

# Homework 2 sample solution/explanation

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

February 22, 2007

## Question 1

Googlehacking is an attempt to find queries that return exactly one result when searched by google. It has become a popular sport among some computer fanatics.

## Question 2

Noting thatarchie only indexed file titles, to answer this question, one should have taken a look at [www.goduke.com](http://www.goduke.com). Since the title of this document is not “duke football” Archie wouldn’t have returned this result.

## Question 3

There are many measurements that can be used to evaluate a search engine. Here I will list 3, in their order of importance.

- Quality of results. Although it is impossible to meter quality, it is “the” most important performance metric that separates the good from the others.
- Speed. As the connection speed to the Internet increased, so did the user’s demand for quicker response. Many users now consider response time slower than 5 seconds as a signal to leave the page.

- Coverage, i.e. from how much of the web does the search engine search from.

#### **Question 4**

1. One way to reduce the load on web servers is by crawling web pages during idle times of the server, for example at night.
2. Another approach is having a smarter crawling algorithm. One example is, as Gray found out later, making a breadth search that crawls to as many sites as possible before retrieving further contents from a particular site. This way the crawler can spread out the load that it gives to a particular web server. Or another approach can be creating a limit on how many pages the crawler can download from a particular web server in one instance.

#### **Question 5**

1. There was little support from the management. Many of the DEC executives didn't understand the potential a of search engine and the advantage of having it, except its positive publicity. Hence they didn't give enough support to research and improve the search engine.
2. Altavista didn't remain true to its principle. Compaq, which later bough DEC, relying on Altavista's popularity turned it into a web portal to become an "Excite and Yahoo killer", instead of focusing on and performing "quality" searches. However Altavista was recognized to public as a search engine and it ultimately had to loose its battle because it didn't have any space in the portal web market where Yahoo and other web portals were already well established and publicly recognized.
3. During mid 90's, when AltaVista was first created, the search engine market was too immature and advertising models were not yet established. Thus it was not possible "to create a pure play in search that was economically viable". Since AltaVista had little support to do further research (for instance, research for an advertisement model), they always had distractions to do other jobs - like AltaVista powered Internet applications - to get extra revenue and prove its worthiness.

### **Question 6**

1. AltaVista ran on a powerful parallel Alpha computers, which had 64 bit memory capability, which allows to quickly process data.
2. It was able to crawl almost the entire web and process the web pages quickly.

### **Question 7**

Excite was able to group search results and instead of showing one ranked list, return multiple grouped rank lists of different concepts. For example if a search on “jaguar” was performed, Excite would return 2 sublists: one containing the highest ranked pages about the car, the other about the animal.

### **Question 8**

First, Garfield’s method stimulates citation inflation and log-rolling. Instead of citing only the most relevant papers, people get motivated to cite many papers with the hope that they will get cited in return.

Second, Garfield’s method doesn’t take into account the weight of the papers that are citing it. Although bad papers will never get cited by a high quality paper, it might get high ranking because it was cited by many other “low class” papers. Clearly a paper cited by Einstein, should weigh higher than a paper cited by a novice and Garfield’s method might result in the other paper having a higher rank.

### **Question 9**

- Overloading web servers. In contrast to a human, crawler retrieves each and every content of that web page. In addition it doesn’t have any think time in between and causes sudden burst of overload to the server.
- Privacy. Some people simply didn’t want to be indexed and they were sceptical of google’s unauthorizd access to their web.
- Copyright issues. As the crawler consumes the entire web page, some companies thought google was trying to steal their data

- Disagreement on ranking. Some renowned web sites were ranked lower because they weren't connected to many other sites.

### **Question 10**

As we noted on Question 9, google had several problems in its early days. To solve these problems google could have had a self policing guideline for crawling. For example in order not to tangle with copyright issues, it could have created a list of museums and avoid crawling them.

### **Question 11**

Search engine optimization is a technique used by web designers to attract more qualified customers by understanding the concept of search algorithms. They can accomplish it, for example, by making the web page more crawler friendly, writing attention grabbing titles and interesting descriptions, or using unique words and phrases that are easily indexed. Search engine optimization is similar to spamming in a sense that they both design the web page so that it will receive higher number of users. However, the big difference is that in web spamming the goal is to manipulate the search engine to receive higher ranking, whereas optimization tries to adjust to it.

### **Question 12**

1. Web pages consist of numerous types of documents: pictures, videos, plain text and so on. Even the plain texts in an HTML file can have multiple formats and meanings. An effective search engine should be able to deal with all types of documents so that it can return the best result.
2.
  - “Alt text” and “Meta” tags. By using these tags google can find non-text documents like images.
  - Support for parsing non-text documents. Google can return text results not only in html or text page, but also in word documents, pdf and power point presentations.
  - Anchor text. By paying special attention to anchor texts, google can find non-text documents that don't have an appropriate “alt text” tag, or from a page it hasn't crawled.

