OCSE – zbkn Team

Optimized Course Search Engine
Overview

- Project description
  - Objective
- Project methodology
- Key findings / results
  - Okapi BM25
  - PLN
  - Okapi BM25 Axiomatic & Comparison
- Future Work
- Conclusion
- Questions
Project Description

OCSE Objectives

1. Expand content and not just MOOC
2. Improve relevance
   a. Scoring functions
   b. Feedback and judgments
3. Document Summarization
Procedure / Methodology

General IR Framework

Collection

Tokenizer

Inverted Index

Query

Tokenizer

Scorer | Ranker

Feedback

Results

Ranked List of Documents

Offline Process

Online

Evaluation
Procedure / Methodology – cont’d

General IR Framework

Collection
Tokenizer
Inverted Index
Query
Tokenizer
Scorer | Ranker
Feedback
Results
Ranked List of Documents

Custom Crawler and Cleaned Corpus

Query expansion with Rocchio Feedback

Okapi BM25 Axiomatic

Added document summarization

Offline Process
Online
Evaluation
Procedure / Methodology – objective 1

Expand content and not just MOOC

- Crawling – Focused Crawling
  - phantomjs – for non authentication pages Stanford, City University of NY
  - Javascript – Authenticated pages like UIUC, used BFS with 3-level queuing

- Cleaned up - Corpus
  - Affects Document Length Normalization and term frequency.
  - Analyzed each domain, trimmed & picked relevant content
  - Built custom solutions using Python & Java
Procedure / Methodology – objective 2.a

**Improve relevance** - Scoring functions

- Vector Space Model
  - Okapi BM25
  - Pivot Length Normalization

- Okapi BM25 Axiomatic
  - Equation:
    \[
    S(Q,D) = \sum_{t \in D \cap Q} \frac{C_t^Q \times \text{weight}(t)}{\frac{2}{avdl} + |D| + s + C_t^D}
    \]

- Best results MAP 0.563820635
- Significance testing on System A, B (different parameters)
  - p-value: 0.014771603
Procedure / Methodology – objective 2.b

Improve relevance – Feedback and judgments

- Rocchio Feedback
  - Equation: 
    \[
    Q_m = (a \cdot Q_o) + \left( b \cdot \frac{1}{|D_r|} \sum_{\overrightarrow{D_j} \in D_r} \overrightarrow{D_j} \right) - \left( c \cdot \frac{1}{|D_{nr}|} \sum_{\overrightarrow{D_k} \in D_{nr}} \overrightarrow{D_k} \right)
    \]
  - Query Expansion
    - Before: “literacy in children”
    - After: “literacy children language learn education”

- Average Precision – Query “literacy in children”
  - Before: 0.133
  - After: 0.20
Procedure / Methodology – objective 3

Document Summarization

- Existing results are top k lines or matched sentences from the document
- Single sentence summary for user
- Summarization done offline
- Integrated results with
  - Key: url, value: summary

Example:

**default result:** ****Fantastic**** ****Art****, Mail ****Art****, Black and White Portraiture Photography, Installation ****Art**** ...

**summarized sentence:** Convey a personal appreciation for art concepts, techniques, and approaches through the creation and sharing of your own original artwork.
Experiment Results - Okapi BM25
Experiment Results - Pivoted Length Normalization (PLN)
Experiment Results

Okapi BM25 Axiomatic

Comparison
<table>
<thead>
<tr>
<th>k</th>
<th>0.1</th>
<th>0.5</th>
<th>0.3</th>
<th>0.1955</th>
<th>score</th>
<th>0.5184</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>0.1</td>
<td>0.5</td>
<td>0.3</td>
<td>0.0563</td>
<td>0.0595</td>
<td>0.05</td>
</tr>
<tr>
<td>q</td>
<td>0.0000000</td>
<td>0.0000000</td>
<td>0.0000000</td>
<td>0.0000000</td>
<td>0.0000000</td>
<td>0.0000000</td>
</tr>
<tr>
<td>f</td>
<td>0.0000000</td>
<td>0.0000000</td>
<td>0.0000000</td>
<td>0.0000000</td>
<td>0.0000000</td>
<td>0.0000000</td>
</tr>
</tbody>
</table>

**Experiments Data – Okapi BM25 Axiomatic**

- **k** and **r** vary to adjust the weighting in the Okapi BM25 formula. 
- **q** and **f** represent the query and document frequencies, respectively. 

Additional variables include:
- **score**: The calculated score of the query-document relevance. 
- **nat**: Natural log of the document frequency. 
- **len**: Document length. 
- **dF**: Document frequency of the term. 
- **idf**: Inverse document frequency. 
- **df**: Number of documents containing the term. 
- **tf**: Term frequency. 

Each variable plays a crucial role in the calculation of document relevance within the Okapi BM25 framework.
Future Work

- Full Rocchio Feedback & Summarization Integration
- Model Reviews & Class Responses from ex: Piazza
- Sentiment Analysis
Conclusion

- Course Search Engine with broad set of data
- Better Results through
  - Okapi BM25 Axiomatic Scoring
  - Rocchio Feedback
  - Document Summarization
- Better tool for teachers and students
Questions?