Overview of The NTCIR-11 IMine Task

Yiqun Liu, Ruihua Song, Min Zhang, Zhicheng Dou, Takehiro Yamamoto, Makoto P. Kato, Hiroaki Ohshima, Ke Zhou
Background: Diversified Search

• Given an ambiguous/underspecified query, produce a single result page that satisfies different user intents!

• Challenge: balancing relevance and diversity with results from heterogeneous information sources

---

- Highly relevant near the top
- Cover many intents
- Give more space to popular intents?
- Give more space to informational intents?

小米 (Millet, Xiaomi)
Background: Diversified Search

- Possible framework for diversified search
  - Identification of ambiguous/broad/clear queries
  - Generation of subtopics for ambiguous/broad queries
  - Search result diversification for better ranking
The IMine task

- IMine (曖昧, ambiguous in Japanese) Task Goal
  - To explore and evaluate the technologies of mining and satisfying different user intents behind a Web search query

- A core task in NTCIR-11 and succeeding work of INTENT@NTCIR-9 and INTENT2@NTCIR-10 tasks

- Three subtasks
  - TaskMine (TM) subtask: to find subtasks of a given task described by a query.
  - Subtopic Mining (SM) subtask: automatically estimating different intents of a given query.
  - Document Ranking (DR) subtask: Selectively diversifying search results by balancing between relevance and diversity
Differences from Previous Tasks

• Mining and evaluating hierarchical user intents
• More subtopic candidates provided (from commercial search engine, user behavior log mining and result page analysis)
• New corpus (ClueWeb12-B13), More public user behavior data (doubled size)
  • 1.85GB => 3.85GB, over 40M user clicks
• User preference test v.s. Cranfield-like evaluation with professional assessment in diversified search evaluation
IMine Task Timetable

- Corpus available: Aug 31, 2013
- Call for participants: Aug 31, 2013
- Task participant registration Due: Jan 20, 2013
- Topics and non-diversified baseline DR runs released: Jan 21, 2014
- SM and DR submissions due: May 23, 2014
- Evaluation results available: Aug 15, 2014 (delayed 2 weeks)
- Early draft overview paper available: Aug 22, 2014 (delayed 3 weeks)
- Draft participant paper submission due: Sept 15, 2014
- Final Overview paper available: Oct 1, 2014
- Camera-ready participant paper submission due: Nov 1, 2014
Subtopic Mining Settings

• Goal: a two-level hierarchical list of subtopics for each query topic (5*10 subtopics)
Subtopic Mining Settings

• Query set

<table>
<thead>
<tr>
<th>Language</th>
<th>#Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ambiguous</td>
</tr>
<tr>
<td>English</td>
<td>16</td>
</tr>
<tr>
<td>Chinese</td>
<td>16</td>
</tr>
<tr>
<td>Japanese</td>
<td>17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>#Shared Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>14 shared topics for E/C/J (another 8 for E/C)</td>
</tr>
</tbody>
</table>

• Candidate subtopics provided
  • Query suggestions collected from Bing, Google, Sogou, Yahoo! and Baidu
  • Query dimensions generated by rule-based method (Dou et al., 2011) from search results
  • Query facets generated by keyword extraction from clicked snippets on SERPs (Liu et al., 2011)
Participants’ Techniques (SM)

• Additional candidate sources
  • Disambiguation items from Wikipedia/Baidu Baike (e.g. FRDC, THUSAM)
  • Random walk on query-result bipartite graph with user behavior logs (e.g. THUSAM)
  • Title / keyword / anchor of landing pages (e.g. KUIDL)

• Generating two-level hierarchy
  • Clustering candidates to find similar second-level subtopics (e.g. FRDC, THUSAM)
  • Extracting first-level subtopics from clusters with word embedding, semantic expansion or rule-based methods (e.g. KLE, KUIDL, hultech)
  • Web page structures are used to identify the matching of first-level and second-level subtopics (e.g. KUIDL)
Subtopic Mining Evaluation

• A new metric considering both the importance of subtopics and the quality of the subtopic hierarchy.

