DCU at the NTCIR-14 OpenLiveQ-2 Task

Piyush Arora & Gareth J.F. Jones
ADAPT Centre, School of Computing
Dublin City University, Ireland
{Piyush.Arora,Gareth.Jones}@dcu.ie

Date: 13th June 2019
Outline

- Task Overview
- Methodology
- Experiments
- Results
- Analysis
- Findings & Future Work
Task Overview

- **Task:** Rank a list of Japanese language questions matching a user’s query

- **Dataset:** Yahoo queries and respective question-answers

- **Goal:** Effectively model information from the user click logs and relevance based metrics

- **Evaluation:**
  - Offline evaluation: metrics such as NDCG, ERR
  - Online evaluation: live Yahoo question answering platform
I am wondering if I should be a smoker. I am a university student man. Wh...

My surroundings smoke cigarettes anyway. There is a smoking area in the university, but even if my friend is in the smoking area and smokes, I hate the smell of cigarettes, and I don't like the side steam smoke, so I don't get into the smoking area, everyone Wait outside until you finish smoking ...

Resolved 2016/02/24 06:16 18 Views 262

Ways of life and love, troubles in relationships > Love consultation, troubles in relationships

I do not know the smoker's feelings at all. If a human being is normal, I nee...

Smoking is a desire that does not require. Because it looks so cool, it looks so cool, so why not start it? As a result, too high money is paid, breath becomes stinking, aerobic exercise ability is also lost, and unnecessary image down is also caused, and smokers are unconditional ...

Resolved 2016/11/14 02 21 Views 401

Manners, ceremonial occasions > manners > smoking manners

Original Japanese page translated using the Google translation
Challenges

- Queries are typically short and ambiguous in nature and might not capture the user’s intention effectively.
- For example, for a Japanese query: “喫煙”, English translation: “smoking”, can have multiple intentions:
  - “dangers of smoking”
  - “smoking health effects”
  - “mechanism to quit smoking”
- Without understanding the user’s intent and focus of the query, it becomes challenging to re-rank the questions.
- **Aim:** Model textual based information and click logs based information to re-rank questions effectively.
Learning To Rank Problem

Training Data

$q_1$

$x_1^{(1)}$

$x_2^{(1)}$

$\vdots$

$x_m^{(1)}$

$y^{(1)}$

$q_2$

$x_1^{(2)}$

$x_2^{(2)}$

$\vdots$

$x_m^{(2)}$

$y^{(2)}$

$q_n$

$x_1^{(n)}$

$x_2^{(n)}$

$\vdots$

$x_m^{(n)}$

$y^{(n)}$

Learning System

Test Data

$q$

$x_1$

$x_2$

$\vdots$

$x_m$

$\vdots$

Prediction

$h(x)\hat{y}$

Resources and Tools

● Resources provided by the task organizers:
  ○ Pipeline for processing Japanese text
  ○ Pipeline for features extraction
  ○ Data set and click logs

● Used Lemur RankLib toolkit

● Total of 77 features
Content based features

<table>
<thead>
<tr>
<th>Features</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>tf_sum</td>
<td>tf_in_idf_sum &amp;</td>
</tr>
<tr>
<td>log_tf_sum</td>
<td>bm25</td>
</tr>
<tr>
<td>norm_tf_sum</td>
<td>log_bm25</td>
</tr>
<tr>
<td>log_norm_tf_sum</td>
<td>lm_dir</td>
</tr>
<tr>
<td>idf_sum</td>
<td>lm_jm</td>
</tr>
<tr>
<td>log_idf_sum</td>
<td>lm_abs</td>
</tr>
<tr>
<td>icf_sum</td>
<td>dlen</td>
</tr>
<tr>
<td>log_tfidf_sum</td>
<td>log_dlen</td>
</tr>
<tr>
<td>tfidf_sum</td>
<td></td>
</tr>
</tbody>
</table>

Question Title

Question Body

Snippet

Body Answer
Click log based features

<table>
<thead>
<tr>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>answer_num</td>
</tr>
<tr>
<td>log_answer_num</td>
</tr>
<tr>
<td>view_num</td>
</tr>
<tr>
<td>log_view_num</td>
</tr>
<tr>
<td>is_open</td>
</tr>
<tr>
<td>is_vote</td>
</tr>
<tr>
<td>is_solved</td>
</tr>
<tr>
<td>rank</td>
</tr>
<tr>
<td>updated_at</td>
</tr>
</tbody>
</table>

User Logs
Methodology

- Learning to Rank (L2R) algorithms:
  - Coordinate Ascent
  - MART

- Feature Selection & Combination:
  - Alternative combinations of the 5 feature set

- Parameters optimization

- Scores Normalisation:
  - Z-score normalization
  - Score average
  - Max based normalization
## Dataset

<table>
<thead>
<tr>
<th></th>
<th>Training set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Queries</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>Number of Questions</td>
<td>986,125</td>
<td>985,691</td>
</tr>
<tr>
<td>Number of click logs</td>
<td>288,502</td>
<td>148,388</td>
</tr>
</tbody>
</table>
Our Submissions

- Total of 14 systems submitted
- Overall 65 participant submissions
- All 65 submissions evaluated & ranked using
  - NDCG@10, ERR@10, Q measure
  - Phase-1 online evaluation
- Top 30 systems selected for final online evaluation
- 5 of our systems selected in top 30 systems
Best Models

