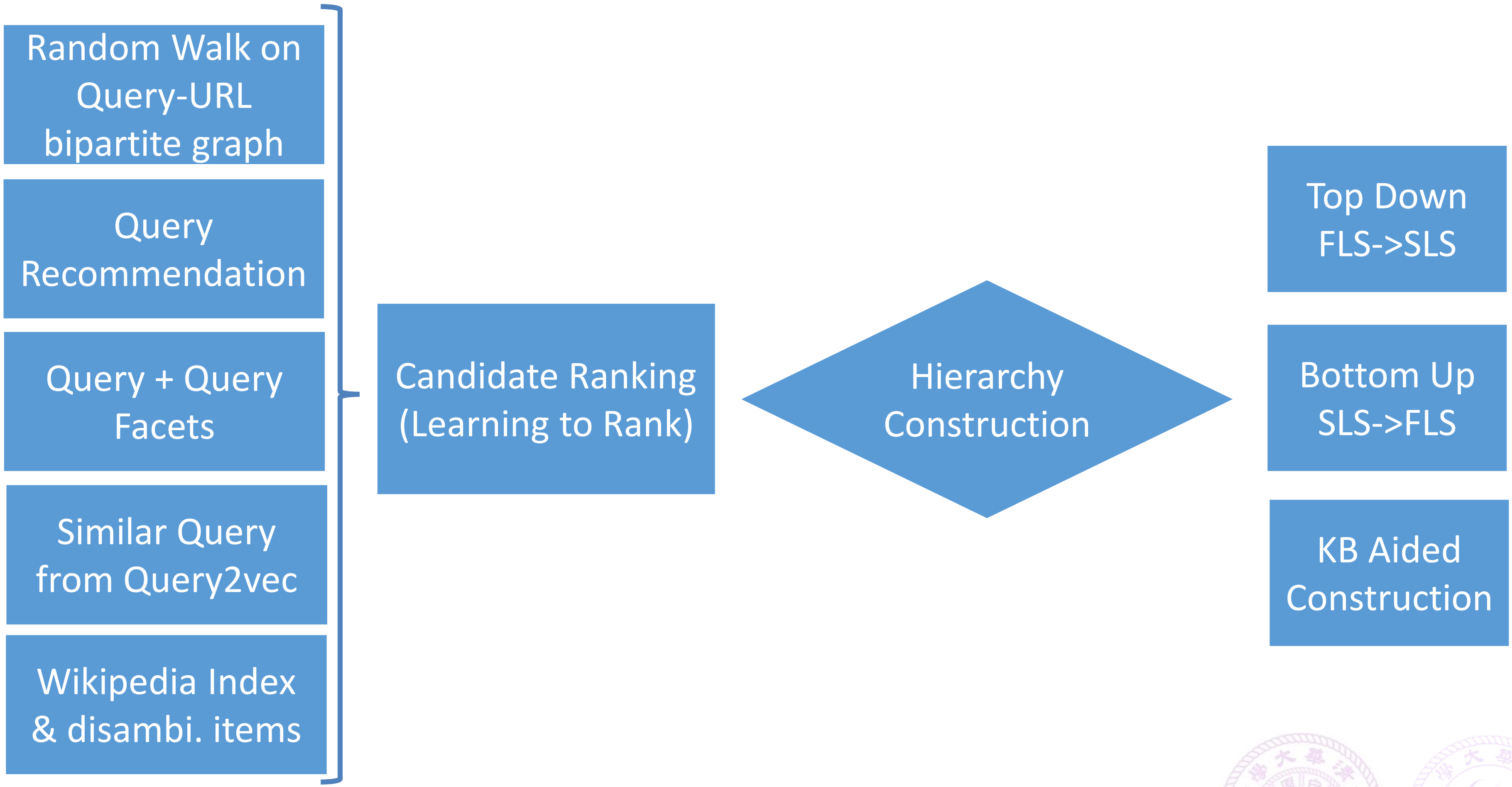
- 37/50 queries can be linked to specific pages on encyclopedia.
- All the disambiguation items and indexes are well organized.
- They also contribute to hierarchy construction

Candidate Mining

- Query2vec
  - Inspired by word embedding approach *word2vec*
  - query <- word session <- text
  - Each query can be represented as a vector
  - Find similar queries by calculating cosine similarity.
  - ***Similarity*** means that the queries carry ***similar intents***.



