Detecting Phrase-Level Duplication on the World Wide Web

Dennis Fetterly, Mark Manasse Marc Najork
Microsoft Research
SIGIR’05

CSE 450 Web Mining Seminar
Presented by Liangjie Hong
March 24th, 2008
BACKGROUND

- **Types of Spam**
  - Content Spam
  - Link Spam
  - Redirection Spam

- **Content Spam**
  - Keyword stuffing
  - Hidden text
  - Meta stuffing
Motivation

- **Keyword Stuffing**
  - Page duplication
  - Word duplication
  - Phrase-level duplication

- **Characteristics**
  - Grammatically well-formed
  - Generated randomly
  - Assembled from various pages
**Finding Phrase Replication**

- **Representation of Documents**

In their practice, $m = 84$ $k = 5$
Finding Phrase Replication

- **Popular Shingles**
  - numbers & letters
  - navigational text
  - copyright notices
  - machine generated

<table>
<thead>
<tr>
<th>Rank</th>
<th>Number</th>
<th>Phrase (five words long)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2366046</td>
<td>head up face 3</td>
</tr>
<tr>
<td>2</td>
<td>1406904</td>
<td>j k l m n</td>
</tr>
<tr>
<td>3</td>
<td>1706905</td>
<td>11 12 13 14 15</td>
</tr>
<tr>
<td>4</td>
<td>1114293</td>
<td>15 16 17 18 19</td>
</tr>
<tr>
<td>5</td>
<td>1132100</td>
<td>t u v w x</td>
</tr>
<tr>
<td>6</td>
<td>1000200</td>
<td>n o p q r</td>
</tr>
<tr>
<td>7</td>
<td>372376</td>
<td>18 19 20 21 22</td>
</tr>
<tr>
<td>8</td>
<td>11615</td>
<td>2 3 4 5 6</td>
</tr>
<tr>
<td>9</td>
<td>746081</td>
<td>powered by phpbb 2.0</td>
</tr>
<tr>
<td>10</td>
<td>742210</td>
<td>copyright 2000 2004 jasoft enterprises</td>
</tr>
<tr>
<td>11</td>
<td>391484</td>
<td>More alphanumeric and numeric sequences, and a Jasoft copyright notice</td>
</tr>
<tr>
<td>12</td>
<td>328015</td>
<td>prev date next thread prev</td>
</tr>
<tr>
<td>25</td>
<td>175387</td>
<td>all rights reserved privacy policy</td>
</tr>
<tr>
<td>29</td>
<td>252044</td>
<td>user name remember not password</td>
</tr>
<tr>
<td>32</td>
<td>240301</td>
<td>bulletin go to http www</td>
</tr>
<tr>
<td>34</td>
<td>226688</td>
<td>today's posts mark forums</td>
</tr>
<tr>
<td>35</td>
<td>221541</td>
<td>may not post attachments you names</td>
</tr>
<tr>
<td>36</td>
<td>200530</td>
<td>nonexistent celebrity videos you_image</td>
</tr>
<tr>
<td>37</td>
<td>21044</td>
<td>f***ing free paris Hilton celebrity</td>
</tr>
<tr>
<td>38</td>
<td>214960</td>
<td>celebrity videos you name it</td>
</tr>
<tr>
<td>41</td>
<td>203478</td>
<td>minnesota mississippi missouri montana nebraska</td>
</tr>
<tr>
<td>46</td>
<td>199408</td>
<td>forum powered by phpbb 2</td>
</tr>
<tr>
<td>47</td>
<td>194739</td>
<td>send a private message to</td>
</tr>
<tr>
<td>55</td>
<td>171197</td>
<td>profile log in to check</td>
</tr>
<tr>
<td>60</td>
<td>159460</td>
<td>product you are looking for</td>
</tr>
</tbody>
</table>
Finding Phrase Replication

- Some Results with Popular Shingles

![Graphs showing distribution of popular shingles.](image)
Covering Sets

- Finding the minimum size of covering set is **NP-complete**
- Using **Greedy heuristic** to approximate
- More likely add documents from other hosts
COVERING SETS

Two Examples of Covering Sets

Table 2. Example covering sets from DS2.

http://www.searchingweb.com/cgi-bin/directory/searchingweb.cgi?Business/Investing/
http://www.moneywebsearch.com/directory/Information_Technology/
http://newhoo.com/Business/Investing/
http://www.specialistweb.com/about.shtml
http://www.dmoz.org/Arts/Literature/Authors/M/Munsey_&_Irenece/
http://www.calculatorweb.com/calculators/transactioncalc.shtml
http://www.forwardingweb.com/forward@join.shtml
http://www.calculatorweb.com/cgi-bin/directory/dclick.cgi?account=ButtonAffiliate
http://www.registerweb.com/cgi-bin/whois/whois.cgi?show_global=1
http://www.specialistweb.com/contact.shtml
http://www.searchingweb.com/cgi-bin/directory/searchingweb.cgi?Reference/
http://www.searchingweb.com/cgi-bin/directory/searchingweb.cgi?Business/
...
http://www.career-training-courses.com/education/TestPreparationStudies.htm
http://www.career-training-courses.com/education/careertraining.htm
http://experts.universaldass.com/dezra
http://www.writing-tools-courses.com/education/writinggenre.htm
http://www.business-software-courses.com/education/computer.htm
http://www.math-courses.com/course.htm
http://www.history-courses.com/education/counseling/sellimprovement.htm
http://www.domestic-violence-courses.com/education/ParentingandFamily.htm
http://www.curricityforkids.com/newthismatch.htm
http://www.ged-test-preparation.com/course.htm
http://www.counseling-courses.com/courses/satmath.htm
http://home.universaldass.com/cim/7514.htm
COVERING SETS

- Some Results about Covering Sets
CONCLUSIONS & FUTURE WORK

- A third of the pages on the web consists of more replicated than original content.
- High fraction of non-original phrases typically feature machine-generated content.
- Most popular phrases are not very interesting.
- Provide a way to estimate how original the content is.

*Cannot distinguish legitimate from spam content!*