Data Cleansing for Web Information Retrieval using Query Independent Features

Yiqun Liu, Min Zhang, Liyun Ru, Shaoping Ma
State Key Lab of Intelligent Tech. & Sys.
Tsinghua University
Outlines

• Data cleansing and its applications in Web IR
• Query-independent features used in data cleansing
• Algorithm and evaluation
• Conclusions and future work
Data cleaning and its applications in Web IR

- Index Size War between Search Engines
  - Billions Of Textual Documents Indexed
  - December 1995-September 2003

From Danny Sullivan, SearchEngineWatch web site
Data cleansing and its applications in Web IR

- Index Size War between Search Engines (cont.)

<table>
<thead>
<tr>
<th>Search Engine</th>
<th>Reported Size</th>
<th>Page Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>8.1 billion (Dec. 2004)</td>
<td>101K</td>
</tr>
<tr>
<td>MSN</td>
<td>5.0 billion</td>
<td>150K</td>
</tr>
<tr>
<td>Yahoo</td>
<td>19.2 billion (Aug. 2005)</td>
<td>500K</td>
</tr>
<tr>
<td>Ask Jeeves</td>
<td>2.5 billion</td>
<td>101K+</td>
</tr>
<tr>
<td>All the Web</td>
<td><strong>152 billion</strong></td>
<td>605K</td>
</tr>
<tr>
<td>All the Surface Web</td>
<td><strong>10 billion</strong></td>
<td>8K</td>
</tr>
</tbody>
</table>

From Danny Sullivan, SearchEngineWatch web site
Data cleansing and its applications in Web IR

- An end to the index size war?
  - No search engine can cover all resources on the Web

<table>
<thead>
<tr>
<th></th>
<th>Google</th>
<th>Yahoo!</th>
<th>MSN</th>
<th>Teoma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Round 1</td>
<td>76.30%</td>
<td>69.28%</td>
<td>62.03%</td>
<td>57.58%</td>
</tr>
<tr>
<td>Round 2</td>
<td>76.09%</td>
<td>69.29%</td>
<td>61.90%</td>
<td>57.69%</td>
</tr>
<tr>
<td>Round 3</td>
<td>76.27%</td>
<td>69.37%</td>
<td>61.87%</td>
<td>57.70%</td>
</tr>
<tr>
<td>Round 4</td>
<td>76.05%</td>
<td>69.30%</td>
<td>61.73%</td>
<td>57.57%</td>
</tr>
<tr>
<td>Round 5</td>
<td>76.11%</td>
<td>69.26%</td>
<td>61.96%</td>
<td>57.56%</td>
</tr>
<tr>
<td>Average</td>
<td><strong>76.16%</strong></td>
<td><strong>69.32%</strong></td>
<td><strong>61.90%</strong></td>
<td><strong>57.62%</strong></td>
</tr>
</tbody>
</table>

- In Sep. 2005, Google removes the number of indexed pages because “absolute numbers are no longer useful”
Data cleansing and its applications in Web IR

• Data quality is more important than quantity for Web IR tools
  – Spams and SEOs
  – Duplicates in Web pages
  – Unreliable, out-dated data

• Current data cleansing algorithms in Web IR
  – Local scale data cleansing
  – Global scale data cleansing
Data cleansing and its applications in Web IR

• Local scale data cleansing
  – To reduce the useless blocks / To find the important blocks inside a Web page
  – Reduce spam hyperlinks / useless hyperlinks (Kushmerick et. al.)
  – Reduce Ad. Contexts (Davison et. al.)
  – VIsion Based Page Segmentation, VIPS, MSRA
  – Site template detecting (Yossef et. al.)
Data cleansing and its applications in Web IR

• Global scale data cleansing
  – To reduce low quality pages / To locate important pages inside a given Web page corpus
  – Hyperlink structure analysis algorithms
    • PageRank, HITS
    • Hypothesis 1: Recommendation
    • Hypothesis 2: Topic locality
    • Challenged by Spam links and SEOs
  – Monika Henzinger (Google Research Director): A better estimate of the quality of a page requires additional sources of information.
Data cleansing and its applications in Web IR

- Our data cleansing method
  - Global scale data cleansing
  - Learn from “what users need”
  - Users’ information requirement is reflected in their search target pages (pages that they want to find)
  - A better data cleansing method should judge the quality of a Web page by whether it can be a search target for a certain user query.
  - Both hyperlink structure features and other kinds of features should be considered in data cleansing
Data cleansing and its applications in Web IR

