Web-Page Summarization Using Clickthrough Data

Sun, Shen, Zeng, Yang, Lu, and Chen, SIGIR 2005

CSE 450 Web Mining Seminar
Jian Wang
Introduction

- Extract-based generic Web-page summarization
- To utilize extra knowledge to improve Web-page summarization, i.e., clickthrough dataset
- To build a thematic lexicon for web pages that have no associate query words
Empirical Study on Clickthrough Data

- The collection of queries is supposed to well reflect the topics of the target Web page.

- Whether the query words are related with the topics of the Web page.
  - 45.5% of keyword occurs in the query words
  - 13.1% of query words appear as keyword.

- To prove clickthrough data is helpful to summarize Web pages.
  - 58% -> 71.3%
  - 1.48 -> 2.0
Adapted Web-Page Summarization Methods

1. Adapted Significant Word (ASW) Method

- Each sentence is assigned a significance factor
- Significant words are selected according to word frequency in a document
- They modified the significant word selection method

\[ w_i = (1 - \alpha) \cdot \frac{t_{f_i}^p}{\max (t_{f_i}^p)} + \alpha \cdot \frac{t_{f_i}^q}{\max (t_{f_i}^q)} \]
Adapted Web-Page Summarization Methods

1. Adapted Significant Word (ASW) Method

• Significant factor of a sentence is computed by:

  1. Set a limit L for the distance
  2. Find a proportion in the sentence
  3. Count the number of significant words contained in the portion and divide the square of this number by the total number of words within the portion
Adapted Web-Page Summarization Methods

2. Adapted Latent Semantic Analysis (ALSA) Method

- A term-sentence matrix is constructed from the original text document
  \[ X_k = U_k \Sigma_k V_k^T \]

- LSA analysis is conducted on the matrix
  \[ \varphi_i = [v_{i1}, v_{i2}, \ldots, v_{ir}]^T \]

- For the most important concept, the sentence having the largest importance factor is selected into the summary
Adapted Web-Page Summarization Methods

2. Adapted Latent Semantic Analysis (ALSA) Method

- If a term occurs as query words, its weight is increased according to its frequency in query word collection.

- A term frequency TF approach:

\[ w_i = (1 - \beta) \cdot \frac{tf_i^p}{\max (tf_i^p)} + \beta \cdot \frac{tf_i^q}{\max (tf_i^q)} \]
Summarizing Web Pages Not Covered by Clickthrough Data

- Build a hierarchical lexion using the clickthrough data
  - (1) TS for each category is empty
  - (2) the query words are added into TS, its frequency is added to its original weight in TS
  - (3) term weight in each TS is multiplied by its Inverse Category Frequency (ICF)

- Weights of the terms in TS can be used to select significant words or to update the term-sentence matrix.
Experiments

- Data Set
  - MSN search engine & ODP directory
  - 1,125,207 web pages; 260,763 are clicked; 1,586,472 queries

- DAT1
  - 90 pages
  - summarized by 3 human evaluators

- DAT2
  - 10,000 pages
  - The description is the ideal summary
Performance Evaluation

- Precision, Recall, and F1

\[ P = \frac{|S_{ref} \cap S_{cand}|}{|S_{cand}|}, \quad R = \frac{|S_{ref} \cap S_{cand}|}{|S_{ref}|}, \quad F_1 = \frac{2PR}{P+R} \]

- ROUGE Evaluation

\[ \text{ROUGE} - N = \frac{\sum_{S \in \{\text{ref}\}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \{\text{ref}\}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)} \]
Experimental Results and Analysis

- Summarization Results on DAT₁

Figure 1: Summarization results using ASW method, on DAT1 with query words
Experimental Results and Analysis

- Summarization Results on DAT1

Figure 2: Summarization results using ALSA method, on DAT1 with query words
Experimental Results and Analysis

- Summarization Results on DAT1

Figure 3: Summarization results using ASW method, on DAT1 without queries
Experimental Results and Analysis

- Summarization Results on DAT1

Figure 4: Summarization results using ALSA method, on DAT1 without queries
Experimental Results and Analysis

- Summarization Results on DAT2

![Graph showing ROUGE-1 results for different methods on DAT2](image)

**Figure 5:** Summarization results on DAT2, evaluated by ROUGE-1 measure
### Experimental Results and Analysis

- **Discussions**

<table>
<thead>
<tr>
<th>Computer/Software</th>
<th>Computer/Hardware</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Before ICF</strong></td>
<td><strong>After ICF</strong></td>
</tr>
<tr>
<td>mapquest</td>
<td>mapquest</td>
</tr>
<tr>
<td>com</td>
<td>kaza</td>
</tr>
<tr>
<td>www</td>
<td>winmx</td>
</tr>
<tr>
<td>free</td>
<td>webshot</td>
</tr>
<tr>
<td>kaza</td>
<td>winzip</td>
</tr>
<tr>
<td>wallpaper</td>
<td>winrar</td>
</tr>
<tr>
<td>screensav</td>
<td>imesh</td>
</tr>
<tr>
<td>christma</td>
<td>winamp</td>
</tr>
<tr>
<td>screen</td>
<td>spybot</td>
</tr>
<tr>
<td>download</td>
<td>microsoft</td>
</tr>
</tbody>
</table>
Conclusions and Future Work

- We leverage extra knowledge from clickthrough data to improve Web-page summarization.
- Trade-off parameter
- How to leverage other types of knowledge
- Evaluate the methods using extrinsic evaluation metrics and much larger data sets.