The Evolution of Web Content and Search Engines

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

Álvaro Pereira Jr
DCC/UFMG/Brazil

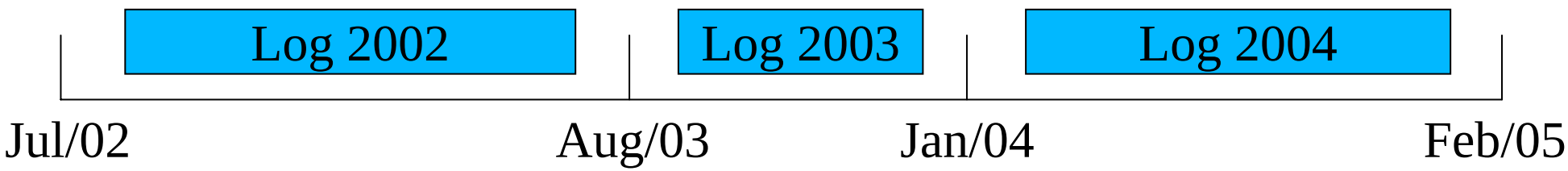
Nivio Ziviani
DCC/UFMG/Brazil

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

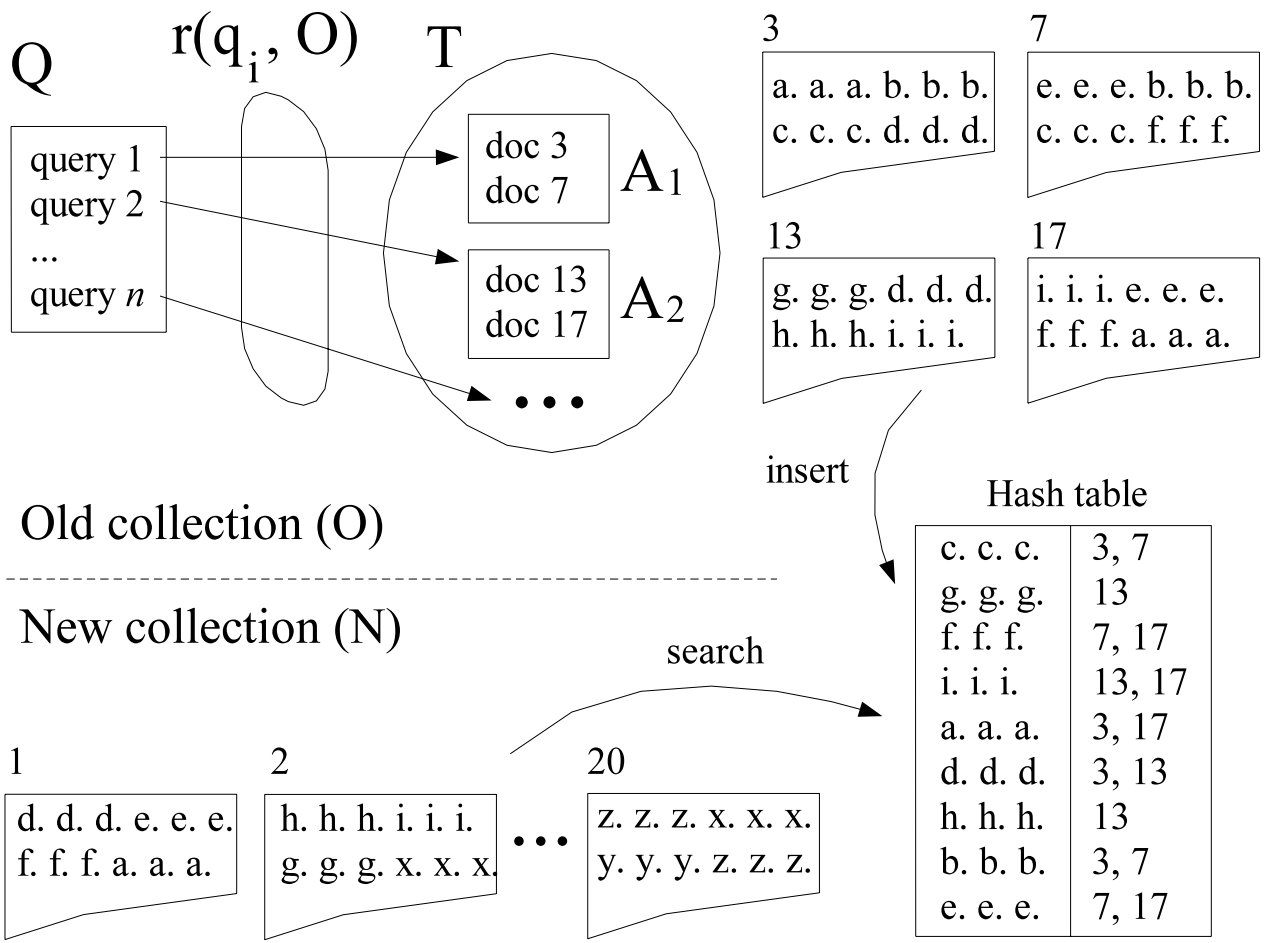
Collection	Crawling date	Number of documents	Text size (Gbytes)
2002	Jul 2002	892,000	2.3
2003	Aug 2003	2.86 mi	9.4
2004	Jan 2004	2.80 mi	11.8
2005	Feb 2005	2.88 mi	11.3



