Learning-based Web Data Cleansing for Information Retrieval

Yiqun Liu, Canhui Wang, Min Zhang, 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 cleansing 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><strong>8.1 billion</strong></td>
<td>101K</td>
</tr>
<tr>
<td></td>
<td>(Dec. 2004)</td>
<td></td>
</tr>
<tr>
<td>MSN</td>
<td>5.0 billion</td>
<td>150K</td>
</tr>
<tr>
<td>Yahoo</td>
<td>19.2 billion</td>
<td>500K</td>
</tr>
<tr>
<td></td>
<td>(Aug. 2005)</td>
<td></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?
  – In Sep. 2005, Google removes the number of indexed pages because “absolute numbers are no longer useful”
  – 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>76.16%</td>
<td>69.32%</td>
<td>61.90%</td>
<td>57.62%</td>
</tr>
</tbody>
</table>
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
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

![Graph showing in-link anchor text length distribution for ordinary and retrieval target settings.](image-url)
Query-independent features used in data cleansing

• Other features

<table>
<thead>
<tr>
<th></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

• Difficulties in algorithms
  – Web page classification
  – Lack of negative examples (uniform sampling is difficult and sometimes not possible)
  – Learning with unlabeled data and positive examples
  – Previous work:
    • O-SVM
    • PEBL: Positive Example Based Learning
    • Not quite suitable for learning based on topic-independent features
Algorithm and evaluation

• Why is k-means used here?
  – Learn without negative examples
  – Independent of prior positive proportion knowledge

• Differences with traditional K-means
  – Fixed cluster number: true or not.
  – Initial positive example centroid is provided
Algorithm and evaluation

• Algorithm

$S_{key}$: key resource training set
$R$: estimated proportion of the positive examples

1. Choose 2 initial cluster centroids:
   
   – Positive centroid: $M_1 = \frac{1}{S_{key}} \sum_{X \in S_{key}} X$
   
   – Negative centroid: $M_2 = \frac{M(Whole \ Collection) - R \times M_1}{1 - R}$

2. In the kth iterative, instance $X$ will be assigned to the jth cluster $S_j^{(k)}$ if:

   $$||X - M_j^{(k)}|| = \min(||X - M_1^{(k)}||, ||X - M_2^{(k)}||) \quad (j = 1, 2)$$

3. For $S_j^{(k)}$, calculate $M_j^{(k)}$, which is defined as:

   $$M_j^{(k+1)} = \frac{1}{N_j} \sum_{X \in S_j^{(k)}} X \quad (j = 1, 2)$$

4. If $M_1^{(k+1)} = M_1^{(k)}$, exit. Else go to 2.
Algorithm and evaluation

- Algorithm converges with different initial $R$
  - Algorithm doesn’t require prior knowledge of $R$
Algorithm and evaluation

- Evaluation (Based on .GOV corpus)
  - Algorithm can cover almost all high quality pages with less than half whole collection size

<table>
<thead>
<tr>
<th></th>
<th>K-means Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole Collection (.GOV) Coverage</td>
<td>44.30%</td>
</tr>
<tr>
<td>High Quality Page Test Set Recall</td>
<td>89.70%</td>
</tr>
<tr>
<td>High Quality Page Test Set Precision</td>
<td>67.50%</td>
</tr>
<tr>
<td>F2-measure</td>
<td>53.89%</td>
</tr>
</tbody>
</table>

- Retrieval Experiment Settings
  - 20% navigational type queries
  - 80% informational/transactional type queries
Algorithm and evaluation

• Evaluation

<table>
<thead>
<tr>
<th></th>
<th>P@10 for Topic Distillation queries</th>
<th>MRR for Navigational query</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole Collection</td>
<td>0.1025</td>
<td>0.7443</td>
</tr>
<tr>
<td>K-means</td>
<td>0.1275</td>
<td>0.7278</td>
</tr>
<tr>
<td>PageRank</td>
<td>0.1134</td>
<td>0.6533</td>
</tr>
<tr>
<td>Authority</td>
<td>0.1100</td>
<td>0.6700</td>
</tr>
<tr>
<td>Hub</td>
<td>0.1250</td>
<td>0.6357</td>
</tr>
</tbody>
</table>

– Cleansed set gains better performance than whole collection
– K-means based cleansing outperforms link-analysis criterion
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:
  – Data cleansing based on K-means clustering is effective in reducing unimportant pages.
  – Cleansed set (half size of total collection) retains useful information of the Web collection.
  – Retrieval on result set gets better overall retrieval performance than the whole collection.
Conclusions and future work

- Future work
  - Algorithm Efficiency Problem
    - Naïve Bayes based learning method
      "Data Cleansing for Web Information Retrieval using Query Independent Features", to be appeared in JASIST, Jan, 2007
  - Hyper link analysis in the cleansed corpus
    - The cleansed corpus retains almost all hyper link information
  - A learn-based algorithm to reduce spam pages / low quality pages
    - Similar way: learn from positive example and unlabelled data
Thank you!

Questions or comments?