NJIT New Jersey's Science & Technology University

A. V. Gerbessiotis

CS 485-101

Fall 2011

Exam 1(250 points)

120 minutes

Name: .....

ID Number: ..... **Exam Number:** .....

Grade: 1: ... 2: ... 3: ... 4: ... 5: ... 6: ... 7: ... 8: ... Total: .....

Solve **ALL the problems** in the space provided

Read the Problems CAREFULLY!

There are 7 (SEVEN) pages this page included

Read and Sign the statement below at the end of the exam

Unsigned exams will be marked with grade 0 (zero).

STATEMENT

On my honor, I pledge that I have not violated the provision of the NJIT Student Honor Code.

Sign below at the end of the exam

Signature .....

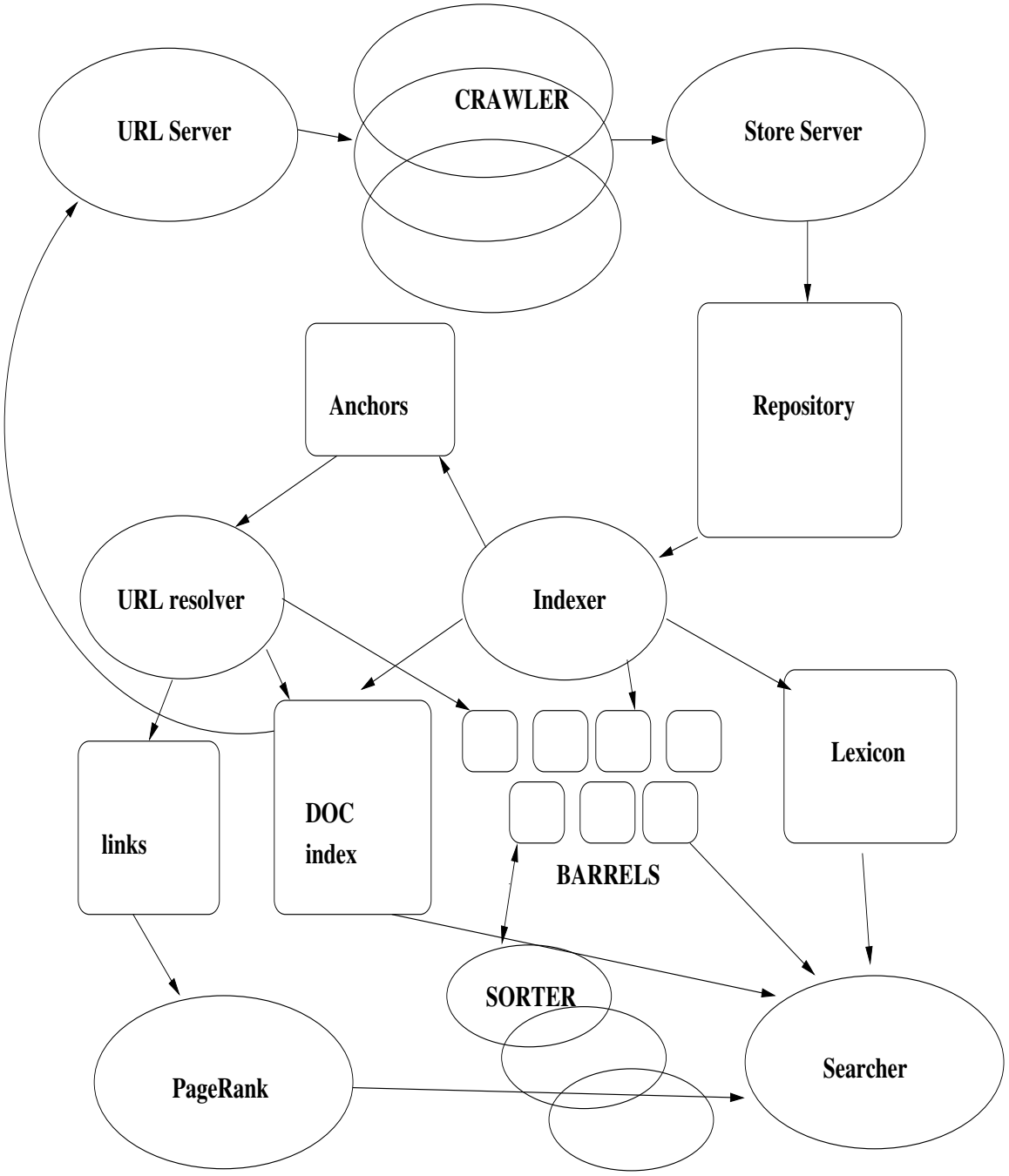

Figure 1: Google System Architecture.

