Overview of the NTCIR-5 WEB Query Term Expansion Subtask

Masaharu Yoshioka
Hokkaido University
National Institute of Informatics
Background

- **Query term expansion**
  - Difficulties for selecting appropriate query terms to represent the information need
  - Estimation of useful query terms for filling the gap between the given query terms and the information need

- **Evaluation of query term expansion technique**
  - Usage of test collection
    - Most of query term expansion techniques improve retrieval results in general

However, the effectiveness of this technique depends on the quality of query terms in the initial query
Background (continue)

- Analysis on Topic difficulties
  - Analysis on NTCIR-1 test collection (Eguchi, 2002)
    - Correlation between topic difficulties and average of each initial query term’s IDF
  - Clarity measure based on a language model (Cronen-Townsend et al., 2002)
    - Identifying ambiguous or ill-formed queries
    - Usage of this measure to decide whether the initial query terms requires query term expansion or not

However, these approaches do not deal with the gap between initial query terms and information need directly
Objectives

- Proposal of an evaluation framework of query term expansion technique
  - Analysis on the several features that affect the quality of the techniques
    - Focusing on the variation of initial query term types.
      - Mismatch between the initial query and relevant documents

- New approach for evaluation of the techniques
  - From evaluation in average to evaluation of strong and weak topic types
Why Query Term Expansion Works (Buckley, 2004)

- (Buckley, 2004) hypothesized a possible reason why query expansion improves the query performance.
  - one or two good alternative words to original query terms (synonyms)
  - one or two good related words
  - a large number of related words that establish that some aspect of the topic is present (context)
  - specific examples of general query terms
  - better weighting to original query terms

- First 4 reasons are related to the query term expansion technique

Question: All topics requires each type of expansion terms in same ratio?  

\[\text{No}\]
Mismatch between the Query Terms and Relevant Documents

- Existence of relevant documents that do not contain a part of initial query terms
  - In NTCIR-4 Web test collection survey type
    - Almost 40% relevant documents do not satisfy title query (2-3 query terms with simple Boolean expression)
    - Many queries are not precise enough to narrow relevant document candidates

- Quality of initial query term sets affect on the quality of the query expansion technique

  Evaluation of each query term sets based on this mismatch
Analysis on Mismatch between the Initial Query and Relevant Documents

- Analysis based on Boolean IR model
  - Comparison between documents that satisfy a initial Boolean query and relevant documents

Boolean query satisfied documents  Relevant documents

(1)  (2)  (3)

Ideal query: Both (1) and (3) are empty set
Large (1) = Ambiguous : needs context terms
Large (3) = Too strict : needs alternative terms
An Example of the Mismatch

- NTCIR-4 Web test collection
  - Title field with Boolean expression for 35 Survey type topics
  - Index of ABRIR (organizer reference system) is used for this calculation

\[
(2)/\{(1)+(2)\}
\]

Appropriateness of initial query quality from the viewpoint of precision

\[
(2)/\{(2)+(3)\}
\]

Appropriateness of initial query quality from the viewpoint of recall

Boolean Relevant Documents

Ideal Query
Feature Quantities for Evaluating Effectiveness of the Query Term

- Feature quantities for both initial query terms and query expansion terms
  - Appropriateness of the term that characterizes the relevant documents

- Feature quantities for query expansion terms
  - Appropriateness of the alternative term for each initial query term
  - Appropriateness of the context definition term for the query
Appropriateness of the term that characterizes the relevant documents

- Comparison between the characteristic terms of relevant documents and query terms
  - Mutual information content between relevant documents $r$ and the term $w$.
    \[
    G'(w) = p(w | r) \log_2 \frac{p(w | r)}{p(w)}
    \]
    - The characteristic terms are terms with higher $G'(w)$.
      - A initial query term with higher $G'(w)$ means good initial query term.
      - A query expansion term with higher $G'(w)$ means good query term for defining context.
Feature quantities for query expansion terms

- A good alternative term
  - should exist for relevant documents that do not contain the initial query term.
  - Therefore, the number of documents that have a query expansion term and do not have an initial query term is useful for evaluation.