• A mixture of three factors:
  • **H-score**: evaluate the matching of first-level and second-level subtopics (accuracy-based)
  • **F-score**: evaluate ranking of first-level subtopics (D#-nDCG based)
  • **S-score**: evaluate ranking of second-level subtopics (D#-nDCG based)

\[
H - \text{measure} = Hscore \times (\alpha \times Fscore + \beta \times Sscore), \quad (\alpha + \beta = 1)
\]

• For ambiguous queries, \(\alpha = \beta = 0.5\)
• For broad queries, \(\alpha = 0, \beta = 1.0\)
Document Ranking Settings

• Goal: a diversified ranked list of no more than 100 results for each query topic

• Chinese corpus: SogouT (ver. 2008)
  • 130M Chinese pages
  • Organizer provided a non-diversified baseline (adopted by TUTA and THUSAM)

• English corpus: ClueWeb12-B13
  • 52M English Web pages
  • A search interface is provided by Lemur project
  • Many thanks to Prof. Jamie Callan and his team

• Evaluation: D#nDCG (weight=0.5)

\[ D\#nDCG = \lambda \cdot I - recall + (1 - \lambda) \cdot D - nDCG \]
Participants’ Techniques (DR)

• External sources adopted
  • Query logs, Wikipedia, ConceptNet and query suggestions

• Result diversification based on subtopics
  • Result combination via filling up multiple knapsacks (TUTA)
  • Result selection based on pruned exhaustive search (THUSAM)
  • Result selection based on greedy search (UM13, SEM13)
  • Result aggregation based on original ranking (udel)

• Result diversification based on novelty detection or redundancy detection
  • Result re-ranking with HITS (THUSAM)
Document Ranking Evaluation

- User preference test v.s. Cranfield-like approach
- 30 students were recruited to finish the preference test
- Each pair of results are annotated by 3 students
Result Submissions

- 10 teams submitted results
  - Universities and research institutes from Canada, China, France, Japan, Korea and U.S.

<table>
<thead>
<tr>
<th>Group</th>
<th>SM-C</th>
<th>SM-J</th>
<th>SM-E</th>
<th>DR-C</th>
<th>DR-E</th>
</tr>
</thead>
<tbody>
<tr>
<td>UDEL</td>
<td></td>
<td></td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>SEM13</td>
<td></td>
<td></td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>HULTECH</td>
<td></td>
<td></td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>THU-SAM</td>
<td>5</td>
<td></td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>FRDC</td>
<td>5</td>
<td></td>
<td></td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>TUTA1</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>CNU</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KUIDL</td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UM13</td>
<td></td>
<td></td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>KLE</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>#Group</td>
<td>5</td>
<td>2</td>
<td>8</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>#Run</td>
<td>19</td>
<td>5</td>
<td>29</td>
<td>9</td>
<td>15</td>
</tr>
</tbody>
</table>
Evaluation Results (SM)

- KLE performs best in Chinese and Japanese SM task
  - *S-score* oriented ranking (best *S-score* for SME/SMC/SMJ)
  - *H-scores* of CNU, KLE, THUSAM and FRDC are not significantly different from each other for SMC
- KLE approach in SLS extraction: semantic pattern matching in top-ranked search results (KLE oral report: 15:00@Day3, session A-2)
Evaluation Results (SM)

• KUIDL performs best in English SM task
  • *H-score* plays the most important part (KUIDL and THUSAM gain best performance with no significant difference)
  • KUIDL approach: document structure of result pages (KUIDL oral report: 14:30@Day3, session A-2)
  • Similar strategy in KUIDL and THUSAM: FLS is a substring of corresponding SLS
Evaluation Results (DR)

- Fine-grain v.s. Coarse-grain evaluation
  - Evaluation based on second-level or first-level subtopic list
- TUTA performs best for DRC and fine-grain DRE; Udel performs best for coarse-grain DRE
  - No significant differences with other top runs
  - Top performers gain more balanced results compared with previous INTENT tasks (both l-recall and D-nDCG are high)
Evaluation Results (DR)

