Paper Reading Guide

Information Retrieval

Some background on indexing for information retrieval e.g. see http://en.wikipedia.org/wiki/Search_engine_indexing, may be helpful for understanding this paper.

The Anatomy of a Large-Scale Hypertextual Web Search Engine

by Sergey Brin and Lawrence Page

1. What are the challenges that a search engine must handle to scale to the size of the web?
2. What is the main design goal of Google?
3. Is high precision or high recall preferred by Google? Why?
4. What part of the web does Google exploit in the PageRank ranking system? Does it make intuitive sense?
5. What is the form of the PageRank equation? Note that it is defined in terms of PageRank of other pages, and may eventually depend on itself? Does that make sense, and under what conditions?
6. What is the 'random surfer' interpretation of PageRank?
7. How is anchor text used in Google?
8. What are the other features used in ranking?
9. What are the differences between the Web and typical information retrieval corpuses?
10. What are the main components of Google's architecture?
11. Why does Google avoid disk seeks?
12. How are full HTML documents stored?
13. How are documents found given a URL?
14. What are the considerations in designing the encoding of the hit list?
15. How is the forward index encoded?
16. How is the inverted index encoded? Why is there a need for both a forward and inverted index?
17. What are the main challenges in crawling the web?
18. How does the Google ranking system work?
19. How well does Google work?
20. What are the main contributions of the paper?
21. Can you do similar work?

You may be interested in learning more about other aspects of web information retrieval, e.g. ranking – see The PageRank Citation Ranking: Bringing Order to the Web, by Lawrence Page, Sergey Brin, Rajeev Motwani and Terry Winograd for more on Google's Pagerank.