The anatomy of a large-scale hypertextual Web search engine

-Sergey Brin, Lawrence Page

Prototype of Search Engine

-Presented by

Sonam Gupta

What motivated the authors?

- Increasing information on web
- World Wide Web Worm- one of the first search engines
- Limited information provided by other search engines
- Googol started its development with more features and efficiency

Introduction

- Improvement in the idea about good quality search engine
- Increase in number of users on web
- Fast web crawling technology, efficient use of storage space
- Scalability of Google
- Exceptions to hardware performance

Their goals…

- According to **Best of the Web 1994-Navigators**, ”The best navigation service should make it easy to find almost anything on the Web (once **all the data** is entered).”
- Users usually go through top ten results
- Relevance of the resulted information needs to be focused.
- In need of prioritizing the relevant data

Goals (contd.)

- Use of link structure, anchor text, meta data, proximity data
- Link structure- judges relevance
- Towards the academic research
- Rapid increase in the use of .com domains
- Need to develop more research activities on Web
- Interactive environment for researchers

The Algorithm

- Need of a certain way to get high precision of search results
- What is PageRank?
  - Creation of maps
  - PR(A) = (1-d)+d(PR(T1)/C(T1) +..+PR(Tn)/C(Tn))
- Counts the links, on the page
- Shows the importance of web pages
- Iterative algorithm
- Idea of a “random surfer”
- Many pages pointing to a page with high PageRank
- Less pages pointing to a page with high PageRank

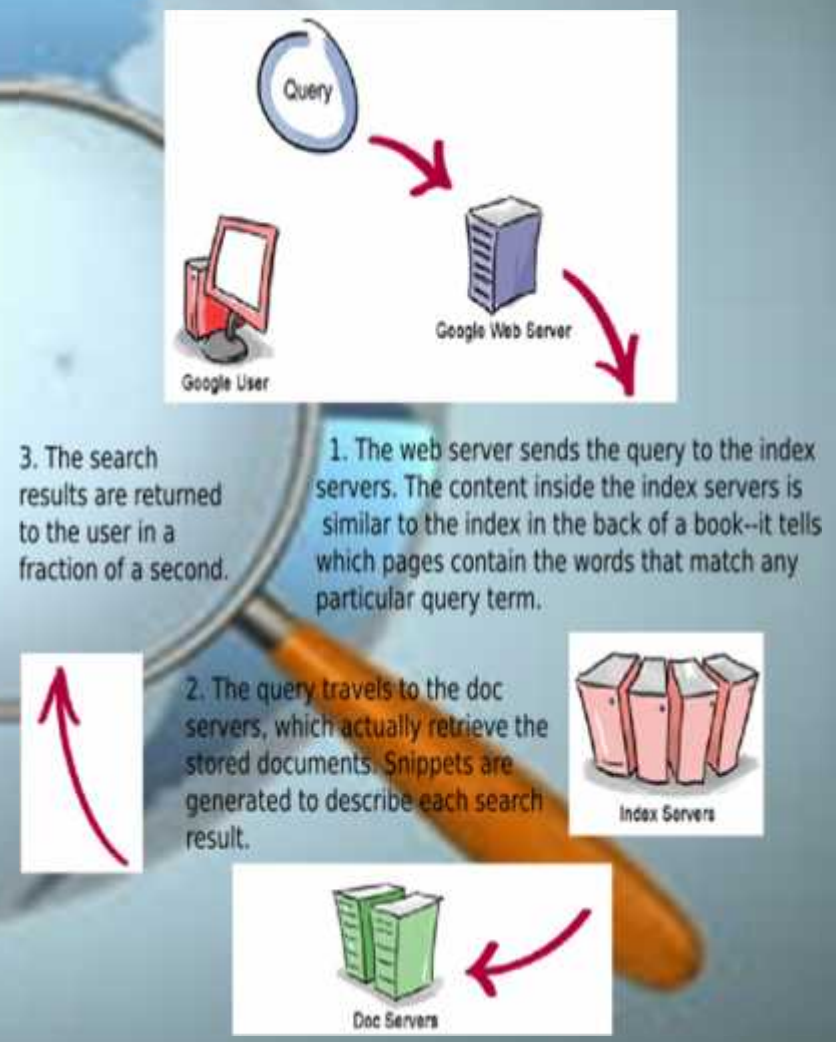
-named after Lawrence Page

Anchor text

- Another feature for high precision
- Association of anchor text with the page and the pointed page
- Non-textual information can be returned as results

Basic architecture

- C or C++ implementation
- Optimized data structures
  - Less cost
  - Avoid disk seek times
- Huge process in between, to return search results



[2]

Performance of Google

Query: bill clinton

http://www.whitehouse.gov/
100.00% ——— (no date) (0K)
http://www.whitehouse.gov/
Office of the President
99.67% ——— (Dec 23 1996) (2K)

http://www.whitehouse.gov/WH/EOP/OP/html/OP_Home.html
Welcome To The White House
99.98% ——— (Nov 09 1997) (5K)
http://www.whitehouse.gov/WH/Welcome.html
Send Electronic Mail to the President
99.86% ——— (Jul 14 1997) (5K)

http://www.whitehouse.gov/WH/Mail/html/Mail_President.html

mailto:president@whitehouse.gov
99.98% ———
mailto:President@whitehouse.gov
99.27% ———
The "Unofficial" Bill Clinton
94.06% ——— (Nov 11 1997) (14K)
http://zpub.com/un/un-bc.html
Bill Clinton Meets The Shrinks
86.27% ——— (Jun 29 1997) (63K)
http://zpub.com/un/un-bc9.html
President Bill Clinton - The Dark Side
97.27% ——— (Nov 10 1997) (15K)
http://www.realchange.org/clinton.htm
$3 Bill Clinton
94.73% ——— (no date) (4K)
http://www.gatewy.net/~tjohnson/clinton1.html

Fig. 2. Sample results from Google.

[1]

In future, they thought to do…

- Aim to index 100 million pages
- As scalable architecture as possible
- Good source for research
- The current identity of Google

Conclusion

- Scalable search engine
- Efficient use of indices
- The ranking function for good quality of search results

References

[1] Brin. S. and Page, L. 1997. The Anatomy of a Large-scale Hypertextual Web Search Engine. In *Proceedings of the 7th International World Wide Web Conference, Vol. 3*. Brisbane, Australia. 107-117.

[2] http://www.googleguide.com/google_works.html

Thank You!