CS5244 Digital Libraries

IR in FAQ System

Raymond Jun ZHENG    HT050659Y
Zheng LU             HT055495R
Research Background

• QA System:
  - Find the most similar question-answer pairs with respect to user’s queries.
  - Rule-based, statistical, and mixed approaches.

• FAQ System
  - Retrieving information from a set of semi structured texts
  - Designed for the retrieval of the very frequent, popular, and highly reusable question-answer pairs, called QA pairs
  - QA pairs are usually provided or verified by domain experts
  - Domain-specific and adopts inference and reasoning to retrieve a more accurate QA pair for a query.

• Traditional information retrieval does not use semantic representation and knowledge
Literature Review and Objective

- There are three prominent parts of the FAQ system: Query Processing Techniques, Knowledge Base Structure and FAQ Retrieval Techniques.

- **FAQ Retrieval Techniques**
  - Statistical similarity approach with keyword match,
  - Statistical similarity approach with prioritized keyword match,
  - Statistical similarity approach with case based reasoning,
  - Statistical similarity approach with vector model,
  - Semantic similarity approach and
  - Database query

- Objective of this study:
  - Discuss and compare the FAQ system answer retrieval techniques based on **statistical similarity approach** and **semantic similarity approach**.
Use of the Statistical Similarity Approach with Vector Model

- **VSM similarity measurement**

    \[
    \text{Sim}(q, d) = \cos(q \cdot d) \\
    = \frac{q \cdot d}{\|d\| \|q\|} \\
    = \frac{\sum_{k \in (q \cap d)} w_{kd} \cdot w_{kq}}{\|d\| \|q\|} \\
    = \frac{\sum_{k \in (q \cap d)} w_{kd} \cdot w_{kq}}{\sqrt{\sum_{k \in d} (w_{kd})^2} \sqrt{\sum_{k \in q} (w_{kq})^2}}
    \]

- **Performance of VSM (Baseline)**

- **The shortfall of VSM Similarity Measure**
  - Documents with similar content but different vocabularies may result in a poor inner product. This is a limitation of keyword-driven IR systems.
Use of Improved Statistical Similarity Approach with Vector Model with Stop-words Removal and Stemming

**Evaluation**
- NN → stop-words removal feature *Not available*
- stemming feature *Not available*
- YN → stop-words removal feature *available*
- stemming feature *Not available*
- NY → stop-words removal feature *Not available*
- stemming feature *available*
- YY → stop-words removal feature *available*
- stemming feature *available*

**Result**
- *stop-words removal does not help*
- *Stemming helps*

<table>
<thead>
<tr>
<th></th>
<th>Mean Reciprocal Rank</th>
<th>MRR improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>0.452885</td>
<td></td>
</tr>
<tr>
<td>NY</td>
<td>0.481566</td>
<td>6.3%</td>
</tr>
<tr>
<td>YN</td>
<td>0.442715</td>
<td>-2.2%</td>
</tr>
<tr>
<td>YY</td>
<td>0.504562</td>
<td>11.4%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14.0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.8%</td>
</tr>
</tbody>
</table>
Use of the Semantic Similarity Approach

- The implementation of the semantic similarity approach
  - Method
    1. Category specific keywords
    2. Documents specific keywords
  - Formula
    \[ \text{Score} = P_1 \cdot M_1 + P_2 \cdot M_2 + M_{vsm} \]

Findings: Best if \( P_1 = 4 \) and \( P_2 = 20 \)
Comparison between the improved model and the baseline model

- **Result**

  ![Graph showing comparison between baseline and improved models](image)

<table>
<thead>
<tr>
<th></th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.504562</td>
</tr>
<tr>
<td>Improved</td>
<td>0.59211</td>
</tr>
<tr>
<td>Improved with Query</td>
<td>17.4%</td>
</tr>
</tbody>
</table>

- **Further Improvement Using Query Expansion**

  ![Graph showing further improvement with query expansion](image)
Conclusion

- Mere Statistical Similarity Approach is not enough.
- Use of the Stemming Feature helps.
- Semantic Similarity Approach with addition of category keywords and sentence keywords help.
- Semantic Similarity Approach with addition of query expansion does not help with regard to the performance.