Republic of the Philippines

**SOUTHERN LEYTE STATE UNIVERSITY-BONTOC**

San Ramon, Bontoc, Southern Leyte

Website: www.slsuonline.edu.ph

Telefax: (053) 382-3121; Email: slsubc@yahoo.com

**ANSWER SHEET**

**Advanced Database**

*First Semester AY: 2020-2021*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name:** | **Jeca G. Calope** | | | **Rating:** |  |
|  |  | |  | |  |
| **Course/Year/Section:** | | **BSIT-3-A** | **Date Submitted:** | |  |

**MODULE 3**

**PRE-TEST**

1. What is structured data and what is unstructured data? Give an example of each from your experience.

**Answer:**

Unstructured data can be thought of as data that’s not actively managed in a transactional system; for example, data that doesn’t live in a relational database management system (RDBMS). Structured data can be thought of as records (or transactions) in a database environment; for example, rows in a table of a SQL database. There is no preference as to whether data is structured or unstructured. Both have tools that allow users to access information. Unstructured data just happens to be in greater abundance than structured data is.

Examples of unstructured data are:

* [Rich media.](https://www.netapp.com/data-storage/unstructured-data/rich-media-content-management) Media and entertainment data, surveillance data, geo-spatial data, audio, weather data
* Document collections. Invoices, records, emails, productivity applications
* [Internet of Things (IoT).](https://www.netapp.com/data-storage/unstructured-data/internet-of-things-iot) Sensor data, ticker data
* Analytics. [Machine learning](https://www.netapp.com/artificial-intelligence/what-is-machine-learning), [artificial intelligence (AI)](https://www.netapp.com/artificial-intelligence/what-is-artificial-intelligence)

Until the advent of [object-based storage](https://www.netapp.com/artificial-intelligence/what-is-machine-learning), most, if not all, of this unstructured data was stored in file-based systems.

2. Give a general definition of information retrieval (IR). What does information retrieval involve when we consider information on the Web?

**Answer:**

Information Retrieval (IR), has been part of the world, in some form or other, since the advent of written communications more than five thousand years ago. IR has as its domain the collection, representation, indexing, storage, location, and retrieval of information-bearing objects. Traditionally these objects have been text-based documents such as articles, reports, and books; however, as multimedia computing has progressed, the list of information-bearing objects has grown to include such things as images, videos, maps, sound, and music recordings. In the modern sense of the term, IR has its roots in the scientific information explosion that accompanied, and followed, [World War II](https://www.encyclopedia.com/history/modern-europe/wars-and-battles/world-war-ii). Predating the computer, early modern IR systems used ingenious manual mechanisms to deal with the millions of scientific and technological research papers that were being written as part of the war against tyranny. Automated information retrieval systems are used to reduce what has been called [information overload](https://en.wikipedia.org/wiki/Information_overload). An IR system is a software system that provides access to books, journals and other documents; stores and manages those documents. [Web search engines](https://en.wikipedia.org/wiki/Web_search_engine) are the most visible [IR applications](https://en.wikipedia.org/wiki/Information_retrieval_applications).

3. Discuss the types of data and the types of users in today’s information retrieval systems.

**Answer:**

4. What is meant by navigational, informational, and transformational search?

**Answer:**

**Navigational search** - refers to finding a particular piece of information (such as the University’s Web site) that a user needs quickly. **Informational search** - to find current information about a topic. **Transactional search** - to reach a site where further interaction happens resulting in some transactional event (such as joining a social network, shopping for products, making online reservations, accessing databases, and so on).

5. What are the two main modes of interaction with an IR system? Describe and provide examples.

**Answer:**

Retrieval is concerned with the extraction of relevant information from a repository of documents through an IR query, while browsing signifies the activity of a user visiting or navigating through similar or related documents based on the user’s assessment of relevance. During browsing, a user’s information need may not be defined a priori and is flexible.

Consider the following browsing scenario: A user specifies ‘Atlanta’ as a keyword. The information retrieval system retrieves links to relevant result documents containing various aspects of Atlanta for the user. The user comes across the term ‘Georgia Tech’ in one of the returned documents, and uses some access technique (such as clicking on the phrase ‘Georgia Tech’ in a document, which has a built-in link) and visits documents about Georgia Tech in the same or a different Website (repository). There the user finds an entry for ‘Athletics’ that leads the user to information about various athletic programs at Georgia Tech. Eventually, the user ends his search at the Fall schedule for the Yellow Jackets foot-ball team, which he finds to be of great interest. This user activity is known as browsing.

6. Explain the main differences between the database and IR systems.

**Answer:**

The main reason for this difference is that informationretrieval usually deals with natural language text which is not always well structured and could be semantically ambiguous. On the other hand, a data retrieval system (such as a relational database) deals with data that has a well-defined structure and semantics.

7. Describe the main components of the IR system.

**Answer:**

1. Indexing system – indexing and searching methods and procedures (an indexing system can be human or automated);

2. Collection of documents – text, image or multimedia documents, or document surrogates (for example bibliographical records);

3. Defined set of queries – which are input into the system, with or without the involvement of a human searcher; and

4. Evaluation criteria – specified measures by which each system is evaluated, for example ‘precision’ and ‘recall’ as measures of relevance. Recall is the proportion of relevant documents in the collection retrieved in response to the query. Precision is the proportion of relevant documents amongst the set of documents retrieved in response to the query.

8. What are digital libraries? What types of data are typically found in them?

9. Name some digital libraries that you have accessed. What do they contain and how far back does the data go?

10. Give a brief history of IR and mention the landmark developments in this field.

11. What is the Boolean model of IR? What are its limitations?

12. What is the vector space model of IR? How does a vector get constructed to represent a document?

13. Define the TF-IDF scheme of determining the weight of a keyword in a document. Why is it necessary to include IDF in the weight of a term?

14. What are probabilistic and semantic models of IR?

15. Define recall and precision in IR systems.

16. How is an F-score defined as a metric of information retrieval? In what way does it account for both precision and recall?

17. What are the different types of queries in an IR system? Describe each with an example.

18. What are the approaches to processing phrase and proximity queries?

19. Describe the detailed IR process

20. What is stopword removal and stemming? Why are these processes necessary for better information retrieval?

21. What is a thesaurus? How is it beneficial to IR?

22. What is information extraction? What are the different types of information extraction from structured text?

23. What are vocabularies in IR systems? What role do they play in the indexing of documents?

24. Describe the process of constructing the result of a search request using an inverted index.

25. Define relevance feedback.

26. Describe the three types of Web analyses.

27. What are the three categories of agent-based Web content analyses mentioned in this chapter?

28. What is the database-based approach to analyzing Web content? What are Web query systems?

29. What algorithms are popular in ranking or determining the importance of Web pages?

Which algorithm was proposed by the founders of Google?

30. What can you learn from Web usage analysis? What data does it generate?

31. What mining operations are commonly performed on Web usage data? Give an