Subtopic Mining Framework



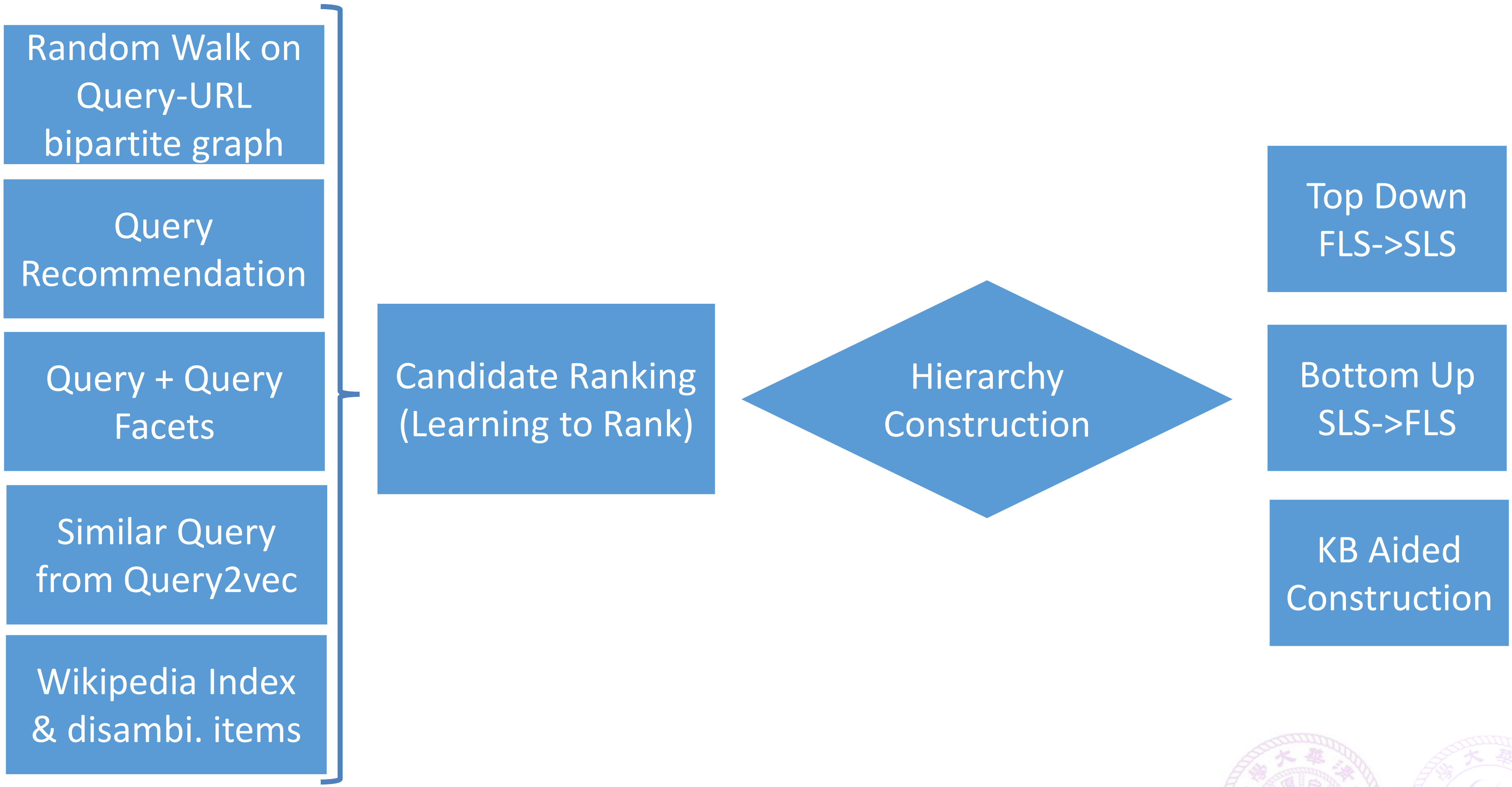
Candidate Ranking

- Background: Candidates are really ***noisssy***!
- Goal: Find the high quality subtopic candidates
- Rank candidates using Learning To Rank algorithm (RankBoost)
- Feature: Similarity between query and candidate and other signals
  - Text similarity: length difference, Jaccard similarity, edit distance…
  - Search Result Similarity: number of shared results…
  - If the candidate act as a query recommendation…
- Metric to optimize: NDCG@50
- Training set: Ranked Subtopics from INTENT-2

Candidate Ranking Examples

Query	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 故事	云轩阁 盘龙	遮天快眼看书
4	波斯猫 习性	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	靥	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

Subtopic Mining Framework



Hierarchy Construction

- Top-Down First Level Subtopic (FLS) Construction
  - Select representative query candidates/n-grams as FLSs
- Bottom-Up FLS Summarization
  - Cluster the candidates
  - For each cluster, summarize a FLS using n-gram information
- Knowledge Base Aided Construction
  - Use the items on wiki indexes/disambi. items as FLSs

Top-Down FLS Construction

- Select ***representative*** query candidates/n-grams as FLSs.
- Consider quality and diversity.
- A Heuristic Greedy Select Algorithm:
  - Consider 1) Novelty Based on chosen Candidates; 2) Query length; 3) Relevance; 4) Query Frequency (if appears in our query log)
  - In each step, select a best candidate with highest score which linearly combine the four factors
  - A group of parameters learnt from INTENT-2 annotations.