- Query-independent Data Cleansing

Data Cleansing Process is independent of Queries
Outlines

- Data cleansing and its applications in Web IR
- Query-independent features used in data cleansing
- Algorithm and evaluation
- Conclusions and future work
Query-independent features used in data cleansing

• Query-independent feature analysis of High Quality Pages
  – Corpus
    • 37M Chinese web pages collected in Nov. 2005
    • Over 0.5 Terabyte.
    • Obtained from Sogou.com
  – High Quality Page (Search Target Page)
    • Training set: 1600 pages
    • Test set: 17000 pages
    • Evaluated manually by Sogou engineers
Query-independent features used in data cleansing

• Hyperlink structure related features
  – PageRank
  – In-link number
  – In-link anchor text length

• Other features
  – Document length
  – Number of duplicates
  – URL length
  – Encode
Query-independent features used in data cleansing

- PageRank
Query-independent features used in data cleansing

- In-link anchor text length
Query-independent features used in data cleansing

- Document length

![Graph showing document length distribution](image)

Ordinary Retrieval Target
Query-independent features used in data cleansing

- URL Length
Query-independent features used in data cleansing

- Other features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Ordinary</th>
<th>High Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>URL contains “?”</td>
<td>13.06%</td>
<td>1.87%</td>
</tr>
<tr>
<td>Encode is not GBK</td>
<td>14.04%</td>
<td>1.39%</td>
</tr>
<tr>
<td>Hub type page</td>
<td>3.78%</td>
<td>24.77%</td>
</tr>
</tbody>
</table>

- The query-independent features can separate high quality pages from ordinary pages
Outlines

• Data cleansing and its applications in Web IR
• Query-independent features used in data cleansing
• Algorithm and evaluation
• Conclusions and future work
Algorithm and evaluation

- A learning based data cleansing algorithm
  - The possibility of one web page being a search target page is:

\[
P(p \in \text{Target page} \mid p \text{ has feature } A)\]

\[
P(p \in \text{Target page} \mid p \text{ has feature } A) = \frac{P(p \text{ has feature } A \mid p \in \text{Target page})}{P(p \text{ has feature } A)} \times P(p \in \text{Target page})
\]

\[
P(p \text{ has feature } A \mid p \in \text{Target page}) = \frac{\#(p \text{ has feature } A \cap p \in \text{Target page})}{\#(\text{Target page})} / \frac{\#(p \text{ has feature } A)}{\#(\text{Ordinary page})}
\]
Algorithm and evaluation

• General information of the cleansed corpus

<table>
<thead>
<tr>
<th></th>
<th>Current Size / Original Size</th>
<th>High Quality Recall (Training Set)</th>
<th>High Quality Recall (Test Set)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduced Page Set</td>
<td>95.04%</td>
<td>7.27%</td>
<td>7.63%</td>
</tr>
<tr>
<td>Cleansed Corpus</td>
<td>4.96%</td>
<td>92.73%</td>
<td>92.37%</td>
</tr>
</tbody>
</table>

• The cleansed corpus contains about 5% pages in the original corpus, but can meet 92% user needs.
Algorithm and evaluation

- Function of different features in our algorithm

Although PageRank plays an important role in the algorithm, we don’t rely on this single feature.
Algorithm and evaluation

- The possibility of reducing spam/low quality pages using our data cleansing algorithm

![Graph showing the reduction of spam and low quality pages for different methods: Data Cleansing, PageRank Only, Indegree Only. The graph indicates a significant reduction in spam and low quality pages for the Data Cleansing method compared to the other two methods.](image)
Outlines

- Data cleansing and its applications in Web IR
- Query-independent features used in data cleansing
- Algorithm and evaluation
- Conclusions and future work
Conclusions and future work

• Conclusions:
  – Query-independent features can separate Search Target Pages from ordinary pages
  – It is possible to reduce 95% web pages with a small loss in key information
  – The data cleansing algorithm can also reduce part of spam pages / low quality pages
Conclusions and future work

• Future work
  – Retrieval in the cleansed corpus
  – Hyper link analysis in the cleansed corpus
  – A learn-based algorithm to reduce spam pages / low quality pages
  – Personalized search
Thank you!

Questions or comments?