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

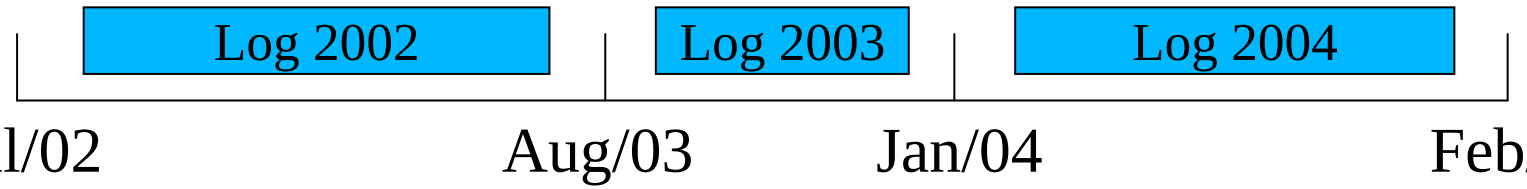
Algorithm – Step 1: Finding Candidates



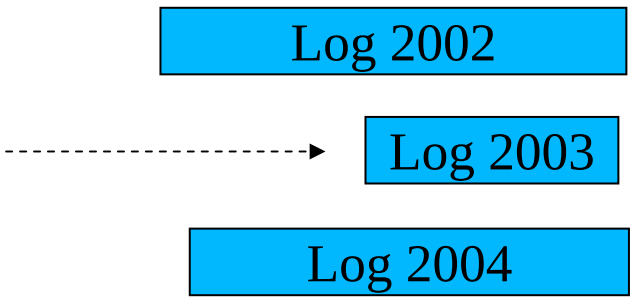
Algorithm – Step 2: Filtering

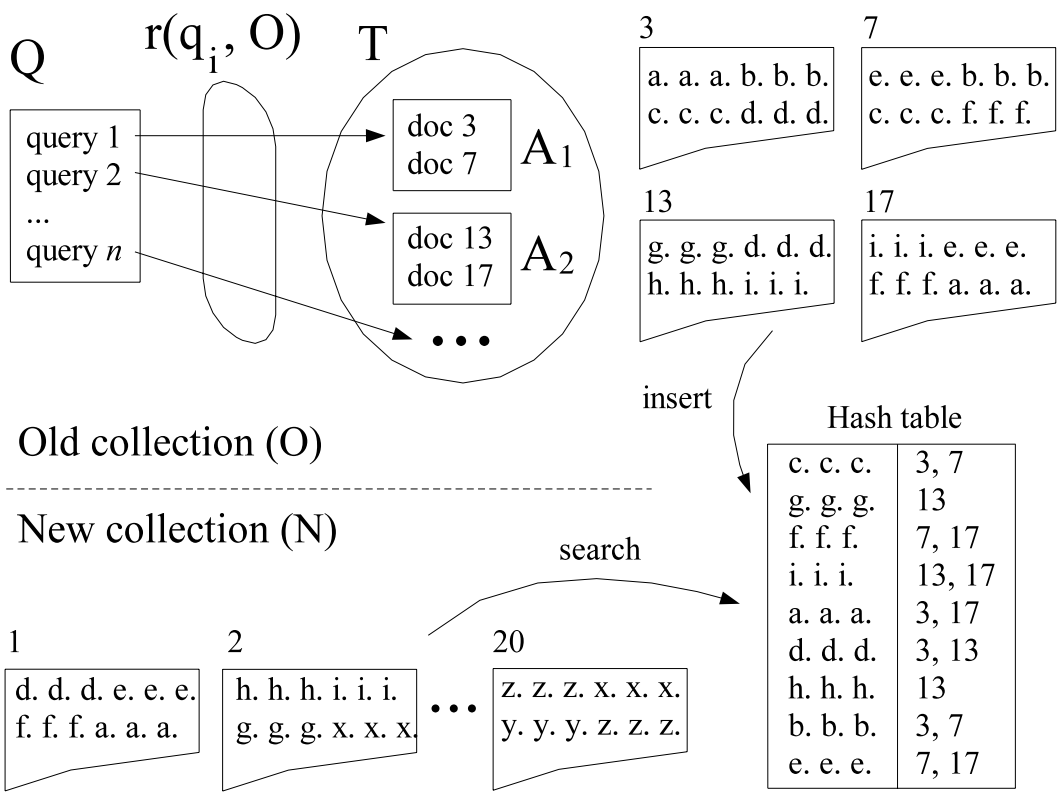
- Number of paragraphs in both old and new documents
- New document composed by two old documents returned by the same query
  - At least two distinct paragraphs from each old document
- The new document URL cannot exist in the old collection
- Duplicates are not allowed for both old and new documents