Problem 1. (70 POINTS)
Give short answers to the following questions. (References related to Google are to the Google System Architecture of 1997/1998.)

(1) Does `Anchors` store information in the form of `URLs` or `docIDs`?

(2) Where are web-pages stored in Google?

(3) Does Google store web-pages in compressed or uncompressed format?

(4) Name (do not further discuss) 5 data structures used by Google.

(5) Give a data structure in Google that stores only `docIDs` and neither URLs nor URL checksums.

(6) What is the input for the `PageRank` computations?

(7) Why does the `Searcher` access the `DOCIndex`?

(8) Where does Google store the forward index of `anchor text`?

(9) What do we call the input provided to the `SORTER` and what do we call the output generated by the `SORTER`?

(10) Name an operation that can map two different `keywords` into the same `index term`.

Problem 2. (20 POINTS)
(a) What is the difference between a `keyword` and an `index term`? Give an example, if it helps.

(b) What is the difference between a `word` and a `keyword` in the context of information retrieval? Explain by providing an example.

Problem 3. (20 POINTS)
(a) Give an example of a bit-aligned code, and an example of a byte-aligned code.

(b) Give an example of a lossless compression method, and an example of a lossy one.

Problem 4. (20 POINTS)
For a specific query involving term t, the corpus (collection of documents) includes the twenty documents $d_1, \ldots, d_{20}$. The first five $d_1, \ldots, d_5$ contain index term t and are thus relevant, whereas the remaining fifteen $d_6, \ldots, d_{20}$ do not contain t at all.
As a response to the query the search engine returned documents

$$d_1, d_3, d_8, d_4, d_7, d_9, d_{10}, d_{11}, d_{12}, d_{13}$$

in that specific order.
(a) What is the precision of the search query? Explain.
(b) What is the recall of the search query? Explain.

Problem 5. (32 POINTS)
Which one of the 4 marks `AND`, `OR` , `NOT`, and `-` are operators in (i) Google, and (ii) Bing?

Problem 6. (28 POINTS)
Give the unary code for 0, and binary, Elias-γ, Elias-δ codes for 30.

Problem 7. (30 POINTS)
You have a machine with 1000MBytes of free memory. We want to use all this memory to store the `docID`, `URLhash` and `URL`s of a search engine, similarly to the structure used by the lexicon to store `wordID`s and `words` in which the latter are stored in consecutive memory location of an array separated by NULLs; in our case, URLs will be separated by NULLs. The Operating System consumes 70MB and other user programs including the one dealing with the table another 30MB. The average length of a `URL` in the search engine is around 99 characters ASCII. The hash table will store a 16-byte `URLhash`, a 5-byte `docID`, and a pointer/index p that points to the location in the table where the strings for the URLs are stored delimited by ASCII NULLs. Organize the hash table and the table of URLs so that you can extract from them maximum performance. How many URLs can be accommodated? What is the length m of the hash table? How many bytes for p? How much space in MBytes will you scheme consume? Explain.

Turn page

Problem 8. (30 POINTS)
We typed the following queries in Google and Bing and recorded the following number of hits in millions for the two search engines. Answer the following questions with respect to the Table below.

No	Query Term(s)	GOOGLE (number of hits in '000,000)	BING (number of hits in '000,000)
1	⟨a b or c⟩	2530	2680
2	⟨a b OR c⟩	22100	5150
3	⟨A B OR C⟩	22100	5150
4	⟨a (b OR c)⟩	25270	6870
5	⟨(a b) OR c)⟩	22100	5150
6	⟨a or b⟩	3230	3090
7	⟨a OR b⟩	25270	11500
8	⟨b or a⟩	3270	73
9	⟨b OR a⟩	25270	11500

(a) Why the difference between query 1 and query 2 ?

(b) Why the similarity between query 2 and query 3 ?

(c) Why the similarity between query 2, query 3, and query 5 ?

(d) What's the difference between query 6 and query 7?

(e) Why the similarity between query 7, and query 9 ?

Turn page

Intentionally left blank. You can use it for scratch paper by tearing it off carefully; do not return it then.

End of Exam 1