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.
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

- Random Walk on Query-URL bipartite graph
- Query Recommendation
- Query + Query Facets
- Similar Query from Query2vec
- Wikipedia Index & disambi. items

Mining → Ranking → Organizing

Candidate Ranking (Learning to Rank) → Hierarchy Construction

Top Down FLS->SLS
Bottom Up SLS->FLS
KB Aided Construction
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

- Random Walk on Query-URL bipartite graph
- Query Recommendation
- Query + Query Facets
- Similar Query from Query2vec
- Wikipedia Index & disambi. items

Hierarchy Construction:
- Top Down FLS->SLS
- Bottom Up SLS->FLS
- KB Aided Construction
Candidate Ranking

• Background: Candidates are really noisy!

• 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

<table>
<thead>
<tr>
<th>Query</th>
<th>波斯猫</th>
<th>云轩</th>
<th>遮天</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>波斯猫歌词</td>
<td>云轩阁 阳神</td>
<td>4399太古遮天</td>
</tr>
<tr>
<td>2</td>
<td>波斯猫歌曲</td>
<td>云轩阁小说</td>
<td>遮天txt下载</td>
</tr>
<tr>
<td>3</td>
<td>波斯猫 故事</td>
<td>云轩阁 盘龙</td>
<td>遮天快眼看书</td>
</tr>
<tr>
<td>4</td>
<td>波斯猫 习性</td>
<td>云轩阁小说网</td>
<td>太古遮天官网</td>
</tr>
<tr>
<td>5</td>
<td>波斯猫的眼睛</td>
<td>云轩阁txt下载</td>
<td>百度遮天官网</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>-5</td>
<td>孟买</td>
<td>德州</td>
<td>书库</td>
</tr>
<tr>
<td>-4</td>
<td>嵊</td>
<td>嘉兴</td>
<td>题材</td>
</tr>
<tr>
<td>-3</td>
<td>浓度</td>
<td>海口</td>
<td>媒体</td>
</tr>
<tr>
<td>-2</td>
<td>洛威拿</td>
<td>邢台</td>
<td>类型</td>
</tr>
<tr>
<td>-1</td>
<td>眼大</td>
<td>金华</td>
<td>唐砖</td>
</tr>
</tbody>
</table>
Subtopic Mining Framework

Random Walk on Query-URL bipartite graph

Query Recommendation

Query + Query Facets

Similar Query from Query2vec

Wikipedia Index & disambi. items

Candidate Ranking (Learning to Rank)

Hierarchy Construction

Top Down FLS->SLS

Bottom Up SLS->FLS

KB Aided Construction
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/disambig. 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

<table>
<thead>
<tr>
<th>Query</th>
<th>First Level Subtopics</th>
</tr>
</thead>
<tbody>
<tr>
<td>先知</td>
<td>虚空先知 先知电子狗 先知x 808 DOTA 先知出装 先知 电影</td>
</tr>
<tr>
<td>波斯猫</td>
<td>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</td>
</tr>
<tr>
<td>猫头鹰</td>
<td>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</td>
</tr>
<tr>
<td>Adobe</td>
<td>Adobe Reader X Adobe Lightroom Adobe Flash Player Adobe Photoshop Adobe Acrobat Professional</td>
</tr>
<tr>
<td>传奇</td>
<td>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</td>
</tr>
<tr>
<td>小米</td>
<td>小米m2 小米粥 小米红米2代 小米+官网 小米1s青春版</td>
</tr>
<tr>
<td>中国水电</td>
<td>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 中国水电建设集团 中国水电建设集团港航建设有限公司</td>
</tr>
<tr>
<td>三字经</td>
<td>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</td>
</tr>
</tbody>
</table>
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.
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

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

• Ranking Models

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

\[
R(Q,D) = W_{BM25} + \alpha_I \cdot W_{WP}
\]

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

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

• 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.

**THEOREM:** 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 documents $L$
1 $S\leftarrow \Phi$, $maxG\leftarrow 0$
2 function recursion_full_search(curD,leftD,$d_i$,curG)
3     if (leftD is $\Phi$ 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 $/$ $\{d_j\}$, $d_j$, $G_j$)
11         end function
12     foreach $d_i$ in D
13     recursion_full_search( $\{d_i\}$, $D$ $/$ $\{d_i\}$, $d_i$, $G_{ii}$)
14 return $S$
```
## Document Ranking Results

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

luochengleo@gmail.com
www.thuir.cn