# Homework 1 solutions

February 9, 2007

### Question 1

1. Because anchor texts tend to have more accurate descriptions of web pages they point to than the actual web pages, search quality can be improved.

2. Anchor texts help search non-text information as the only way to know that a media file is present in the web page is by looking through its anchor text. For example, let's assume you are searching for "J.J.Reddick's picture". This picture can only be found if it has a correctly describing anchor text.

3. By using anchor texts it is possible to return web pages which have not been actually crawled. A corresponding example can be found at "The Anatomy of a Large-Scale Hypertextual Web Search Engine" paper.

### Question 2

1. **Position**. Words in certain positions tend to have more importance than others. For example words in the title usually have the description of the whole content so they desribe that web page better than any other words in the page.

2. **Capitalization**. People usually capitalize words to grab attention of the reader. Hence, often more important words are capitalized.

3. **Font size**. Words in bigger font sizes also tend to have more importance. Normally people increase the font sizes of titles or sub titles, which usually contain the main description of the web pages.

### Question 3

After crawling web pages, Google stores their full HTML codes in a repository. So when users perform a search, they are able to access the cached HTML codes at the Google repository instead of going to the original sources. It enables the users to access the information even if the original source's servers are temporarily down and in some situations users can have faster access to the information if the original content is stored in a slow server.

### Question 4

1. As Google deals with web pages, it deals with unstructured data. On the other hand Google crawls web pages, parses them and stores them in the barrels and the lexicon, which are structured data. Therefore Google deals with both types of data.

2. Data about students in a university is usually very structured data. For example, this data may contain a record for each student, which in turn contains, e.g., the name of the student, her date of birth, her current year of study, the courses she has done, the grades she got in those courses, and so on. An example query over this data would be to find all students who were born before 1985.

3. The data in typical HTML web pages is unstructured data. The best examples of queries over this data are keyword-based search queries (the same queries that Google has to deal with).

### Question 5

- Yahoo contains only human selected pages, i.e. it maintains lists of pre-calculated query results.

- The main difference between Google and Yahoo is that Yahoo contains human created lists, whereas Google returns automatically calculated results.

### Question 6

Web pages are considered hubs if they point to many good authorities regardless of how many web pages point to them. Authorities, on the other hand, are pages to which many good hubs point to. Because Google's PageRank is calculated based on how many pages point to that particular web page, authorities will have a higher PageRank.

**Question 7**

There could be different answers to this question. However, the justification has to be correct depending how you view it.

1. (iii) Both hub and authority. Wikipedia, itself, has references to, i.e. points to, many good sources about the search term - "compression", which makes it a hub. Many good web pages, on the other hand, reference Wikipedia for explanation about "compression", which makes it an authority.
2. (iv) mail.google.com doesn't point to many different web pages about emails. And the rival emails do not point to it as well. Hence it cannot receive enough scores to be either a hub or an authority.
3. (i) This page points to many quality websites about automobiles, but does not give authoritative information about automobiles. Thus, it is a hub.
4. (ii) Our class website contains authoritative information about the search term, therefore it is an authority.
5. (iv) The web page itself is not Duke football's official web site. It is a site that has all of the Duke athletics news. Therefore, it is valid to argue that this page is neither a hub nor an authority for "duke football".

**Question 8**

There are many differences in Google's and Clever's approaches. Hence there can be many different valid answers to this question. Here, we list only the main three.

1. Google's PageRank ranks websites regardless of the search query, whereas Clever's hubs and authorities ranks the websites based on the search query.

2. Unlike Google, Clever first gets search results from a regular search engine, then processes these results to compute hubs and authorities. Therefore, Clever has more work to do than Google when the query is submitted.

3. In addition to how the hyperlinks point to one another, Google uses other information like text position and font size while computing the final ranking of results for a search query. In contrast, once Clever obtains 200 results for a search query from Altavista, Clever uses only the link structure of pages pointed to by these 200 pages and pages that these 200 pages point to, to compute the good hubs and good authorities for the search query.

**Question 9**

The reason that the author argues that Google is a media company is because Google uses advertisements as its main source of income. With the whole Database of Intentions in its hands, Google is able to determine trends over time, intentions and interests of users and use the information to market itself to commercial websites.

**Question 10**

1. As the Government is now able to obtain personally identifiable information due to the USA Patriot Act, privacy of the users is compromised. If that information gets to the Government, it can track all actions of each person.

2. The behavior of the community of users can be extracted from the "Database of Intentions" and used as a mean of discrimination.