Information Storage and Retrieval Exam 3 Name: _____

1. What is the goal of providing a collection-level interface to users? Why is it needed?
2. Identify a likely difficulty with providing users with a list of collections.
3. What type of collection overview does MedLine, the ACM digital library, and Yahoo! provide?
4. How does Scatter-Gather work? What information is provided to users?
5. Describe the Themescapes collection overview.
6. Identify three forms of querying interface.
7. Describe faceted queries.
8. How are Venn diagrams limiting when used for query expression?
9. Describe the Filter Flow interface for queries.
10. What are Magic Lenses? How can they be used as queries?
11. What are four methods providing users with an understanding of returned documents prior to their viewing those documents?
12. What is a document surrogate? Give two examples.
13. How does Popout Prism indicate a document's relation to the query? How does TileBars indicate a document's relation to the query? How are these different?
14. How does InfoCrystal indicate returned document's relation to query?
15. Identify two systems that provide categories for search results?
16. How does Cha-Cha provide an understanding of relations between returned documents?
17. How do table views provide an understanding of relationships between returned documents?
18. What characteristics of the Web make it a difficult environment (collection) for IR purposes?
19. Describe the basic search engine architecture.
20. Explain the concept of hubs and authorities on the Web.
21. How can Web crawling be distributed over several servers?
22. How can Web site administrators specify what content should not be crawled?
23. Identify four potential crawling goals.
24. Describe two forms of politeness guarantees that can be provided by crawlers.

25. Why might human-generated directories be better than full-text search engines for IR on the Web?
26. Identify two advantages and disadvantages of metasearch engines when compared to search engines.
27. What was the first Web search engine?
28. What were Google's three design goals?
29. How many bits did Google use to represent each "hit" in 1997?
30. How does Google support the indexing of non-textual Web pages (e.g. jpgs, gifs, etc.)?
31. Provide a high-level overview of Google's PageRank algorithm.
32. Why are subgraphs a problem for PageRank?
33. Why are dangling links a problem for PageRank?
34. Name four factors used by Google's search engine to rank results.
35. When is PageRank most useful? Why?
36. How does PageRank provide an estimation of WebTraffic?
37. Why did Google need to develop the Google File System?
38. What assumptions drive the design of the Google File System?
39. How does the Google File System deal with hardware failure?
40. What types of metadata are (semi-)automatically assigned in the INFOMINE system?