Improving Pseudo-Relevance Feedback in Web Information Retrieval

Using Web Page Segmentation

Shipeng Yu, Deng Cai, Ji-Rong Wen and Wei-Ying Ma
Microsoft Research Asia
Peking University
Tsinghua University
• Problems of web page as
  – Web page usually contains not only pure content
    • Noise: navigation, decoration, interaction, …
  – One web page may contain multiple topics
    • Usually only a few topics are relevant to user query
Segmenting Web Pages to Detect Noise and Multi-topics

- *A News Page*:
  - Dolphin dies at Clearwater Marine Aquarium
  - Dolphin named Halma died at the Clearwater Marine Aquarium on Monday, the Tampa Tribune reported.
  - Halma died from a bacterial infection, which officials suspect was caused by a fisherman orangler that punctured her abdomen. The dolphin did not respond to any antibiotics.
  - The bottlenose dolphin at the aquarium in 1994 for long-term rehabilitation after receiving severe shark bites and cuts to its finning lines. Halma was also known as Sonnet Sam’s companion, the aquarium’s other resident dolphin.

- *Yahoo! Shopping Page*:
  - Leadership: How to Run Your Business Like the Pros
  - Format: Hardcover
  - About the book: Writing in his familiar voice — a New Yorker's bluntness, leavened by his passion for ideas — Rudolph Giuliani demonstrates in Leadership how the leadership skills he practices can be employed successfully by anyone who has to do nothing. After all, until 9/11 attacks on the...
Pseudo-relevance Feedback & Web Page Segmentation

• Pseudo-relevance feedback
  – Two rounds of retrieval
  – Choosing expansion terms from top-ranked documents
    • Quality of expansion terms heavily relies on the relevance of contents

• Web page segmentation can help to separate noisy information and distinguish multiple topics
  – Choosing expansion terms
  – Improving the quality of expansion terms
Vision-based Page Segmentation (VIPS)

- **Goal**: Construct a vision-based content structure for web page

- **Assumption**: Content structure based on visual display reflects semantic partition of content, and visual cues help to build content structure
  - Visual cues: position, size, distance, color, font, …
  - Can be extracted from HTML elements and attributes
Related Work: DOM-based Page Segmentation

• Page segmentation using DOM
  – Extract structural tags such as P, TABLE, UL, TITLE, H1~H6, etc

• Page segmentation using DOM, content and link
  – Record boundary discovery by heuristics
  – Fine-grained topic distillation by link analysis

• Using FOM (Function-based Object Model)
  – Define a function for each object and partition the page based on these functions

• *DOM does not necessarily reflect semantic partition*
An Example of Vision-based Content Structure
The VIPS Algorithm

Steps:

1. Visual Block Extraction
2. Visual Separator Detection
3. Content Structure Construction
4. Iterating the Above Steps

Steps:

• Iteratively find all appropriate visual blocks
• Visual cues in this stage
  - Tag cue
  - Color cue
  - Text cue
  - Size cue
• Visual Separator
  - horizontal or vertical line
  - visually cross with no blocks
• Weight of Separator
  - Set weight to each separator according to some patterns
• Maximally-weighted separators are chosen as the real separators
• Blocks that are not separated are merged
• Calculate DOC for each block
• Each block is checked if they meet the granularity requirement (i.e., if $DOC > PDOC$)
  - For those that fail, iteratively partition them
Pseudo-Relevance Feedback Based on VIPS

- **Idea:** Choosing expansion terms from top ranked segments

- **Steps:**
  - Initial Retrieval
    - Use any traditional IR methods
  - Page Segmentation
    - Apply VIPS on top $N$ documents to get a segment set
  - Segment Selection
    - Select top segments based on some rank methods
  - Expansion Term Selection
    - A PRF approach to select top terms for expansion
  - Final Retrieval
    - Adjust term weights and do retrieval with expanded query
• Perform pseudo-relevance feedback based on three methods
  – **VIPS**: our vision-based page segmentation
  – **DOMPS**:
    • A DOM-based page segmentation
    • Some structural tags are extracted as segments
  – **FULLDOC**
    • No segmentation is performed
Experiments – Setup

• **Data**
  – Web pages: TREC’s WT10g corpus, 1.69 million pages
  – Queries: 50 TREC-2001 ad hoc topics (topics 501-550)

• **Retrieval System**
  – Okapi: term weight BM2500

• **Preprocessing**
  – Reject pages larger than 700K
  – A small stop-word list (about 220 words)
  – Do not use stemming
  – Do not consider phrase

• **Baseline of average precision**: 16.55%
Comparison of average precision among three methods