- A good term for context definition
  - is a distinct term that exists in relevant documents.
  - Therefore, the number of documents that have a query expansion term in the relevant documents, the Boolean satisfied documents, and total documents is useful for evaluation.
Retrieval Experiments

- Test collection
  - NTCIR-4 Web test collection survey type topics: 35 topics
  - Title field (2-3 query terms with simple Boolean expression)
  - Additional relevance judgment in almost same way in NTCIR-4
    • There were several submission results whose top-ranked documents are not included in the judged document list

- Type of retrieval experiments
  - Number of query expansion terms
    • No query term expansion
    • query term expansion
      – Limited number (10)
      – No limitation
  - Feedback Type
    • Pseudo-relevant documents
    • User selected relevant documents
      – Each participant uses relevant document for simulating document selection
List of Participants

- **JSWEB**
  - Experimented with relevant document vectors that were generated based on the existence of the keyword in the relevant documents. They also proposed combining relevant document vectors (one from the user selected relevant documents and the other from the pseudo-relevant documents). The retrieval method was based on a vector space IR model.

- **NCSSI**
  - Experimented with a clustering technique for the initial retrieval results and a named entity recognition technique for selecting query expansion terms from the appropriate cluster (user selected cluster or pseudo-relevant cluster). They used an organizer reference model, ABRIR, based on a probabilistic model as an IR system.

- **R2D2**
  - Experimented with Robertson's Selection Value (RSV) for selecting query expansion terms using pseudo relevant documents. The retrieval method was based on the modified Okapi. They also used link information for scoring the retrieved documents.

- **ZKN**
  - Experimented with Larvenko's relevance model for selecting query expansion terms using pseudo relevant documents. The retrieval method was based on the inference network and language model.

- **ABRIR: Organizer Reference System**
  - Experimented with mutual information between terms and relevant documents for selecting query expansion terms. The retrieval method was based on the Okapi.
Overall Evaluation Results of the Experiment

- Most of the cases, the retrieval results improve in average.
- However, there is no system that improves all topics.
Topics where query term expansion works well

- Comparison between 10 query term expansion and no expansion
  - Count number of runs where expansion results is better than no expansion one
    - Average precision (AP)
    - R-Precision (RR)
    - Relevance Retrieved (RR)

<table>
<thead>
<tr>
<th>Topic</th>
<th>AP</th>
<th>RR</th>
<th>RR</th>
<th>AP</th>
<th>RR</th>
<th>RR</th>
<th>AP</th>
<th>RR</th>
<th>RR</th>
<th>AP</th>
<th>RR</th>
<th>RR</th>
</tr>
</thead>
</table>
Topics where query term expansion works well

- No direct correlation between mismatch and effectiveness of the query term expansion

![Diagram showing topics where many and small number of performance measures improve]
Effect of Feedback Documents Quality

- Comparison between automatic feedback and user feedback
  - Existence of topics whose retrieval results for user feedback is lower than automatic one
    - Run-Id: 0021, 0058, 0076, 0082
    - Those topics have higher R&B/B values compared with other topics
    - Good query expansion terms for these topics are terms that can be used as alternative terms for the initial query terms.
  - This result shows that non-relevant documents may be useful for finding alternative terms for initial query terms.
Discussion

- Topic types characterized by mismatch between the initial query and relevant documents may affect the performance of the query term expansion technique.
  - Parameter tuning for query term expansion technique may affect the mixed ratio of topic types.

For further analysis on the query term expansion, it is necessary to pay attention to this issue.
Summary

- Proposal a new evaluation framework for query expansion technique
  - Topic type classification based on the mismatch between initial query term set and relevant documents
  - Enumeration of feature quantities that may affect the performance of query term expansion technique
- Further analysis is necessary for evaluating this framework
Errata

Data for zkn with no query term expansion has a problem. Please replace following tables in the proceedings and “Summary of evaluation results” in a CD-ROM data.
Table 1 in the proceedings
Summary of Evaluation Results (Page 5) in CD-ROM

