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

Atsushi Fujii (Univ of Tsukuba)
Katunobu Itou (Nagoya Univ)
Tomoyoshi 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
  – In NTCIR-5 Navi-2 Subtask, a hypothetical user requires to find the representative pages of an item (e.g., person and product)

• Informational
  – A user searches for Web sites that provide knowledge for his/her information need
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 $P(d)$
  – The 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

$$\frac{1}{\alpha \times \frac{1}{3} + \beta \times \frac{1}{1} + \gamma \times \frac{1}{4}}$$
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
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

$P(q \mid d)$

#inlinks of $d$

#inlinks in collection
Computation of $P(q \mid d)$

- We assume independence of terms in $q$

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

- We use $P(t)$ if term $t$ is not modeled
- ChaSen is used to extract term $t$

- We compare the effectiveness of 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 \( d \)

\[ \text{surrogate document } SD \]

\[
P(t \mid d) = \frac{Freq(t, SD)}{\sum_{t} Freq(t, SD)}
\]
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
- Computation of $P(t \mid d)$ is affected by $a_2$ significantly

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
Query expansion in $P(t \mid d)$

- If term $t$ is not modeled in our system, we use synonym term $s$

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

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

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

```
www.google.co.jp
googles
グーグル（guuguru）
```

“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]

\[
google / グーグル (guuguru) \rightarrow Yes
\]
\[
google / エンジン (engine) \rightarrow No
\]
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>
## Evaluation result (TYPE=AB)

<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>2.203</td>
<td>0.529</td>
</tr>
<tr>
<td>AM+Syn</td>
<td>2.182</td>
<td>0.524</td>
</tr>
<tr>
<td>AM</td>
<td>2.152</td>
<td>0.521</td>
</tr>
<tr>
<td>DM+Syn</td>
<td>2.148</td>
<td>0.518</td>
</tr>
<tr>
<td>DM</td>
<td>2.124</td>
<td>0.516</td>
</tr>
<tr>
<td>C</td>
<td>0.333</td>
<td>0.070</td>
</tr>
</tbody>
</table>