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

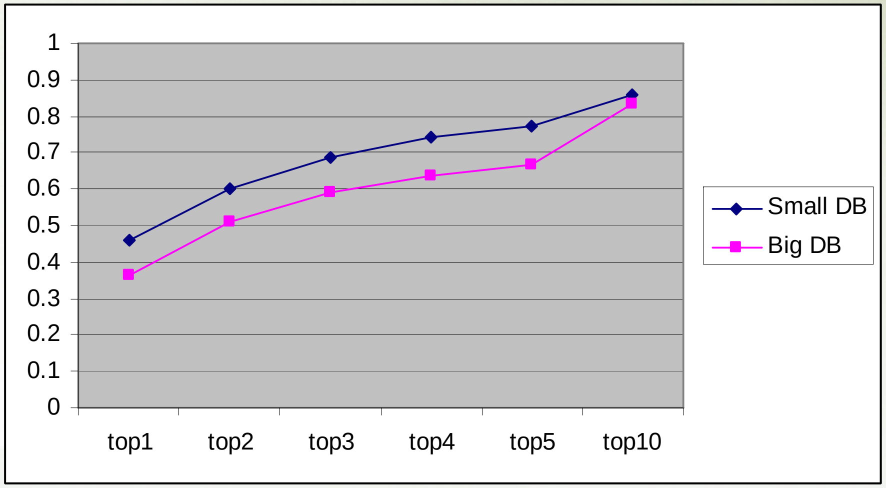
- Objective of this study:
  - Discuss and compare the FAQ system answer retrieval techniques based on *statistical similarity approach* and *semantic similarity approach*.

Use of the Statistical Similarity Approach with Vector Model

- VSM similarity measurement

$$\begin{aligned} Sim(\mathbf{q},\mathbf{d}) &= \cos(\mathbf{q}\angle\mathbf{d}) \\ &= \frac{\mathbf{q}\cdot\mathbf{d}}{\|\mathbf{d}\|\ \|\mathbf{q}\|} \\ &= \frac{\sum\limits_{k\in(q\cap d)} w_{kd}\cdot w_{kq}}{\|\mathbf{d}\|\ \|\mathbf{q}\|} \\ &= \frac{\sum\limits_{k\in(q\cap d)} w_{kd}\cdot w_{kq}}{\sqrt{\sum\limits_{k\in d}(w_{kd})^2}\sqrt{\sum\limits_{k\in q}(w_{kq})^2}} \end{aligned}$$

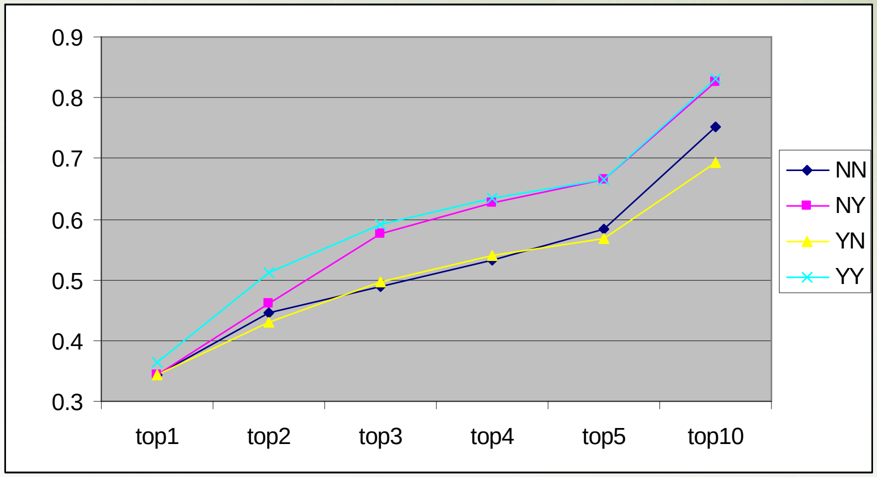
Performance of VSM (Baseline)



- 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*

	Mean Reciprocal Rank
NN	0.452885
NY	0.481566
YN	0.442715
YY	0.504562

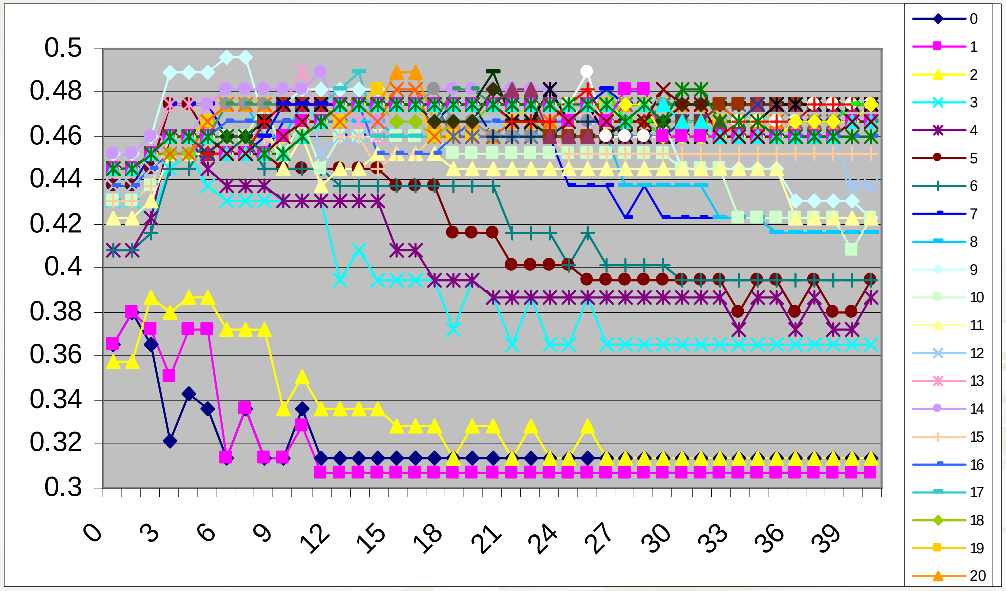
	MRR improvement
NN -> NY	6.3%
NN -> YN	-2.2%
NN -> YY	11.4%
YN -> YY	14.0%
NY -> YY	4.8%

