CPS196.03 Information Management and Mining, Spring 2009, Notes for Web Crawling, Indexing, and Search

Suppose the entire World Wide Web is as shown in Figure 1. There are 5 web pages. Each web page shows the words on the page. Each arrow shows a hyperlink, and the text on the arrow shows the anchor text associated with that link.

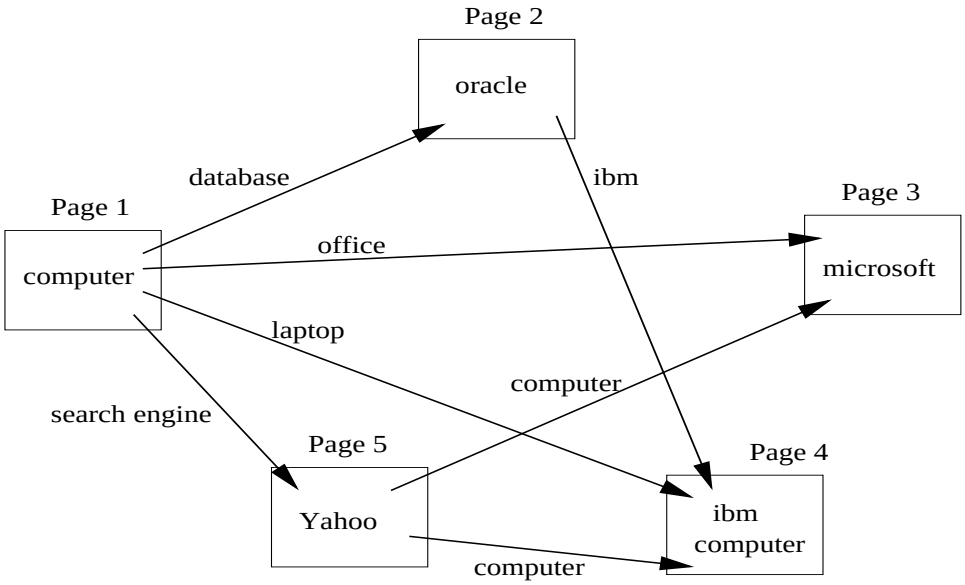

Figure 1: Pages and hyperlinks for example

1 Crawling

Which is a good bootstrap page to start crawling from?
Suppose the crawler starts at Page 1 and downloads it. In terms of the high-level Google architecture (Figure 1 in the Anatomy paper), what all steps are involved in processing the downloaded page?

2 Indexing

Give the hit lists for Pages 1, 2, 3, 4, and 5.
What will the forward index contain for the pages in Figure 1?
What will the inverted index contain for the pages in Figure 1?
Compute the PageRank of the pages in Figures 2 and Figure 3.
Compute the PageRank of Pages 1, 2, 3, 4, and 5.

3 Searching

What will be the lexicon for the Web in Figure 1?

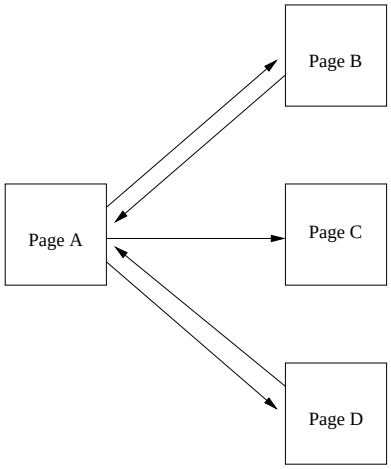

Figure 2: Link graph (1)

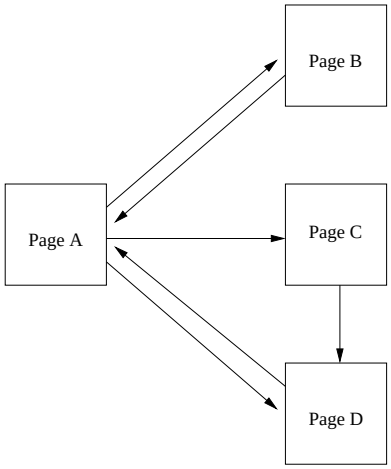

Figure 3: Link graph (2)

Which pages will be returned if we search for "computer" using a first-generation search engine? Which pages will be returned if we search for "computer" using a second-generation search engine? How does Google enable phrase searches? How does Google enable phrase searches based on anchor text?