- User preference test v.s. Cranfield-like approach

<table>
<thead>
<tr>
<th>Run A</th>
<th>Run B</th>
<th>A&gt;B</th>
<th>A=B</th>
<th>A&lt;B</th>
</tr>
</thead>
<tbody>
<tr>
<td>TUTA1-D-C-1B</td>
<td>FRDC-D-C-1A</td>
<td>53.7%</td>
<td>19.5%</td>
<td>26.8%</td>
</tr>
<tr>
<td>TUTA1-D-C-1B</td>
<td>FRDC-D-C-2A</td>
<td>48.7%</td>
<td>28.2%</td>
<td>23.1%</td>
</tr>
<tr>
<td>TUTA1-D-C-1B</td>
<td>THUSAM-D-C-1A</td>
<td>29.2%</td>
<td>22.9%</td>
<td>47.9%</td>
</tr>
<tr>
<td>TUTA1-D-C-1B</td>
<td>THUSAM-D-C-2A</td>
<td>45.8%</td>
<td>14.6%</td>
<td>39.6%</td>
</tr>
<tr>
<td>THUSAM-D-C-1A</td>
<td>FRDC-D-C-1A</td>
<td>56.1%</td>
<td>31.7%</td>
<td>12.2%</td>
</tr>
<tr>
<td>THUSAM-D-C-1A</td>
<td>FRDC-D-C-2A</td>
<td>51.3%</td>
<td>20.5%</td>
<td>28.2%</td>
</tr>
<tr>
<td>THUSAM-D-C-1A</td>
<td>THUSAM-D-C-2A</td>
<td>54.2%</td>
<td>39.6%</td>
<td>6.3%</td>
</tr>
<tr>
<td>FRDC-D-C-1A</td>
<td>FRDC-D-C-2A</td>
<td>32.4%</td>
<td>43.2%</td>
<td>24.3%</td>
</tr>
<tr>
<td>FRDC-D-C-1A</td>
<td>THUSAM-D-C-2A</td>
<td>31.7%</td>
<td>12.2%</td>
<td>56.1%</td>
</tr>
<tr>
<td>FRDC-D-C-2A</td>
<td>THUSAM-D-C-2A</td>
<td>28.2%</td>
<td>15.4%</td>
<td>56.4%</td>
</tr>
</tbody>
</table>
Evaluation Results (DR)

• User preference test v.s. Cranfield-like approach (cont.)
  • TUTA1-D-C-1B v.s. THUSAM-D-C-1A: difference is not significant (two-tailed paired t-test p-value=0.13)

• FRDC-D-C-1A/2A v.s. THUSAM-D-C-2A: FRDC systems sometimes return less than 10 results
Lessons Learned from This Round

• Less second level subtopic should be required
  • In this year, 5 first-level subtopics per query and 10 second-level subtopics per first-level subtopic is required
  • Too much assessment cost => reduced to 10-20 in practice
  • Not so reasonable in Web search scenario

• A more recent corpus should be adopted
  • One query from DRC fails to return any valid results: Android 2.3 game download
  • SogouT is crawled in 2008 (Android 2.3 didn’t exist)

• Heterogeneous information sources (e.g. verticals) should be involved
  • Vertical results are necessary; users are not familiar with SERP without verticals
Take-home messages from SM and DR

• Subtopic structure is studied in SM task
  • Two-level hierarchy of subtopics requires much annotation efforts and sometimes cause confusions
  • KLE performs the best in SMJ and SMC with highest SLS mining performance (Sscore)
  • KUIDL performs the best in SME with highest FLS-SLS matching performance (Hscore)

• User preference test results are compared with Cranfield-like approach in DR task
  • Most results are similar with each other
  • $D\#-nDCG$ may not produce credible ranking when performances are close or lengths of result lists are different

• Top results are more balanced than previous tasks
Task Mining Subtask (TaskMine)

- As a subtask of IMine -

Takehiro Yamamoto
Makoto P. Kato
Hiroaki Ohshima
(Kyoto University)

December, 2014@NII, Tokyo
Background

Information needs of searchers are sometimes Complex

Lose Weight

Do physical exercise
  - fitness center
  - swimming school

Control calories intake
  - healthy recipes
  - weight loss foods

Have diet pills
  - diet pills
  - HCG drops

Have surgery
  - lose weight surgery
  - Lap band

...
Goal of TaskMine subtask

• Understanding *the relationship among tasks* for supporting the Web searchers.

