RUCIR at NTCIR-14 WWW Task

Speaker: Jiaqing Liu

School of Information
Renmin University of China

Author: Xue Yang and Zhicheng Dou
Outline

• WWW @ NTCIR-14
• Overview
• Model
• Results and analysis
• Conclusion
WWW @ NTCIR-14

• Goal: An ad hoc web search task
  • Ranking Web pages with their relevance
• Subtask 1: Chinese
  • SogouT-16 Corpus, SogouQCL Corpus
• Subtask 2: English
  • ClueWeb12-B13 Corpus
Data-Flow Overview

• Four steps

Data Preprocessing

Training Data

Feature Extraction

Model

Test Data

Predict Result

Feature vectors
Data Preprocessing

• Pre-processing web corpus: cleaning, parsing and indexing using Solr

• Collecting official and previous TREC and NTCIR Competition labeled data for training models.

• We do not use user behavior data
Feature Extraction

• Traditional Features
  • Traditional relevance features for different fields

• Embedding Features
  • Cosine similarity between the distributed representations of query and document

• Deep Neural Features
  • Matching scores of unlabeled query-document pair by deep neural matching models
Feature Extraction (Cont’d)

• Traditional Features
  • Relevance features for four fields
    • Anchor, title, URL, and body
  • Relevance features for the whole document

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Fields</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25</td>
<td>BM25 with default parameters</td>
<td>(anchor), title, URL, body, whole</td>
</tr>
<tr>
<td>TF-IDF</td>
<td>TF-IDF model</td>
<td>(anchor), title, URL, body, whole</td>
</tr>
<tr>
<td>LMIR</td>
<td>Language model with Dirichlet smoothing</td>
<td>(anchor), title, URL, body, whole</td>
</tr>
<tr>
<td>TF</td>
<td>Sum of term frequency</td>
<td>(anchor), title, URL, body, whole</td>
</tr>
<tr>
<td>IDF</td>
<td>Sum of inverse document frequency</td>
<td>(anchor), title, URL, body, whole</td>
</tr>
<tr>
<td>DL</td>
<td>Document length</td>
<td>(anchor), title, URL, body, whole</td>
</tr>
<tr>
<td>PM</td>
<td>Perfect match</td>
<td>(anchor), title, URL, body, whole</td>
</tr>
<tr>
<td>CM</td>
<td>Complete match</td>
<td>(anchor), title, URL, body, whole</td>
</tr>
</tbody>
</table>
Feature Extraction (Cont’d)

• Embedding Features
  
  • Word2Vec (Mikolov et al., 2013)
  
  • Get representations of query and document by averaging the word embedding of terms
    
    • \( V_{di} = \frac{1}{n} \sum_{j=1}^{n} Term_{ji}, j \in [1 \ldots n] \)
  
  • Cosine similarity between query representation and document representation as feature
  
  • Basic use of pre-train word embedding
Feature Extraction (Cont’d)

• Deep Neural Features (Matching Score)
  • ARC-I (Hu et al., 2014)

• Learn representation vectors of query and document with CNNs
• Get the matching score by a multi-layer perceptron layer (MLP)
Feature Extraction (Cont’d)

• Deep Neural Features (Matching Score)
  • ARC-II (Hu et al., 2014)

- Learn interaction representation vectors for query and document
- Get the matching score by a MLP after 2D pooling and convolution
Feature Extraction (Cont’d)

• Deep Neural Features (Matching Score)
  • $DRMM$ (Guo et al., 2016)

Matching histograms: interaction between query term with document
Matching score: based on MLP and calculated by a softmax function
Feature Extraction (Cont’d)

• Deep Neural Features (Matching Score)
  • aNMM (Yang et al., 2016)

  - Use value-shared weighting rather than position-shared (ARC-II)
  - Integrate the results of each query term with a softmax function
Feature Extraction (Cont’d)

- Deep Neural Features (Matching Score)
  - *MV-LSTM* (Wan et al., 2016)

- Learn representation of query and document by bi-LSTMs
- Build interaction matrix with cosine and get score by MLP
Feature Extraction (Cont’d)