Use of the Semantic Similarity Approach

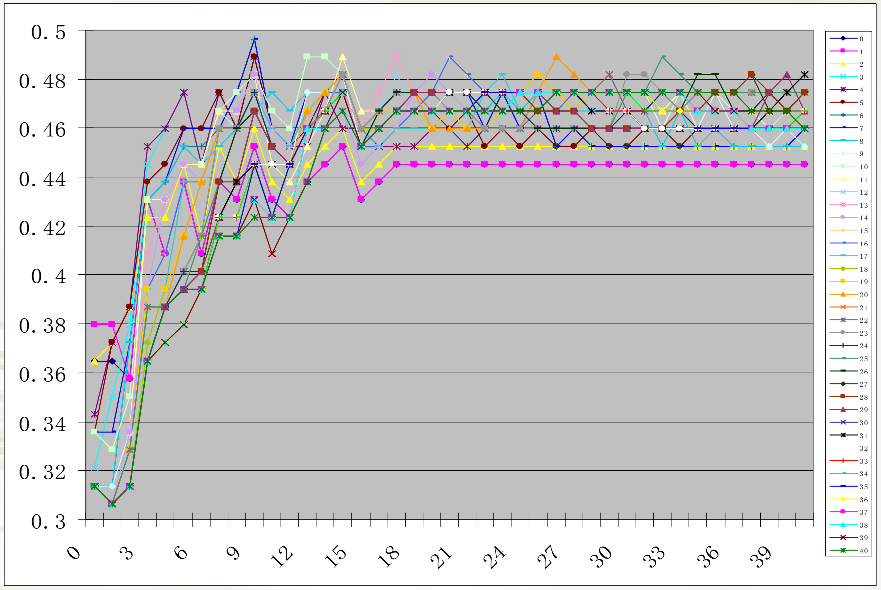
- The implementation of the semantic similarity approach
  - Method
    1. Category specific keywords
    2. documents specific keywords
  - Formula

    Score = $P_1 * M_1 + P_2 * M_2 + M_{vsm}$

Performance with respect to P_1



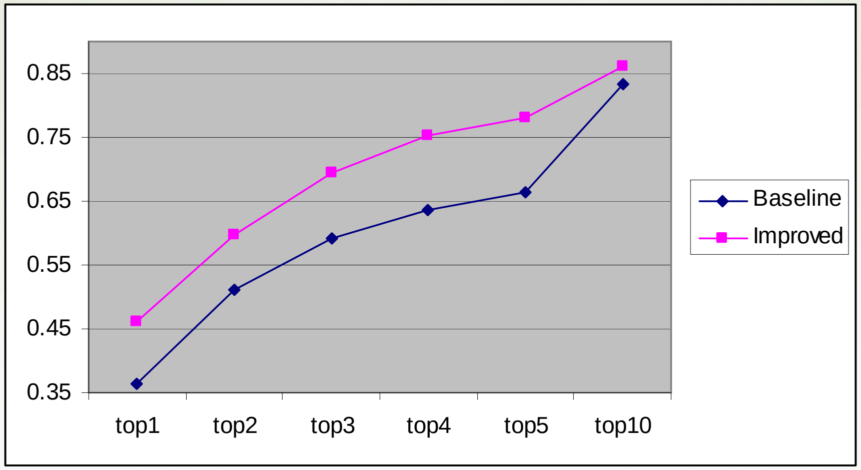
Performance with respect to P_2



Findings: Best if P_1 = 4 and P_2 = 20

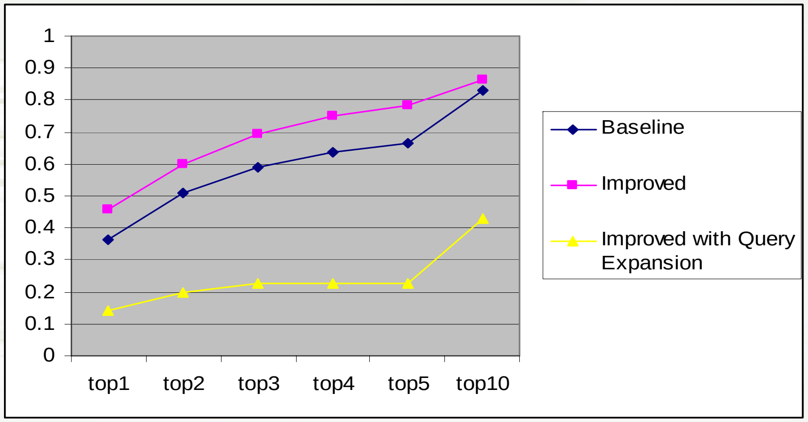
Comparison between the improved model and the baseline model

- *Result*



	MRR
Baseline	0.504562
Improved	0.59211
⇧	**17.4%**

- *Further Improvement Using Query Expansion*



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.