QIC: Incorporating Context into a User Query

Min Song and Lori Watrous-deVersterre
Information Systems
College of Computing Sciences
New Jersey Institute of Technology
Outline

• Research Problems
• Research Goals
• QIC Overview
  – QEIQIC
  – Concept Extraction
  – Learning to Rank and Dynamic Clustering
• Evaluation
• Conclusions and Future Work
Research Problems

People want search results to reflect exactly what they meant, all that they meant, and only what they meant, and they want it quickly.

- There are gaps….
  - Gap between what the user wants (information need) and the query that the user formulates
  - Gap between what the document represents and indexes that the IR engine built
Research Goals

• The purpose of the project is three-fold:

  1. Incorporating inference of user preferences in Query Expansion
     
     Our approach: QEQIC

  2. Capturing meanings embedded in documents
     
     Our approach: Concept Extraction

  3. Ranking search results with context-enriched features
     
     Our approach: Learning to Rank and Dynamic Clustering
Data Collection

• Two different types of data:
  
  1. **Ohsumed** – biomedical data collection for proof-of-concept at the initial development phase.
  2. **Ensemble Pathway** – computing sciences data collections at [http://www.computingportal.org/collections](http://www.computingportal.org/collections)

  **jOAI**, an OAI harvesting tool built in Ncore was used to crawl Ensemble.

  Note: The size of data is small. This may influence the overall performance of our approach.
Query Expansion: QEQIC

- Query is initially represented as a tuple of \{context, named entities\}.

  - **Named entities** detected using Boosted Dictionary-based Entity Spotter (BDES).
  - A **concept tuple** consists of \{Computing concept, description, class\}.
    - *Class* assigned to a concept manually based on ACM Classification
Boosted Dictionary-based Entity Spotter

• Dictionary-based approach: tackles the problem of lack of contextual cues but:
  – too many false recognitions
  – takes too long to look up the dictionary entry.

• Our approach resolves these issues by:
  – Approximate String Distance Algorithm to retrieve candidate entries
  – Shortest-path Distance Algorithm
  – Part-Of-Speech (POS) tag
  – Syntactical properties of terms
Demonstrate the algorithm for simultaneously finding the minimum and maximum values in an array.

- **Concept:**
  - *Algorithm*

- **Class:**
  - *Theory of Computation*

- **Description:**
  - *Model of computation and algorithm*
QEQIC: static user profile data

- Incorporate Static User Preference data into query expansion
  - Subject terms stored in user profile are matched with titles of data set in the Latent Semantic Index (LSI) space.
  - $N$ top terms relevant to subject terms in the user profile are compared with a query.
    - If there is a good match (based on string similarity between a top term and query terms), the terms are weighted higher.
QEQIC: Semantic Query Expansion Algorithm

**Offline Indexing Phase**
- Build Concept Tuple Index
- Build Context Term Index
- Build Static User Context Datastore
- Build a Dictionary for NER

**Real-time Searching Phase**
- Parse a Query by Boolean Operators
  - AND, OR, NOT
- Identify Left/Right Context Terms
- Extract Named Entities from the Query
- Use a string distance algorithm
- Select the Best Concept Tuple that matches with the Named Entities
  - BDES NER
- Match context terms in the query with Context Index
- Select the relevant terms to context terms by Random Project (LSI)
- Retrieve relevant terms from the title field to Static User Context data
- Select top N matched from the retrieved terms with the query
- Combine expanded query terms

**QEQIC Data Flow**
Preliminary Results of BDES

<table>
<thead>
<tr>
<th>BDES</th>
<th>Genia</th>
<th>Geinia+MeSH</th>
<th>Genia+MeSH+UMLS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Precision</strong></td>
<td>0.93</td>
<td>0.878</td>
<td>0.88</td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td>0.573</td>
<td>0.72</td>
<td>0.68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BDES without POS/syntactic properties</th>
<th>Genia</th>
<th>Geinia+MeSH</th>
<th>Genia+MeSH+UMLS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Precision</strong></td>
<td>0.76</td>
<td>0.56</td>
<td>0.51</td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td>0.62</td>
<td>0.82</td>
<td>0.78</td>
</tr>
</tbody>
</table>
Concept Extraction

• Probabilistic Combinatorial Markov Random Fields (PCMRF):
  – A supervised learning technique
  – PCMRF is a non-generative graph model.

• Training data:
  – 5000 sentences from Ensemble Pathway and other computing sciences related digital libraries.
    • These 5000 sentences are **positive** examples (meaning containing concepts in the sentence).
    • Combined with 5000 more sentences (**negative** examples), we build a concept extraction model.
Concept Extraction

- RESTful Web Services for Concept Extraction


<tagging>
  <token>
    <name>Testing</name>
    <class>O</class>
  </token>
  <token>
    <name>internet</name>
    <class>Web</class>
  </token>
  <token>
    <name>tagging</name>
    <class>O</class>
  </token>
  <token>
    <name>service</name>
    <class>O</class>
  </token>
</tagging>
Dynamic Clustering of Search Results

• Clustering approach:
  – Based on a supervised learning technique - Probabilistic Combinatorial Markov Random Fields (PCMRF) technique
    • Same as our concept extraction technique
  – Requires a small set of initial training examples.
  – For performance reasons, input for clustering is a set of concepts extracted from Ensemble and stored in a database.
Rank search results with context features

• **Learning to rank** – apply supervised learning techniques to rank search results.

