The Evolution of Web Content and Search Engines

Ricardo Baeza-Yates
Yahoo! Research, DCC/UChile, UPF/Spain

Álvaro Pereira Jr
DCC/UFMG/Brazil

Nivio Ziviani
DCC/UFMG/Brazil
Objectives

- To state the following hypothesis:
  - When pages have sources (content originated from other pages), in a portion of pages there was a query that related the sources and made possible the creation of the new page
    - Part of the web content is biased by the ranking function of search engines

- To study how new content is generated in the web
  - How old content is used to compose new pages
  - Definition of genealogical trees for the web
## Web Collections and Query Logs

<table>
<thead>
<tr>
<th>Collection</th>
<th>Crawling date</th>
<th>Number of documents</th>
<th>Text size (Gbytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>Jul 2002</td>
<td>892,000</td>
<td>2.3</td>
</tr>
<tr>
<td>2003</td>
<td>Aug 2003</td>
<td>2.86 mi</td>
<td>9.4</td>
</tr>
<tr>
<td>2004</td>
<td>Jan 2004</td>
<td>2.80 mi</td>
<td>11.8</td>
</tr>
<tr>
<td>2005</td>
<td>Feb 2005</td>
<td>2.88 mi</td>
<td>11.3</td>
</tr>
</tbody>
</table>
Algorithm

- **Objective:**
  - To find in the new collection documents that were created using content from old documents, returned by the same query.

- For that we simulate a user performing a query in the search engine (TodoCL) in the past.

- We used a set of the most frequent queries of each query log:
  - We had access to the query processor of the search engine.

- Algorithm divided into two steps:
  - **First step:** finding new documents that has content from the old documents.
  - **Second step:** filter the documents.
Algorithm – Step 1: Finding Candidates

Old collection (O)

New collection (N)

Hash table
Algorithm – Step 2: Filtering

- Number of paragraphs in both old and new documents
- New document composed by two old documents returned by the same query
  - At least two distinct paragraphs from each old document
- The new document URL cannot exist in the old collection
- Duplicates are not allowed for both old and new documents
Experiments Summary

Log 2002  Log 2003  Log 2004
Jul/02  Aug/03  Jan/04  Feb/05

Real query log

2003 Old collection (O)

2004 New collection (N)

Q r(q₁, O) T
query 1
query 2
... query n
doc 3
doc 7
doc 13
doc 17
A₁
A₂

3 a. a. a. b. b. c. c. c. d. d. d.
13 g. g. g. d. d. d. h. h. i. i. i.
7 e. e. e. b. b. c. c. c. f. f. f.
17 i. i. i. e. e. e. f. f. f. a. a. a.

3, 7 13, 17
7, 17
13, 17
3, 17
13, 13
3, 7
7, 17

HASH TABLE

WEBKDD'06, August 20, 2006, Philadelphia, USA
An Experimental Result

- Different query logs on old collection 2003 and new collection 2004
Main components of the tree considering collection 2002 as the old collection

Sample of 120,000 documents

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of parents</td>
<td>5,900</td>
<td>4,900</td>
<td>4,300</td>
</tr>
<tr>
<td>Number of children</td>
<td>13,500</td>
<td>8,900</td>
<td>9,700</td>
</tr>
<tr>
<td>Number of survived pages</td>
<td>13,900</td>
<td>10,700</td>
<td>6,800</td>
</tr>
</tbody>
</table>
Conclusions

- We have presented evidences that a portion of the web is biased by the ranking function of search engines.
- A significant portion of the Web has evolved from old content.
- The number of copies from previously copied web pages (or content) is indeed greater than the number of copies from other pages.
  - Do search engines contribute to this situation?
Thank You!
Bimonthly Logs

Bimonthly logs 2002

1 2 3 4 5

Jul/02 Aug/03

Bimonthly logs 2004

1 2 3 4 5

Jan/04 Feb/05
Bimonthly Logs on the Same Collection

- Bimonthly logs 2002
- Bimonthly logs 2004

![Graphs showing new documents found vs. minimal number of identical paragraphs for bimonthly logs 2002 and 2004. Each log has a different line indicating the number of documents found at various thresholds of identical paragraphs.](attachment:image.png)
Bimonthly Logs in Different Collections

- Bimonthly logs 4 and 5 used for collection 2002
- Bimonthly logs 4 and 5 used for collection 2004

![Graph showing the relationship between the minimal number of identical paragraphs and the new documents found between log 2002 on col. 2002 and log 2004 on col. 2002.]

![Graph showing the relationship between the minimal number of identical paragraphs and the new documents found between log 2004 on col. 2004 and log 2002 on col. 2004.]
Chilean Web Genealogical Tree (2/2)

- Main component of the tree considering collection 2003 as the old collection
- Sample of 120,000 documents

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of parents</td>
<td>5,300</td>
<td>5,000</td>
</tr>
<tr>
<td>Number of children</td>
<td>33,200</td>
<td>29,100</td>
</tr>
<tr>
<td>Number of survived pages</td>
<td>19,300</td>
<td>10,500</td>
</tr>
</tbody>
</table>