Experiments Summary



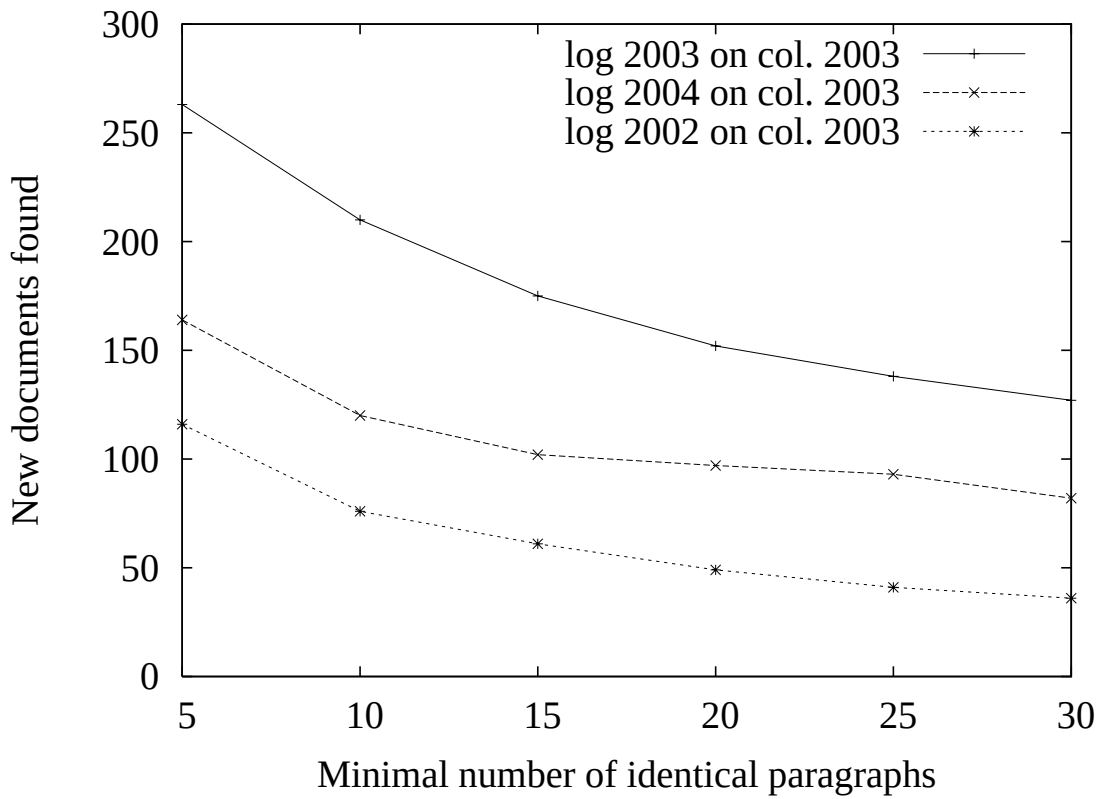
Real query log





An Experimental Result

- Different query logs on old collection 2003 and new collection 2004



Chilean Web Genealogical Tree

- Main components of the tree considering collection 2002 as the old collection
- Sample of 120,000 documents

Collection pairs	2002-2003	2002-2004	2002-2005
Number of parents	5,900	4,900	4,300
Number of children	13,500	8,900	9,700
Number of survived pages	13,900	10,700	6,800