- Pairwise Evaluation in Selected Candidates: Error rate 11.2%

Top-Down FLS Examples

Query	First Level Subtopics				
先知	虚空先知	先知电子狗	先知x 808	DOTA先知出装	先知 电影
波斯猫	波斯猫糖果	波斯猫价格	波斯猫 女士黑皮衣	波斯猫 舞蹈教学视频	波斯猫论坛
猫头鹰	猫头鹰nh u9b	猫头鹰视频看看	DNF猫头鹰	猫头鹰生活习性	电影猫头鹰
Adobe	Adobe Reader X	Adobe Lightroom	Adobe Flash Player	Adobe Photoshop	Adobe Acrobat Professional
传奇	王菲+传奇	星际传奇2	传奇世界私服	传奇客户端下载	传奇小说
小米	小米m2	小米粥	小米红米2代	小米+官网	小米1s青春版
中国水电	中国水电权益变动报告书	中国水电融资融券信息	中国水电十五局四公司	中国水电集团	中国水电建设集团港航建设有限公司
三字经	三字经的译文	三字经mp3下载童声	三字经全文带拼音	幼儿三字经舞蹈	三字经儿歌

Bottom-Up FLS Construction

- Cluster the candidates
- Extract N-grams
  - Extract N-grams from the titles on SERPs of all the candidates in the cluster
  - N ranges from *query.length+1* to *query.length+10*
- Name cluster
  - Choose the shortest N-gram that matches a candidate in the cluster
  - Rank N-grams using LTR

Choose N-gram with LTR

- Goal: Find the best intent to represent the candidate cluster.
- Features:
  - Intent.length – query.length
  - Whether intent appears in SLS
  - Average reciprocal of the intent first shown position in SERP title of SLS
  - Accumulative score of intent in SERP title of SLS
  - Average reciprocal of the intent first shown position in SERP summary
  - Accumulative score of intent in SERP summary of SLS
  - Text Jaccard similarity with second level subtopics
  - URL Jaccard similarity with second level subtopics
- Metric to optimize: P@5

Knowledge Base Aided FLS Construction

- 37/50 Queries has KB pages.
- For the queries which have disambiguation pages, we use the disambiguation items as FLSs.
- All of the Indexes are used as candidates.

