Exploiting Anchor Text for the Navigational Web Retrieval at NTCIR-5

Atsushi Fujii (Univ of Tsukuba)
Katunobu Itou (Nagoya Univ)
Tomoyosi Akiba (TUT)
Tetsuya Ishikawa (Univ of Tsukuba)

WEB-7
Introduction

Taxonomy of queries on the Web

• Navigational
  – A user has a Web site in mind and requires to reach that site

• Informational
  – A user searches for Web sites that provide knowledge for his/her information need

Focus of today’s talk

In NTCIR-5 Navi-2 Subtask, a hypothetical user requires to find the representative pages of an item (e.g., person and product)
Contribution of our research

• Exploiting anchor text in Navigational Web Retrieval
  – Modeling anchor text
  – Extracting synonyms from anchor text for query expansion

• Combining different information types
  – anchor text, page content, link structure
System overview

• We use different information types to compute the score (RSV) of document $d$ with respect to query $q$

• Anchor text: $P(d \mid q)$

• Page content: Okapi BM25
  – Indexing with word and bi-word

• Link structure: PageRank $P(d)$
  – Probability that a user surfing on the Web reaches document $d$
Combining different scores

• Scores computed by different information types have different meanings
• The **final** score of $d$ is determined by a weighted harmonic mean of the ranks

$$\text{Final score of } d = \frac{1}{\alpha \times \frac{1}{3} + \beta \times \frac{1}{1} + \gamma \times \frac{1}{4}}$$
Reality is …

• The best performance was obtained with

\[ \alpha = 0.8 \quad \beta = 0.2 \quad \gamma = 0 \]

Anchor Content Structure

• Anchor text was definitely effective for Navigational Web Retrieval

• In the remaining of this talk, we focus only on exploiting anchor text
Exploiting anchor text

• Modeling anchor text
• Extracting synonyms from anchor text for query expansion
Modeling anchor text: Basis

- $P(d \mid q)$: probability that document $d$ is the representative page for the item expressed by query $q$

\[
\arg\max_d P(d \mid q) = \arg\max_d P(q \mid d) \times P(d)
\]

Computation of $P(q \mid d)$ is crucial
Computation of $P(q \mid d)$

- We assume independence of terms in $q$

$$P(q \mid d) = \prod_{t \in q} P(t \mid d)$$

- ChaSen is used to extract term $t$

- If $P(t \mid d)$ is not modeled, we use $P(t)$ for smoothing

- We compare two alternative methods to model $P(t \mid d)$
First method: Document model
• Use all anchor texts linking to \( d \) as a single surrogate document for \( d \) [Westerveld et al., TREC 2001]

\[
P(t \mid d) = \frac{\text{Freq}(t, SD)}{\sum_t \text{Freq}(t, SD)}
\]

\( a_1 \) \hspace{1cm} \( a_2 \) \hspace{1cm} \ldots \hspace{1cm} \( a_N \)
Problem of DM

• $P(t \mid d)$ is same for ヤフー (yafuu) and “Japan”
• But, “Japan” is useless without “Yahoo”

surrogate document SD

Yahoo    ヤフー    Yahoo Japan

www.yahoo.co.jp
Problem of DM (cont.)

• Document model is spammable
  – A user can change $P(t \mid d)$ purposefully
• Computation of $P(t \mid d)$ is affected by $a_2$

surrogate document $SD$
Second method: Anchor model

• Use anchor texts linking to \( d \) independently to compute \( P(t \mid d) \)

\[
P(t \mid d) = \sum_a P(t \mid a) \times P(a \mid d)
\]

Probability of term \( t \) is normalized on an anchor-by-anchor basis
Exploiting anchor text

• Modeling anchor text
• Extracting synonyms from anchor text for query expansion
Query expansion in $P(t \mid d)$

- If $P(t \mid d)$ is not modeled, synonym term $s$ is used

$$P(t \mid d) = P(t \mid s, d) \times P(s \mid d) \approx P(t \mid s) \times P(s \mid d)$$

Probability that $s$ is replaced with $t$

- We need synonym pairs to compute $P(t \mid s)$
Extracting synonym pairs

- Multiple anchor texts linking to the same document have same/similar meaning
- We extract phonetic equivalents (transliterations) from those anchor texts

"google" and "グーグル" are phonetic equivalents
Identifying phonetic equivalents

• Our transliteration method is used to determine whether two words are phonetic equivalents

