Quiz 1 sample solution

Azbayar Demberel
Department of Computer Science
asic@cs.duke.edu

February 9, 2007

Question 1

1. An inlink of a web page is a link pointing to that particular web page from other web sites. Page 4 has the most (3) inlinks.

2. An outlink of a web page is a link from that particular web page to other web sites. Page 1 has the most (4) outlinks.

3. Recall that first generation search engines ranked sites based only on their page content. Also you should note that an anchor text appears as a "Text" on the page. Therefore not only pages 1 and 4, which contain texts "computer" and "ibm computer" respectively, but also page 5 which contains anchor text "computer" will be returned as a result.

4. Second generation search engines will associate a link's anchor text with the page pointed to by the link. (Of course, the anchor text will also be associated with the source page of the link.) Hence, a second-generation search engine will return pages 1, 3, 4, and 5 in the result. Note that Page 3 is returned in the result because "computer" is part of the anchor text for the link from Page 5 to Page 3.

Question 2

1. Web page spamming is a technique used by web site developers to trick automated search engines so that their web pages receive a favorable ranking. As spamming makes it hard to maintain an effective search engine, it is bad for search engines.

2. One way of spamming is by inserting phrases multiple times, like "blue devils blue devils blue devils blue devils..." so that next time when people search for "blue devils", your web page will get a high rank because of the number of hits of the search term.

   Another way of spamming is inserting phrases as titles (or with a big font size or capitalization) in invisible color/font, like "**blue devils**"

Question 3

Google uses the lexicon to map words to the location in the inverted index where the document list for the word is stored. Note that the inverted index is stored in the barrels.

Question 4

A search engine with an access to semantic network could determine the synonyms of the search term and retrieve all web pages containting both the actual term and its synonyms. However, those synonyms can have multiple meanings, i.e. polysems, therefore pages containing the polysems of the synonyms, which are totally irrelevant to the search term , can not only be returned but also get high ranks. For example if we search for "laptop", the search engine would determine that "notebook" is its synonym and return results about "Notebook"-the movie or "Notebook"-a book of blank pages, and because the movie "Notebook" was such a hit, web sites regarding that movie will probably have the highest rank and although it is completely irrelevant to our search term "laptop" it will return as the most relevant result.

Question 5

Let's assume that Pepsi was the official sponsor of the Superbowl and it paid a search engine to advertise its web sites. Then a user searches for "football" to get recent football results or maybe to get information about football to write an essay. But because the search engine received money from Pepsi - the sponsor of the biggest "football" game of the year, it will have to give pepsi.com high ranking although its relevance to the search term was little or close to nothing.