<table>
<thead>
<tr>
<th>Description</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
<th>Value 5</th>
<th>Value 6</th>
<th>Value 7</th>
<th>Value 8</th>
<th>Value 9</th>
<th>Value 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Column 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Column 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Column 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Column 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Column 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Column 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Column 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Column 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Column 9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Column 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3 in the proceedings
Summary of Evaluation Results (Page 2) in CD-ROM

<table>
<thead>
<tr>
<th>Column 1</th>
<th>Column 2</th>
<th>Column 3</th>
<th>Column 4</th>
<th>Column 5</th>
<th>Column 6</th>
<th>Column 7</th>
<th>Column 8</th>
<th>Column 9</th>
<th>Column 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data 1</td>
<td>Data 2</td>
<td>Data 3</td>
<td>Data 4</td>
<td>Data 5</td>
<td>Data 6</td>
<td>Data 7</td>
<td>Data 8</td>
<td>Data 9</td>
<td>Data 10</td>
</tr>
<tr>
<td>Data 11</td>
<td>Data 12</td>
<td>Data 13</td>
<td>Data 14</td>
<td>Data 15</td>
<td>Data 16</td>
<td>Data 17</td>
<td>Data 18</td>
<td>Data 19</td>
<td>Data 20</td>
</tr>
<tr>
<td>Data 21</td>
<td>Data 22</td>
<td>Data 23</td>
<td>Data 24</td>
<td>Data 25</td>
<td>Data 26</td>
<td>Data 27</td>
<td>Data 28</td>
<td>Data 29</td>
<td>Data 30</td>
</tr>
<tr>
<td>Data 31</td>
<td>Data 32</td>
<td>Data 33</td>
<td>Data 34</td>
<td>Data 35</td>
<td>Data 36</td>
<td>Data 37</td>
<td>Data 38</td>
<td>Data 39</td>
<td>Data 40</td>
</tr>
<tr>
<td>Data 41</td>
<td>Data 42</td>
<td>Data 43</td>
<td>Data 44</td>
<td>Data 45</td>
<td>Data 46</td>
<td>Data 47</td>
<td>Data 48</td>
<td>Data 49</td>
<td>Data 50</td>
</tr>
<tr>
<td>Data 51</td>
<td>Data 52</td>
<td>Data 53</td>
<td>Data 54</td>
<td>Data 55</td>
<td>Data 56</td>
<td>Data 57</td>
<td>Data 58</td>
<td>Data 59</td>
<td>Data 60</td>
</tr>
<tr>
<td>Data 61</td>
<td>Data 62</td>
<td>Data 63</td>
<td>Data 64</td>
<td>Data 65</td>
<td>Data 66</td>
<td>Data 67</td>
<td>Data 68</td>
<td>Data 69</td>
<td>Data 70</td>
</tr>
<tr>
<td>Data 71</td>
<td>Data 72</td>
<td>Data 73</td>
<td>Data 74</td>
<td>Data 75</td>
<td>Data 76</td>
<td>Data 77</td>
<td>Data 78</td>
<td>Data 79</td>
<td>Data 80</td>
</tr>
<tr>
<td>Data 81</td>
<td>Data 82</td>
<td>Data 83</td>
<td>Data 84</td>
<td>Data 85</td>
<td>Data 86</td>
<td>Data 87</td>
<td>Data 88</td>
<td>Data 89</td>
<td>Data 90</td>
</tr>
<tr>
<td>Data 91</td>
<td>Data 92</td>
<td>Data 93</td>
<td>Data 94</td>
<td>Data 95</td>
<td>Data 96</td>
<td>Data 97</td>
<td>Data 98</td>
<td>Data 99</td>
<td>Data 100</td>
</tr>
<tr>
<td>Data 101</td>
<td>Data 102</td>
<td>Data 103</td>
<td>Data 104</td>
<td>Data 105</td>
<td>Data 106</td>
<td>Data 107</td>
<td>Data 108</td>
<td>Data 109</td>
<td>Data 110</td>
</tr>
</tbody>
</table>

Note: The table above shows the summary of evaluation results as per the proceedings. Each column represents different aspects or criteria of the evaluation.