Conclusions

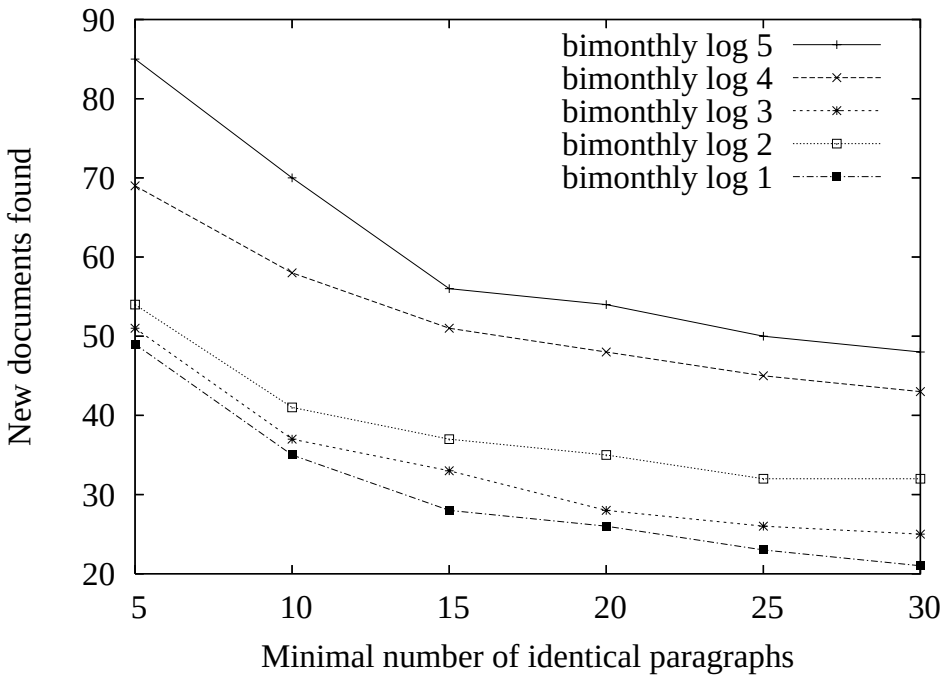
- We have presented evidences that a portion of the web is biased by the ranking function of search engines
- A significant portion of the Web has evolved from old content
- The number of copies from previously copied web pages (or content) is indeed greater than the number of copies from other pages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

Thank You!

