Midterm Exam

406.424 Internet Applications

Spring, 2010

Date: April 22, 2010

YOUR NAME:

1. (40 points) True / False Questions

1-1. PageRank is a query dependent feature used for Google ranking.

1-2. Unlike the hyperlink, the link direction is formed from an old post to a new post in the trackback mechanism.

1-3. HTTP maintains a state to facilitate communication between web browser and server.

1-4. A valid URL must start with string "http://".

1-5. Every response message under HTTP must include entity body.

1-6. The URL, http://imlab.snu.ac.kr, entered to a web browser must be resolved into an IP address before the web browser contacts a web server.

1-7. Cookies are usually stored at web server.

1-8. Web spiders are mobile agents in a sense that they move around on the web to discover new web content.

1-9. Multi-threaded web server architecture provides better scalability than the process-driven architecture.

1-10. When the web servers are replicated, the requests made by browsers need to be dispatched to the servers by L4 switch instead of L7.

1-11. Structured blogging made creation of a new blog post a lot easier.

1-12. The XML file for an RSS feed may contain more than one item.

1-13. Digg.com is a social bookmarking site.

1-14. The content of a widget can be dynamically generated based on the content of a page that contains the widget.

1-15. AJAX is a technology for supporting active document.

1-16. Ranking is one of important components in the indexing process.

1-17. Transformation from "Fishing" to "Fish" is a job of the stopper.

1-18. PageRank is widely used in Boolean retrieval.

1-19. UTF-32 uses a variable length encoding.

1-20. Even if two documents have the same checksum value, they may actually contain different content.

2. (10 points) Consider a web proxy that implements a cache-consistency protocol. Assume that the time cached, the current time, and the time the resource was last modified are 500, 700, and 200, respectively.

2-1. Consider the following equation for the squid web proxy.

$$T_{expire} = \alpha * (T_{cached} - T_{last\_modified}) + T_{cached}$$

Assuming that α is 2.0, how many times will the resource be downloaded to the proxy if there are 10 successive requests during the time interval [700, 720]?

2-2. Now, suppose that the pull-based protocol is used for maintaining cache-consistency. Assuming that the resource has been modified at time 690, how many times will the resource be downloaded to the proxy if there are 10 successive requests during the time interval [700, 720]?

3. (15 points) Suppose that the total number of documents you have is 1,000,000. You would like to estimate how many documents have both the terms "LiveK" and "Blog". After randomly scanning 1,000 documents, you found out that the number of documents containing "LiveK" is 100 while the number of documents containing "Blog" is 500. Given the current information, what is the best estimation you can make? Justify your answer.

4. (15 points) Consider a document that has the following text:

"2PM Again And Again And Again And Again"

Assuming that 4 bit hash values for "2PM", "Again", and "And" are "1000", "0110", and "0001", respectively, compute the 4 bit simhash fingerprint by using the term frequency as a word weight.

5. (10 points) Consider a corpus with 10,000 documents and 1,000,000 words. Under the Zip's law, what is the estimated number of words occurring 10 times in the corpus?

6. (10 points) Your IR system has a collection of 100 documents of which 50 are relevant to the query "2NE1". Suppose that your IR system returned 30 documents for your query "2NE1" but only 10 documents were relevant.

6-1. What is the precision of your IR system?

6-2. What is the recall of your IR system?