• Particularly, aims to explore the methods of automatically *finding subtasks of a given task*.

• Subtopic Mining subtask
  • Focus on the topical intent of a query
    • "I want to *find* this information!"

• TaskMine subtask
  • Focus on the task-oriented intent of a query
    • "I want to *accomplish* this task!"
Task

• Given a query (task), participants are required to return a ranked list of subtasks that help to achieve the query.

Input: **Query**

Lose Weight

Output: **List of Subtasks**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Subtask</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>physical exercise</td>
</tr>
<tr>
<td>2</td>
<td>healthy food</td>
</tr>
<tr>
<td>3</td>
<td>diet pills</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

• Documents
  • Participants are allowed to use **any resources** on the Web
Queries

- Queries
  - 50 Japanese queries

<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health</td>
<td>禁煙する (quit smoking) ストレスを解消する (relief stress)</td>
</tr>
<tr>
<td>Education</td>
<td>中国語を勉強する (learn Chinese) 九九を覚える (master 9x9 table)</td>
</tr>
<tr>
<td>Daily Life</td>
<td>ペットを預ける (leave pet) ホエールウォッチングをする (whale watching)</td>
</tr>
<tr>
<td>Sequential</td>
<td>神社でお参りをする (pray in shrine) 食パンを作る (make bread)</td>
</tr>
</tbody>
</table>
Evaluation Methodology

1. Preparing gold-standard task
   • Ask assessors to create gold standard task for each query
   • Also ask assessors to vote the importance of each task
     • How the task effectively helps to achieve the given goal?

2. Matching gold-standard task and participant task
   • For each participant task, assessors were asked to select at most one corresponding gold-standard task

3. Evaluation Metric
   • nDCG (adopted to penalize the redundant output)
Participating Teams

• **uhyg** (University of Hyogo)
  • Use Web search engines
  • Query expansion
  • Dependency parsing

• **InteractiveMediaMINE** (Kogakuin University)
  • Use *Community Q&A corpus* (Yahoo! Chiebukuro)
  • Dependency parsing

• **Organizer's baseline**
  • Use Web search engines
  • Simple syntactic pattern with tf-idf weighting
InteractiveMediaMINE significantly performed better than the other runs.

Community Q&A corpus is useful resource for task mining.
uhyg worked well on sequential category

Query modification and dependency parsing are effective to mine sequential tasks
Summary of TaskMine

• Lessons Learned
  • Community Q&A corpus is a strong resource for the task mining
  • Query expansion and dependency parsing were effective on sequential types of query
  • InteractiveMediaMINE and uhyg have oral presentations @Day3 14:05〜16:05

• Open Questions
  • How do the existing subtopic mining technique work well on TaskMine?
  • Can we aggregate heterogeneous information to find more effective tasks?
Future Plans of IMine-2

• In IMine-2, we will keep the basic task design in IMine-1, but more focus on vertical intents behind a query
  • More realistic to actual Web searches

• Query Understanding subtask
  • ≡ Subtopic Mining Subtask
  • Given a query, participants are required to identify its subtopics and their relevant verticals (web, news, image, movie, etc)

• Vertical Incorporating subtask
  • ≡ Document Ranking Subtask
  • Given a query, participants are required to return a diversified result and decide which result should be displayed with vertical results.
Thank you

http://www.thuir.org/IMine/
http://www.dl.kuis.kyoto-u.ac.jp/ntcir-11/taskmine/