HITSZ-ICRC at NTCIR-11 Temporalia Task

Temporal Information Retrieval

Yongshuai Hou, Cong Tan, Jun Xu, Youcheng Pan, Qingcai Chen and Xiaolong Wang

Intelligent Computing Research Center
Key Laboratory of Network Oriented Intelligent Computation
Harbin Institute of Technology Shenzhen Graduate School

2014/12/11
Outline

• Introduction
• Temporal Information Retrieval
  – Candidate document retrieval
  – Temporal relevant judging
  – Search subtopic classification
  – Relevant document re-ranking
• Results Evaluation
• Summary
Introduction

• Who are we?
  – HITSZ-ICRC group: we are from Intelligent Computing Research Center (ICRC), Harbin Institute of Technology Shenzhen Graduate School (HITSZ), Shenzhen, China.

• Which task/subtask did we participate?
  – Temporal Query Intent Classification (TQIC)
  – Temporal Information Retrieval (TIR)
Introduction

• Temporal Information Retrieval (TIR)
  – For a given topic, participants retrieve relevant documents for different kinds of temporal subtopics
  – A given search topic contains title, description, query date, and 4 different class subtopics including class information (atemporal, past, recency, future)
**Title** | **Girl with the Dragon Tattoo**
--- | ---
**Description** | I've recently watched a film called Girl with the Dragon Tattoo, and really liked it. Therefore, I would like to gather information about the movie.

**Past question** | How did the casting of the film develop?

**Recency question** | What did the recent reviews say about the film?

**Future question** | Is there any plan about its sequel?

**Atemporal question** | What are the names of main actors and actresses of the film?

**Search date** | 28 Feb 2013 GMT+0:00
Document Collection

- Supplied “LivingKnowledge news and blogs annotated sub-collection” corpus, contains 3.8M documents from blogs and news sources.

- Available information of a document in Corpus
  - Document create time
  - Named entity tags in content
  - Time expression tags in content
  - Normalized value for time expression
  - ...

2014/12/11
New Oscar rules: Can the Academy curtail awards season excess?

The Academy of Motion Picture Arts and Sciences is set to announce the new Oscar rules—rules that are aimed at curbing the over-the-top campaigning that has become amedia norm. The Academy's official statement reads:

"The Academy has always been committed to maintaining the integrity of our awards process. With the introduction of these new rules, we aim to ensure that the focus remains on the films and performances, rather than on the spectacle of the awards season. These changes will help to preserve the Academy's reputation for fairness and impartiality, and to maintain the integrity of our awards for years to come."
TIR Process

1. **User Query**
2. **Document Retrieve**
3. **Document Index**
4. **Temporal Classify**
5. **Candidate document set**
6. **Document Re-Rank**
   - Weight Sum
   - Learning to Rank
7. **Relevant Document Re-Rank List**
Candidate Document Retrieval

• Create an index on the document collection and retrieve for each subtopic using Lucene toolkit with BM25 language model
• Save top 500 retrieval results in result list as candidate documents for each subtopic
• Save BM25 score for each candidate document as content relevant score $CR$
Temporal Relevant Judging

• Judge temporal relevant between subtopic and candidate document
  – Calculate the date distance between search date and each time expression in document
  – Classify each time expression into class *past, recency or future*

\[
d_{i} = Dq - DX_i
\]

\[
C_i = \begin{cases} 
\text{future}, & \text{if } d_{i} < 0 \\
\text{past}, & \text{if } d_{i} > B_p \\
\text{recency}, & \text{if } 0 \leq d_{i} \leq B_r
\end{cases}
\]

Where \( Dq \) is search date of the topic, \( DX_i \) is normalized time expression in document, \( B_p \) is the classification boundary for *past* class time expression, \( B_r \) is the classification boundary for *recency* class time expression. \( B_p=300 \text{ (days)} \) here.

– Judge temporal relevant according to whether the document contains same class time expressions as subtopic class
– If have, the temporal relevant score \( TR \) for the document is 1, otherwise, the score \( TR \) is 0
Document Re-Ranking

• Relevant score weight sum
  – Sum content relevant score and temporal relevant score as final relevant score for candidate document ranking

\[ R = \alpha R_c + (1 - \alpha) R_t \]

Where \( R \) is the document final relevant score to the search subtopic, \( R_c \) is the content relevant score, \( R_t \) is the temporal relevant score, \( \alpha \) is the weight coefficient and \( \alpha \geq 0, \alpha \leq 1 \).