- 先知：汉语词语
- 先知：宗教先知
- 先知：动画《喜羊羊与灰太狼》角色
- 先知：游戏光晕系列人物
- 先知：游戏《魔兽争霸III》的英雄
- 先知：游戏dota中的英雄
- 先知：纪伯伦著书籍
- 先知：美国亚历克斯·普罗亚斯执导电影
- 先知：先知电子有限公司
- 先知：阿拉伯散文诗
- 先知：游戏《星际争霸2》中兵种
- 先知：关智斌演唱的歌曲

[yù gāng]
浴缸

1 按摩类
2 铸铁类
3 折叠类
4 发展现状
5 选用要点
- 分类
- 规格
- 类型比较
- 执行标准
- 要点
- 相关产品
- 发展趋势

6 选购须知
- 款式材料
- 尺寸形状

7 材质分析
- 亚克力类
- 铸铁类
- 木质类
- 钢板类

8 折叠特点
9 选择技巧

Clustering & Classification

- Clustering
  - Candidate Enrichment with the SERP of all candidates
  - Use TF-IDF of words as feature
  - K-means (6 clusters)
- Classification
  - Features: Text Similarity/ #Shared Results (URL)
  - Accuracy: F-measure 0.59 on 6 categories
  - A linear regression classifier learnt from INTENT-2 annotations.

Comparison of the 3 Strategies

- Top-Down
  - Advantage: Readability
  - Disadvantage: There is not necessarily a candidate that can represent the subtopic
- Bottom-Up
  - Advantage: Representative
  - Disadvantage: Not necessarily readable
- KB Aided
  - Advantage: Well organized
  - Disadvantage: the *cold* items

Subtopic Mining 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

- Ranking Models
  - Leveraged for document ranking, which is based on BM25 and combined with our previous proposed *word pair* model.

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

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

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

- Reranking with HITS algorithm

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

Pruned Exhaustive Search

- Previous studies have demonstrated that finding the optimal solution for diversified search is NP-hard (max-cover) problem.

> **THEORM**: *Given k=l+1, if there exist a document pair d_l and d_k that satisfies:*
>
> $$(G_{kl}-G_{kk}) - (G_{ll}-G_{lk})>0$$
>
> *where G_{kl} denotes the score for d_k in the l-th slot. The document list containing d_l in its l-th slot and d_k in its k-slot cannot be optimal diversified search result.*

- We can stop search in this branch if such *Ordered Pair* detected.

Pruned Exhaustive Search

- Based on this observation, we proposed a Pruned Exhaustive Search algorithm.
- Decrease the complexity without performance loss.
- Further optimize with *Search Window Strategy,* only need to exhaustively search for the optimum within the *W* slots. *SW cannot* guarantee *to optimal results.*

ALGORITHM Pruned Exhaustive Search

```
INPUT all the selected documents D, the required number of docments L
S←Φ, maxG←0
function recursion_full_search(curd,leftD,d_i,curG)
    if(leftD is Φ or |curd|=L) and curG>maxG
        maxG←curG
        S←curD
    else
        n←|curD|
        foreach d_j in leftD
            if (Gin-Gi(n+1)) – (Gjn-Gj(n+1))≥0
                recursion_full_search(curD ∪ {d_j}, leftD / {d_j} ,d_j, G_j1)
end  function
foreach d_i in D
  recursion_full_search( {d_i}, D / {d_i} ,d_i , G_i1)
return S
```

Document Ranking 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

Thank you!

luochengleo@gmail.com
www.thuir.cn