### Question 13

Human language is very rich, and it would be very hard (currently not even possible), for a machine to completely understand human language. In that sense manual indexers have an advantage over machines and we have seen several examples regarding it in our textbook (like search on cars and automobiles, sneakers and tennis shoes, and so on) that show the weakness of automated indexers. Humans are, however, error prone and are subjective. And although they can pick up on the nuances, their knowledge is limited or simply they can forget. Hence, in some situations a machine can outperform humans in language processing. Therefore, the best solution might be a combination of these two approaches. Currently modern search engines are starting to implement language processing capabilities like stemming or recognize synonyms and polysems. However, much research is still needed in this area.

### Question 14

Words can have multiple meanings and many words can have the same meaning. So although documents contain different words, i.e. they are not lexically connected, they can have similar or even the same content.

For example New Yorkers have a special dialect and some of their lexicon is completely different than the general English. Some New Yorkers may say that they made a mistake “on accident” as opposed to “on purpose”<sup>1</sup>. And if a lawyer born in New York searches for “Perjury on purpose”, he may enter “Lie on accident”. A search engine in this case will have an immense difficulty trying to connect these 2 terms.

### Question 15

- Pros of having smaller data: Less disk space, faster = have more time to process other contents.
- Cons: Less information = less accuracy, requires powerful and more effective indexer

### Question 16

---

<sup>1</sup>[en.wikipedia.org/wiki/New\\_York\\_Dialect](http://en.wikipedia.org/wiki/New_York_Dialect)

Metathesaurus is a thesaurus of medical terms. Having a thesaurus in search engines would be useful, especially for a dedicated search engine for a specific topic or field. However, it would be extremely complicated to build one for a big scale search engine.

### **Question 17**

Stemming is the process of removing suffixes or prefixes from a word to reduce it to its root form. Stemming gives both advantages and disadvantages to search engines. Here are some of the advantages:

1. By using stemming all the words that have the same root will be stored as their root form. Thus it will save the index space significantly. Hence, search speed can be improved as well.
2. Stemming can help the user if she misspells a word. For example if she entered "hippopotameus", the search engine can determine that she misspelled the word, and return results for "hippopotamus"

Some of its disadvantages are:

1. Stemming could cause to return irrelevant results, as some words which have the same root can have different meanings when used in different fields.
2. It is not intuitive how to perform stemming. "Stemming can be a tedious undertaking, especially considering that decisions must be made and rules developed for thousands of words in the English Language".
3. Search engine could have trouble processing, i.e. parsing, the query because it will also have to stem the search terms. Hence the performance of the processing speed can be downgraded.

### **Question 18**

Recall that stop words are words that have little or no value as a search term. Some common guidelines to create stop lists would be:

- Articles and prepositions tend to not have any impact on search results

- Any word that doesn't add any meaning to the context can be included in stop list. Some examples of these words are: become, be, clearly, enough.
- Process the database and count the occurrences of words in sentences. If a word exists in more than some percentage of the sentences, it could be included in the stop list.
- Some people also think that singletons, or words that appear very rarely, should be included in the stop list.

### Question 19

Contiguous phrases are combinations of words that appear together to express a new meaning. If a user searches for a contiguous phrase, for example "big mac", it will clearly make no sense to return results which contain the words separately (or not next to each other). However, to search for these phrases, the search engine needs additional information like proximity of the words and will need to create extra storage space for each word in each document (2 bytes for 5 billion documents that contain approximately 1000 words! I will leave rest of the math to you). Later whenever a user performs search, the search engine must in addition to finding the phrase, calculate the proximity of the words. Therefore in systems that have strict storage and computational requirements, it might be better to avoid storing proximity. In a search engines, however, the advantages will clearly outweigh the negative impact as the quality of search is a higher requirement.

### Question 20

1. Parse the query: Separate words "duke", "freshman", "seminar"
2. Using the lexicon, convert words into wordID: "duke"="103512", "freshman"="150352", "seminar"="190325". (Note that these are not the real values) and find in which barrel the words are located.
3. Using the short inverted index, go to the start of the barrels for each word
4. Scan through the barrel until there is a document that matches all the search terms, e.g "<http://www.aas.duke.edu/trinity/49s/>"

5. Using the page rank, compute the rank of that document for the query.
6. If more than 40,000 results have been produced go to step 9
7. If we are in the short barrels and at the end of any doclist, seek to the start of the doclist in the full barrel for every word and go to step 4.
8. If we are not at the end of any doclist go to step 4.
9. Sort the documents that have matched by rank and return the top k