– Different class use different coefficient value to calculate relevant score

<table>
<thead>
<tr>
<th>Subtopic Class</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>past</td>
<td>0.85</td>
</tr>
<tr>
<td>recency</td>
<td>0.73</td>
</tr>
<tr>
<td>future</td>
<td>0.76</td>
</tr>
<tr>
<td>atemporal</td>
<td>1</td>
</tr>
</tbody>
</table>
Document Re-Ranking

– Coefficient choosing

nDCG and $\alpha$ relation for different subtopic class

- past
- recency
- future

nDCG@20

$\alpha$ value
Document Re-Ranking

– Deficiency of relevant score weight sum method
  • Difficult to get the best coefficient for each subtopic
  • Subtopic class information is necessary

• Learning to rank
  – Features
    • content relevant feature
    • Temporal relevant feature
Document Re-Ranking

• Learning to rank
  – Feature List for learning to rank
    ➢ similarity between search topic and document title
    ➢ similarity between search topic and document content
    ➢ similarity between search subtopic and document title
    ➢ similarity between search subtopic and document content
    ➢ BM25 relevant score between search topic and document
    ➢ BM25 relevant score between search subtopic and document
    ➢ temporal relevant
Document Re-Ranking

• Learning to rank
  – Training data: 15 search topic each with 4 subtopics and relevant documents in qrels file in task dry run step
  – Algorithm: LambdaMART algorithm in RankLib toolkit
  – Model
    • Train independent model for each subtopic class
  – Deficiency: Lack of training data
Subtopic classification

- Use method in TQIC subtask to classify search subtopic class
Subtopic classification

- Classification accuracy of method in TQIC subtask is poor
  - Accuracy on TIR subtopic classifying is about 80%

<table>
<thead>
<tr>
<th>runID</th>
<th>atemporal</th>
<th>future</th>
<th>past</th>
<th>recency</th>
</tr>
</thead>
<tbody>
<tr>
<td>PrW</td>
<td>70.67%</td>
<td>64.00%</td>
<td>78.67%</td>
<td>62.67%</td>
</tr>
<tr>
<td>PrWsQW</td>
<td>69.33%</td>
<td>66.67%</td>
<td>77.33%</td>
<td>57.33%</td>
</tr>
<tr>
<td>qRPrHNB</td>
<td>57.33%</td>
<td>68.00%</td>
<td>81.33%</td>
<td>61.33%</td>
</tr>
</tbody>
</table>

- Training one rank model for all the 4 subtopic classes to avoid the subtopic classification step
Results Evaluation

• We submit 3 runs: Run BW (run1), Run BWCC (run2), Run LTRNC2 (run3)

• Run description
  – All 3 runs used search title, subtopic domain as input information
  – run1: used the relevant score weight sum method, and used the original class information of search subtopic.
  – run2: used the relevant score weight sum method, used the our classifiers in TQIC to get subtopic class, and did not use the original class information of search subtopic.
  – run3: used 1 model for 4 subtopic classes learning to rank method (LambdaMART algorithm here), and did not use the original class information
Formal Results Evaluation

- *Run run3* is better than *run1* and *run2*
- No significant difference between *run1* and *run2*
  - Use time factor can improve re-rank result, no matter which class it is.
  - Same temporal feature for each class is not much suitable

Results evaluation of TIR subtask runs

<table>
<thead>
<tr>
<th>runID</th>
<th>nDCG @20</th>
<th>AP @20</th>
<th>P@20</th>
<th>nERR @20</th>
</tr>
</thead>
<tbody>
<tr>
<td>BW (run1)</td>
<td>0.4544</td>
<td>0.4587</td>
<td>0.5895</td>
<td>0.6056</td>
</tr>
<tr>
<td>BWCC (run2)</td>
<td>0.4554</td>
<td>0.4599</td>
<td>0.5902</td>
<td>0.6064</td>
</tr>
<tr>
<td>LTRNC2 (run3)</td>
<td>0.4768</td>
<td>0.483</td>
<td>0.6018</td>
<td>0.6313</td>
</tr>
</tbody>
</table>

nDCG@20 of each class in TIR formal runs

<table>
<thead>
<tr>
<th>runID</th>
<th>atemporal</th>
<th>future</th>
<th>past</th>
<th>recency</th>
</tr>
</thead>
<tbody>
<tr>
<td>BW (run1)</td>
<td>0.4669</td>
<td>0.4607</td>
<td>0.4005</td>
<td>0.4897</td>
</tr>
<tr>
<td>BWCC (run2)</td>
<td>0.4678</td>
<td>0.4593</td>
<td>0.403</td>
<td>0.4915</td>
</tr>
<tr>
<td>LTRNC2 (run3)</td>
<td>0.5092</td>
<td>0.4804</td>
<td>0.4227</td>
<td>0.495</td>
</tr>
</tbody>
</table>
Results Evaluation

• nDCG values of different subtopics are very different, 4 subtopics in one topic are also different
Results Evaluation

nDCG value and relevant document number in *qrels* file for *recency* class

![Graph showing nDCG and RelNum values](image)
Results Evaluation

• The more number of relevant documents in answer file, the higher nDCG value for most of subtopics
  – The number of relevant document for subtopic 036r, 015r, 050r and 042r is 141, 173, 68 and 24, nDCG value for those subtopics is 0.9159, 0.8887, 0.0772 and 0
  – Fewer relevant documents in data set means smaller chance to get relevant documents for the top 20 results
Summary

• Use date distance as temporal feature
• Tried relevant score weight sum method
• Tried learning to rank method
• Future work
  – Improve temporal feature extraction. Eg. using float data between 0 and 1 to indicate the temporal relevant instead of 0, 1 only
  – Using the NE tags in document content
References

Thanks for your Attention

Q&A

Welcome offline discussion by sending emails to houyongshuai@hitsz.edu.cn