Engaging Content Engaging People

# DCU at the NTCIR-14 OpenLiveQ-2 Task

### **Piyush Arora and Gareth J. F. Jones**

ADAPT Centre, School of Computing, Dublin City University, Dublin 9, Ireland {piyush.arora,gareth.jones}@dcu.ie

### **Task Overview**

• **Challenge:** Rank a list of questions matching a user's query, for

### Dataset

• **Dataset:** Yahoo Queries and respective Question-Answers

- Japanese language
- **Goal:** Effectively model information from the user click logs and relevance based metrics
- **Evaluation:** Offline and Online evaluation

| Q 喫煙 × Search official   | corner Q Question and          |
|--|--------------------------------|
| + Condition specification  | consultation                   |
| Q Smoking 's passive smoking smoking seat smoking office searched  | Search target                  |
| I am wondering if I should be a smoker . I am a university student man. Wh   | All (198, 221)                 |
| My surroundings smoke cigarettes anyway. There is a smoking area in the university, but even if m  | Answering received (752)       |
| y friend is in the <b>smoking</b> area and smokes, I hate the smell of cigarettes, and I don't like the sidestr  | Voting Accepted (109)          |
| eam smoke, so I don't get into the <b>smoking</b> area, everyone Wait outside until you finish smoking   | Solved (197,360 cases)         |
| Ways of life and love, troubles in relationships > Love consultation, troubles in relationships  |                                |
|  | order of display               |
| I do not know the smoker 's feelings at all. If a human being is normal, I nee   | former and the second second   |
| Smoking is a desire that does not require . Because it looks so cool, it looks so cool, so why not sta   | Relevancy order                |
| rt it? As a result, too high money is paid, breath becomes stinking, aerobic exercise ability is also los<br>t, and unnecessary image down is also caused, and smokers are unconditional |                                |
| Resolved  2016/11/14 02  21  21  21  21  21  21  21  21  21  2   | Notice                         |
| Manners, ceremonial occasions > manners > smoking manners  | Wiedom bag soarch DSS function |

Original Japanese page translated using the Google translation

### Main Challenges

- Queries 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" Ο • 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 user click data, to retrieve and present the information effectively to a user

|                      | Training set | Test set |
|----------------------|--------------|----------|
| Number of queries    | 1,000        | 1,000    |
| Number of questions  | 986,125      | 985,691  |
| Number of click logs | 288,502      | 148,388  |

### Methodology

- Learning to Rank algorithms: Explored L2R algorithms including **Coordinate Ascent and MART**
- Feature Selection & Combination: Explored alternative combinations of diverse feature sets capturing relevance of the user query and retrieved ranked list of questions

| Type of Features                                | Features Range |
|---|----------------|
| Title Based Textual Features (Title set)        | [F1-F17]       |
| Snippet Based Textual Features (Snippet set)    | [F18-F34]      |
| Question Body Based Textual Features (Body set) | [F35-F51]      |
| Body Answer Based Textual Features (Answer set) | [F52-F68]      |
| Click Log Features (Click set)                  | [F69-F77]      |

#### More detail on the features is provided in the paper

• **Parameter selection:** Varied L2R model parameters to learn effective hypothesis functions from the dataset

- **Scores Normalisation:** The scale of the features (77 features) varies considerably, some features are on logarithmic scales (log-based values), so we performed three scores normalization functions:
  - Z-score normalization
  - Score average
  - Max based normalization

## **Systems Submission & Results**

- Total of 14 systems submitted
- 5 of our systems were selected in top 30 systems to be evaluated in the final phase, out of a total of 65 participant submissions

### **Overall Results**





### 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, such as BM25 and LM scores, measuring the similarity of queries with the set of questions to be re-ranked

### **Findings & Future Work**

• Ranking of systems based on the online evaluation metric differed from that for the offline evaluation metrics

|  | Average<br>Scores | Best | 152 | 150 | 147 | 143 | 133 | 130 | 128 | 126 | 123 | 118 | 112 | 110 | 106 | 99 |  |
|--|-------------------|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|----|--|
|--|-------------------|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|----|--|

System Submissions

| ID  | NDCG@10 | ERR@10 | Q-Measure | Credit- Phase-1 | Credit- Phase-2 |
|-----|---------|--------|-----------|-----------------|-----------------|
| 106 | 32      | 24     | 26        | 7               | 7               |
| 112 | 36      | 35     | 64        | 8               | 10              |
| 118 | 45      | 38     | 65        | 4               | 6               |
| 126 | 34      | 34     | 32        | 14              | 12              |
| 147 | 21      | 23     | 20        | 29              | 23              |

Our top systems' ranking based on different 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. Thus in future work we would like to explore more effective techniques for the exploitation of user logs and click distributions for ranking questions
- 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 Yasufumi Moriya from the ADAPT centre



The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund. This research is supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre for Digital Content Technology"