• Deep Neural Features (Matching Score)
  • *DUET* (Mitra et al., 2017)

• Local representations: one-hot encoding to exact term match
• Distributed representations: latent embedding based topic model
Model Training

• Input Format

![Input Format Diagram]

• Model
  • Ranklib: LambdaMART
Evaluation Metrics

• nDCG@K
  
  \[ nDCG@K = N_K^{-1} \sum_{i=1}^{n} g(r_i) d(i) \]

• Q@K
  
  \[ Q@K = \frac{1}{\min(K,R)} \sum_{r=1}^{K} J(r) \frac{c(r) + \beta cg(r)}{r + \beta cg^*(r)} \]

• nERR@K
  
  \[ nERR@K = \sum_{r=1}^{K} \frac{1}{r} \prod_{i=1}^{r-1} (1 - R_i) R_r \]
### Results (Chinese)

<table>
<thead>
<tr>
<th>Run</th>
<th>Query</th>
<th>Features</th>
<th>nDCG@10</th>
<th>Q@10</th>
<th>nERR@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>RUCIR-1</td>
<td>Description</td>
<td>Traditional Embedding</td>
<td>0.4515</td>
<td>0.4228</td>
<td>0.5792</td>
</tr>
<tr>
<td>RUCIR-2</td>
<td>Content</td>
<td>Traditional Embedding</td>
<td>0.4866</td>
<td>0.4571</td>
<td>0.6044</td>
</tr>
<tr>
<td>RUCIR-3</td>
<td>Description</td>
<td>Traditional</td>
<td>0.4503</td>
<td>0.4223</td>
<td>0.5630</td>
</tr>
<tr>
<td>RUCIR-4</td>
<td>Description</td>
<td>Traditional Deep Neural</td>
<td>0.4458</td>
<td>0.4226</td>
<td>0.5619</td>
</tr>
<tr>
<td>RUCIR-5</td>
<td>Description</td>
<td>Deep Neural</td>
<td>0.2745</td>
<td>0.2404</td>
<td>0.3832</td>
</tr>
</tbody>
</table>
## Results (English)

<table>
<thead>
<tr>
<th>Run</th>
<th>Query</th>
<th>Features</th>
<th>nDCG@10</th>
<th>Q@10</th>
<th>nERR@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>RUCIR-1</td>
<td>Description</td>
<td>Traditional</td>
<td>0.3137</td>
<td>0.2973</td>
<td>0.4469</td>
</tr>
<tr>
<td>RUCIR-2</td>
<td>Content</td>
<td>Traditional</td>
<td>0.3489</td>
<td>0.3352</td>
<td>0.4917</td>
</tr>
<tr>
<td>RUCIR-3</td>
<td>Description</td>
<td>Traditional</td>
<td>0.3137</td>
<td>0.2973</td>
<td>0.4469</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Embedding</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RUCIR-4</td>
<td>Description</td>
<td>Deep Neural</td>
<td>0.3293</td>
<td>0.3094</td>
<td>0.4602</td>
</tr>
<tr>
<td>RUCIR-5</td>
<td>Description</td>
<td>Deep Neural</td>
<td>0.2876</td>
<td>0.2659</td>
<td>0.4188</td>
</tr>
</tbody>
</table>
Analysis

• [CN & EN] Query content Run > Query description Run (CO > DE)

• [CN] Traditional features Run + Embedding features Run > Other Runs (1 > 3, 4, 5)

• [EN] Traditional features Run + Deep neural features Run > Other Runs (4 > 1, 3, 5)

• [CN & EN] Deep neural features Run << Other Runs (5 << 1, 2, 3, 4)
Conclusion

• We Want Web task
  • Matching with query content is better than matching with query description
  • Traditional text relevance features are still stable and effective
  • Using embedding feature can help a little
  • Using deep neural features can help but less than expectation, which needs future research
Thanks

Speaker: Jiaqing Liu
Author: Xue Yang and Zhicheng Dou
Email: {ruc_yangx, dou}@ruc.edu.cn