Top 5 Systems

<table>
<thead>
<tr>
<th>System-ID</th>
<th>NDCG@10</th>
<th>ERR@10</th>
<th>Q-Measure</th>
<th>Cumulative Gain-1</th>
<th>Cumulative Gain-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>106</td>
<td>0.85</td>
<td>0.70</td>
<td>0.75</td>
<td>0.80</td>
<td>0.75</td>
</tr>
<tr>
<td>112</td>
<td>0.75</td>
<td>0.60</td>
<td>0.65</td>
<td>0.70</td>
<td>0.65</td>
</tr>
<tr>
<td>118</td>
<td>0.65</td>
<td>0.50</td>
<td>0.55</td>
<td>0.60</td>
<td>0.55</td>
</tr>
<tr>
<td>126</td>
<td>0.55</td>
<td>0.40</td>
<td>0.45</td>
<td>0.50</td>
<td>0.45</td>
</tr>
<tr>
<td>147</td>
<td>0.45</td>
<td>0.30</td>
<td>0.35</td>
<td>0.40</td>
<td>0.35</td>
</tr>
</tbody>
</table>

Best Scores: 106, 112, 118, 126, 147
Average Scores: 0.70, 0.60, 0.55, 0.50, 0.45
## Systems Ranking

<table>
<thead>
<tr>
<th>Systems</th>
<th>ID</th>
<th>NDCG@10</th>
<th>ERR@10</th>
<th>Q-Measure</th>
<th>Online Evaluation Phase-1</th>
<th>Final Online Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>System-2</td>
<td>106</td>
<td>32</td>
<td>24</td>
<td>26</td>
<td>7</td>
<td>7**</td>
</tr>
<tr>
<td>System-4</td>
<td>112</td>
<td>36</td>
<td>35</td>
<td>64</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>System-5</td>
<td>118</td>
<td>45</td>
<td>38</td>
<td>65</td>
<td>4</td>
<td>6**</td>
</tr>
<tr>
<td>System-7</td>
<td>126</td>
<td>34</td>
<td>34</td>
<td>32</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>System-12</td>
<td>147</td>
<td>21</td>
<td>23</td>
<td>20</td>
<td>29</td>
<td>23</td>
</tr>
</tbody>
</table>

** No significant differences between the top scored runs using Tukey’s HSD tests
Analysis

- Coordinate Ascent algorithm performs relatively better than the Mart algorithm
- Our best system (ID-130) based on NDCG@10 and ERR@10 was ranked “2” and “3” respectively
- Based on Q-scores our best system (ID-123) was ranked “6”
- Based on the cumulative credit our best system (ID-118) was ranked “4” and “6” for online phase-1 and final phase evaluation
- Most of our submissions were heavily tuned to focus on relevance-based features (for e.g BM25 and LM scores)
Findings & Future Work

- Ranking of systems based on the online evaluation metric differed from that for the offline evaluation metrics.
- Need for more research to understand the factors behind contrary ranking results arising from the use of online and offline evaluation metrics.
- Our best systems in the online phase focused on modelling users click logs.
- **Future work:** explore more effective techniques for the exploitation of user logs and click distributions for ranking questions.
Task Overview
- Challenges: Rank a list of questions matching a user’s query, for Japanese language
- Goal: Effectively model information from the user click logs and relevance-based metrics
- Evaluation: Offline and Online evaluation

Main Challenges
- Questions are typically short and ambiguous in nature and might not capture the user’s intention effectively.
- For example, for Japanese query, “喫煙”, English translation: “smoking”
  - Possible Query Intention-1: "dangers of smoking"
  - Possible Query Intention-2: "mechanism to quit smoking"
- Complex problem to re-rank the questions without understanding the user’s intent and focus of the query.
- Aim: How to model the aspects of textual relevance and information gained through used click data to re-rank and present the information effectively to a user.

Systems Submission & Results
- Submitted 13 systems
- Out of 15 total submissions, 4 systems were selected in top 30 systems to be evaluated in the final phase.

Analysis
- Coordinate Ascent algorithm performs relatively better than the Mert algorithm.
- Our best system (ID: 130) based on NDCG@10 and MAP@10 was ranked 22nd and 13th respectively.
- Based on Q-scores our best system (ID: 123) was ranked 6th.
- Based on the cumulative score our best system (ID: 119) was ranked 4th and 1st for offline phase 1 and the final phase evaluation.
- Most of our submissions were heavily tuned to focus on the relevance-based features such as BMS and EM scores, measuring the similarity of queries with a set of questions to be re-ranked.

Findings & Future Work
- Ranking of systems based on the online evaluation metric correlated to the offline evaluation metric.
- Need for more research and focus to understand the main factors behind the contrary results ranking using offline and online evaluation metrics.
- Our best systems in the offline phase focused on modeling user click logs, thus in the future we would like to explore more effective approaches for modeling user click logs and click distributions for ranking.
- Need for further investigation to find online and offline evaluation metrics that correlate well in order to address the task of ranking questions.

Acknowledgement: We would like to thank the Task Organizers of NTCIR-14 OpenLiveQ-2 and YuShinMoriyama from the ADAPT center.
Acknowledgement:

- NTCIR’14 Organizers
- Task Organizers of NTCIR’14 OpenLiveQ-2
- Yasufumi Moriya from the ADAPT centre