[Fujii and Ishikawa, CHUM 2001]

google / グーグル (guuguru) → Yes

google / エンジン (engine) → No
Example of query expansion

<table>
<thead>
<tr>
<th>Topic ID</th>
<th>Source term</th>
<th>Expanded term</th>
</tr>
</thead>
<tbody>
<tr>
<td>1041</td>
<td>UNESCO</td>
<td>ユネスコ</td>
</tr>
<tr>
<td>1097</td>
<td>エキサイト</td>
<td>excite</td>
</tr>
<tr>
<td>1131</td>
<td>ダンス</td>
<td>dance</td>
</tr>
<tr>
<td></td>
<td>ディライト</td>
<td>delight</td>
</tr>
<tr>
<td>1138</td>
<td>トヨタ</td>
<td>Toyota</td>
</tr>
<tr>
<td>1172</td>
<td>ディレクトリ</td>
<td>directory</td>
</tr>
</tbody>
</table>
## Evaluation result (TYPE=A)

<table>
<thead>
<tr>
<th></th>
<th>DCG-3-0</th>
<th>WRR-1-0</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM+Syn+C</td>
<td>2.522</td>
<td>0.605</td>
</tr>
<tr>
<td>AM+Syn</td>
<td>2.499</td>
<td>0.600</td>
</tr>
<tr>
<td>AM</td>
<td>2.464</td>
<td>0.596</td>
</tr>
<tr>
<td>DM+Syn</td>
<td>2.460</td>
<td>0.593</td>
</tr>
<tr>
<td>DM</td>
<td>2.431</td>
<td>0.590</td>
</tr>
<tr>
<td>C</td>
<td>0.381</td>
<td>0.080</td>
</tr>
</tbody>
</table>
Analysis

• We analyzed the results of AM+Syn+C by topic subcategories
  – Type
  – Category
  – Specialty
## Topic Type (TYPE=A)

<table>
<thead>
<tr>
<th>Topic Type</th>
<th>#Topics</th>
<th>DCG-3-0</th>
<th>WRR-1-0</th>
</tr>
</thead>
<tbody>
<tr>
<td>keyword</td>
<td>145</td>
<td>3.101</td>
<td>0.767</td>
</tr>
<tr>
<td>keywords</td>
<td>96</td>
<td>2.033</td>
<td>0.446</td>
</tr>
<tr>
<td>incomplete</td>
<td>28</td>
<td>1.383</td>
<td>0.356</td>
</tr>
</tbody>
</table>

If a query can be expressed by a single keyword precisely, the performance was better than those for other cases.
### Topic category (TYPE=A)

<table>
<thead>
<tr>
<th></th>
<th>#Topics</th>
<th>DCG-3-0</th>
<th>WRR-1-0</th>
</tr>
</thead>
<tbody>
<tr>
<td>product</td>
<td>49</td>
<td>2.256</td>
<td>0.540</td>
</tr>
<tr>
<td>company</td>
<td>60</td>
<td>3.071</td>
<td>0.717</td>
</tr>
<tr>
<td>person</td>
<td>29</td>
<td>2.376</td>
<td>0.517</td>
</tr>
<tr>
<td>facility</td>
<td>29</td>
<td>2.502</td>
<td>0.637</td>
</tr>
<tr>
<td>sight</td>
<td>16</td>
<td>2.206</td>
<td>0.649</td>
</tr>
<tr>
<td>resource</td>
<td>47</td>
<td>2.403</td>
<td>0.555</td>
</tr>
<tr>
<td>online shop</td>
<td>29</td>
<td>2.329</td>
<td>0.598</td>
</tr>
<tr>
<td>event</td>
<td>19</td>
<td>3.117</td>
<td>0.768</td>
</tr>
</tbody>
</table>
Queries produced by specialists did not match with anchor texts produced by “general” people

<table>
<thead>
<tr>
<th>Specialty</th>
<th>#Topics</th>
<th>DCG-3-0</th>
<th>WRR-1-0</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>62</td>
<td>2.577</td>
<td>0.592</td>
</tr>
<tr>
<td></td>
<td>106</td>
<td>2.720</td>
<td>0.632</td>
</tr>
<tr>
<td>Low</td>
<td>73</td>
<td>2.654</td>
<td>0.669</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>1.594</td>
<td>0.435</td>
</tr>
</tbody>
</table>
Example of mismatch b/w specialist query and anchor text

- Topic 1063
  - query: Yahoo housing information
  - anchor text: Yahoo real estate
Conclusion

• Following methods were effective for Navigational Web Retrieval
  – Anchor model
  – Synonym-based query expansion
  – Combination of anchor and content retrieval

Please visit WEB-7