• **Proposed technique**: **Mixture Support Vector Machines**
  – Combines multiple models
    • Models are built with a set of features (attributes) such as TF-IDF, no. of clicks, the user’s research interest, etc.
    • There are several different ways to select features:
      1) Document-driven model [11,15],
      2) Meta data-driven model [14],
      3) User static context-driven model, and
      4) User search behavior-driven model [13]

Note: The current model is based on document-related features.
The most popular approach in learning to rank.

- Training data is part of the LETOR package \[11\]
  \url{http://research.microsoft.com/en-us/um/beijing/projects/letor/default.aspx}
- 25 features were extracted
  - 10 from title, 10 from abstract, and 5 from ‘title + abstract’
  - TF, TF*IDF, BM25, Language Model ranking scores, IDF, etc

For query id “1” and document id “40626”, the label is “2” (definitely relevant).
Search Behavior-driven Model

- Incorporate users’ search behavior into ranking the results [13].

<table>
<thead>
<tr>
<th>qid</th>
<th>category</th>
<th>search</th>
<th>abstract_text</th>
<th>full_text</th>
<th>no_visits</th>
<th>no_returned_citation</th>
<th>pos_clicked_citation</th>
<th>search_duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2:1:0</td>
<td>3:0:0</td>
<td>4:1</td>
<td>5:0</td>
<td>6:0</td>
<td>7:0.0169444444</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2:2:0</td>
<td>3:0:0</td>
<td>4:2</td>
<td>5:0</td>
<td>6:1</td>
<td>7:4.2805557</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1:1:0</td>
<td>2:4:0</td>
<td>3:0:0</td>
<td>4:5</td>
<td>5:11583</td>
<td>6:10</td>
<td>7:106.29944</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1:8:0</td>
<td>2:8:0</td>
<td>3:11:0</td>
<td>4:27</td>
<td>5:2194</td>
<td>6:33</td>
<td>7:314.75027</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0:0</td>
<td>2:1:0</td>
<td>3:0:0</td>
<td>4:1</td>
<td>5:0</td>
<td>6:0</td>
<td>7:17.611666</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0:0</td>
<td>2:1:0</td>
<td>3:0:0</td>
<td>4:1</td>
<td>5:0</td>
<td>6:0</td>
<td>7:13.123055</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1:2:0</td>
<td>2:0:0</td>
<td>3:0:0</td>
<td>4:2</td>
<td>5:0</td>
<td>6:0</td>
<td>7:20.195278</td>
</tr>
</tbody>
</table>

Sample training data based on search behavior
Evaluation

- Preliminary Results with Ohsumed Data for Query Expansion
  - A set of 348,566 references from MEDLINE consisting of titles and/or abstracts from 270 medical journals over a five-year period (1987-1991).
  - Popular data set to apply supervised learning techniques to IR
  - Contains the 106 queries in test set, with patient and topic information, in the format:
    - .I  Sequential identifier
    - .B  Patient information
    - .W  Information request
  - For the preliminary test, we used 12 out of 106 queries.
Evaluation

• The Approach
  – Use Recall and Interpolated Average Precision to measure the performance.
  – Investigate whether QEQIC performs better than the baseline Language Model technique.
  – Investigate whether adding concepts, semantic types, and context terms to QE improves the performance.
Preliminary Results

<table>
<thead>
<tr>
<th></th>
<th>QE QIC (title only)</th>
<th>baseline LM (title only)</th>
<th>QE QIC (title+abstract)</th>
<th>baseline LM (title+abstract)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter. Avg. Precision</td>
<td>0.135</td>
<td>0.108</td>
<td>0.172</td>
<td>0.139</td>
</tr>
<tr>
<td>Avg. Recall</td>
<td>0.359</td>
<td>0.256</td>
<td>0.407</td>
<td>0.323</td>
</tr>
</tbody>
</table>

Measure by Recall and Interpolated Average Precision
Preliminary Results

- Impact with different feature sets

<table>
<thead>
<tr>
<th></th>
<th>QEIQC+CON</th>
<th>QEIQC+CON+SEM</th>
<th>QEIQC+CON+SEM+CXT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. Precision</td>
<td>0.137</td>
<td>0.107</td>
<td>0.107</td>
</tr>
<tr>
<td>Avg. Recall</td>
<td>0.359</td>
<td>0.321</td>
<td>0.321</td>
</tr>
</tbody>
</table>

CON: Concept
SEM: Semantic Type
CXT: Context Term

Measure by Recall and Interpolated Average Precision
Conclusions and Future Work

• Conclusions
  – We developed a semantic query expansion technique, and tested it on a biomedical data collections.
  – We developed a new ranking technique for the search results with the “Learning to Rank” approach.
  – We developed a concept extraction technique and a dynamic clustering technique with Probabilistic Combinatorial Markov Random Fields.
  – We developed RESTful APIs for our techniques.

• Future Work
  – We plan to conduct a pilot study and the main experiment on Ensemble Pathway data
Acknowledgement


References


References


References

Questions?

Thanks!