<table>
<thead>
<tr>
<th>Number of Segments</th>
<th>Baseline (%)</th>
<th>FULLDOC (%)</th>
<th>DOMPS (%)</th>
<th>VIPS (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>16.55</td>
<td>17.56 (+6.10)</td>
<td>17.94 (+8.40)</td>
<td>18.01 (+8.82)</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>17.46 (+5.50)</td>
<td>18.15 (+9.67)</td>
<td>19.39 (+17.16)</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td><strong>19.10 (+15.41)</strong></td>
<td>18.05 (+9.06)</td>
<td>19.92 (+20.36)</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>17.89 (+8.10)</td>
<td>19.24 (+16.25)</td>
<td><strong>20.98 (+26.77)</strong></td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>17.40 (+5.14)</td>
<td>19.32 (+16.74)</td>
<td>19.68 (+18.91)</td>
</tr>
<tr>
<td>40</td>
<td></td>
<td>15.50 (-6.34)</td>
<td>19.57 (+18.25)</td>
<td>17.24 (+4.17)</td>
</tr>
<tr>
<td>50</td>
<td></td>
<td>13.82 (-16.50)</td>
<td><strong>19.67 (+18.85)</strong></td>
<td>16.63 (+0.48)</td>
</tr>
<tr>
<td>60</td>
<td></td>
<td>14.40 (-12.99)</td>
<td>18.58 (+12.27)</td>
<td>16.37 (-1.09)</td>
</tr>
</tbody>
</table>
Experiments – Results (Cont.)

Different performance peaks
Case Study (1) Impact of Noise: Expansion Terms

- **Query #15: What about Alexander Graham Bell?**

<table>
<thead>
<tr>
<th>Methods</th>
<th>FULLDOC</th>
<th>DOMPS</th>
<th>VIPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expansion Terms</td>
<td>iath notebooks deaf invention gorman simulating 165 1876 inventors Invent</td>
<td>audiotelegraph 1876 mehalik oblon prototype industriously technology tinkerer telephone spattered</td>
<td>invention deaf 1876 telephone divestiture edison telephones lineage 1885 inventors</td>
</tr>
<tr>
<td>Avg. Precision</td>
<td>14.13%</td>
<td>35.06%</td>
<td>50.89%</td>
</tr>
</tbody>
</table>
Case Study (1) Impact of Noise: Sample Pages

- “IATH”: appears on the bottom of the page “IATH WWW Server”
- “Gorman”: the author of the book “Alexander Graham Bell’s Path to the Telephone”
Case Study (2) Impact of Multiple Topics: Expansion Terms

- **Query #17: Titanic what went wrong?**

<table>
<thead>
<tr>
<th>Expansion Terms</th>
<th>FULLDOC</th>
<th>DOMPS</th>
<th>VIPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iceberg, sri, call, maybe, actually, noticed, difference, picked, theater, Reasons</td>
<td>iceberg, sri, sinking, rms, iceberg, hmhs, passengers, tragedy, ballard, sank, seawifs</td>
<td>sinking, unsinkable, iceberg, britannic, rms, passengers, 1912, ship, cyberflix, ocean</td>
<td></td>
</tr>
<tr>
<td>Avg. Precision</td>
<td>16.20%</td>
<td>15.23%</td>
<td>30.43%</td>
</tr>
</tbody>
</table>
Case Study (2) – Impact of Multiple Topics:
Sample Pages

Titanic Effect

<John@HITS-BILLERICA-MULTICS.ARPA>
Wed, 26 Feb 96 12:24 EST
To: risks@SRI-CLL.ARPA

I suppose if I had said to the designer of the maritime accident on record involved the boat compartments, so let's plan on surviving five, specified smaller compartments, so that there is not sixteen, compartments. And the icebergs were compartments...

Computers placing telephone calls

“Art Evans” <Evans@TI-205.ARPA>
Wed 26 Feb 96 14:18:23 EST

The Titanic Effect

In conclusion, it is very easy, with today's thing. Modern technology has even progressed to the point where the numbers until a connection is made, fact that there will never be a machine on the telephone.

We have always had wrong numbers. However, at number, there is (almost) immediate confirmation and a second or third or tenth retry is not at

Maybe what we need as a touch tone code (or so enter into a nodes that says “The number you h

“call” and “maybe” appear in neighboring topics
## Conclusion

- Introduce a vision-based content structure
- Present a vision-based page segmentation algorithm
- Verify that page segmentation can benefit pseudo-relevance feedback
- Indirectly evaluate our page segmentation algorithm
Thank you for your attention