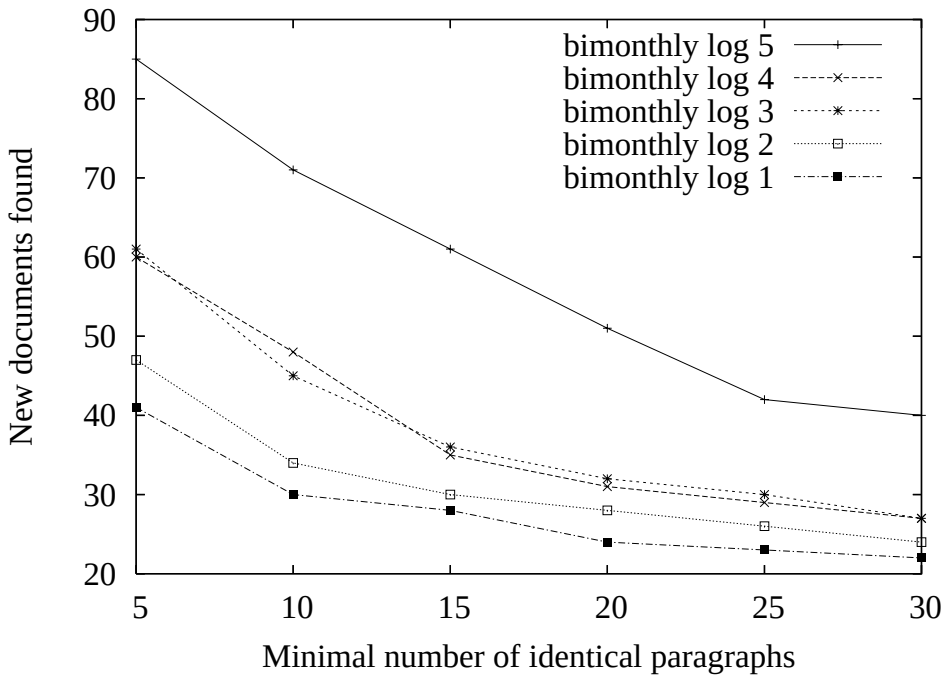
Bimonthly Logs

Bimonthly Logs on the Same Collection

- Bimonthly logs 2002

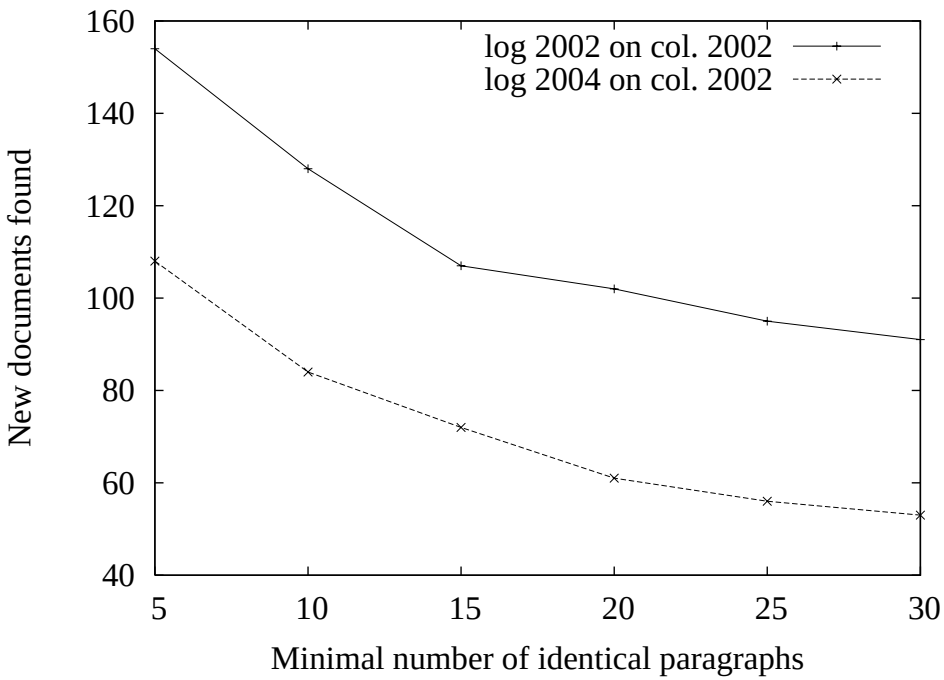


- Bimonthly logs 2004

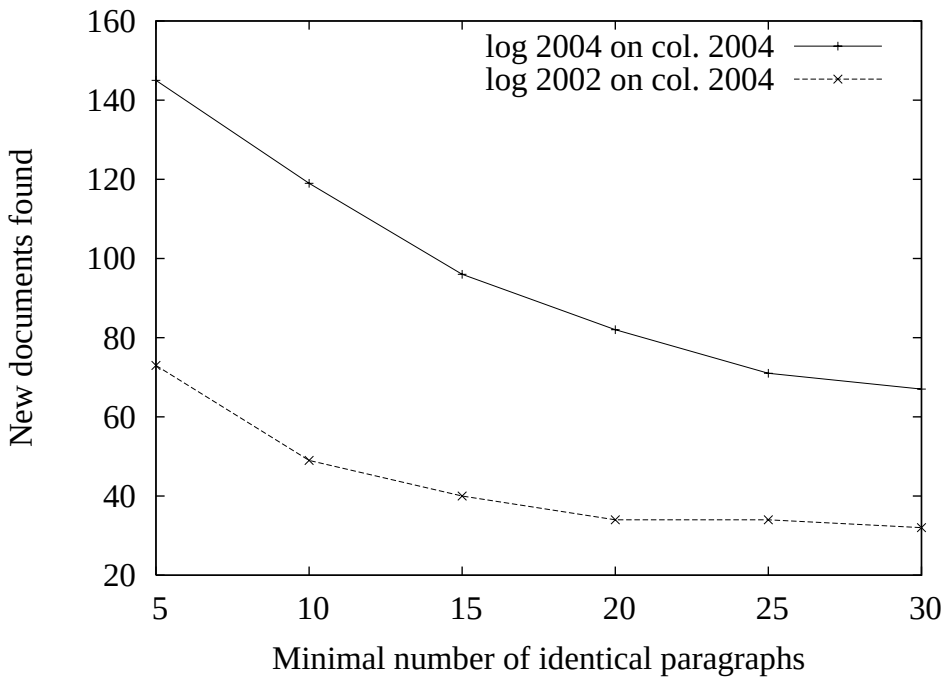


Bimonthly Logs in Different Collections

- Bimonthly logs 4 and 5 used for collection 2002



- Bimonthly logs 4 and 5 used for collection 2004



Chilean Web Genealogical Tree (2/2)

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

Collection pairs	2003-2004	2003-2005
Number of parents	5,300	5,000
Number of children	33,200	29,100
Number of survived pages	19,300	10,500