**Assignment questions**

**Unit 3**

1.Describe Document and Term clustering

2.Describe the Concept of statistical Indexing

3. How the **Term similarity matrix** is represented in the context of term clustering? **3 conecpts**

4.Illustrate the features related to cognitions and perception

5.Describe the structure of Classes of Automatic Indexing. **4classes**

**Unit 4**

1.Summarize the features of Item Clustering. (**types, features, matrics forms)**

2.How the search statements are used in the context of Information Retrieval Systems? **8 functions**

3.Describe the Similarity Measures in terms of Search statements.

**2 measures**

**4.**Explain in detail about various ranking algorithms.

5.What are the features of Information Visualization Technologies? **5technologies**

**Unit 5**

1.What are the Objectives of **Text Search Algorithms** in Information Retrieval Systems?

2.Describe Hardware Text Search Techniques.

3.Discuss the features of Information System Evaluation.**5features 2 measures, 3 techniques, diagrams**

4.Describe about TREC results?

5.Describe different **Brute force, knuth-Morris and Boyer-Moore a algorithm** in the context of Text Searching.

**OBJECTIVE TYPE BITS**

1. \_\_\_\_\_ is the process of analyzing an item to extract the information to be kept permanently in an index.

a. Class Indexing b. Automatic Indexing c. Manual Indexing d. Any Ans: b

2. \_\_\_\_\_\_\_\_ is used mostly in commercial systems.

a. Statistical b. Natural Language c. Concept d. Hypertext Linkages Ans: a

3. \_\_\_\_\_\_\_\_ indexing stores the information that are used in calculating a probability that a particular item satisfies a particular query.

a. Probalistic b. Bayesian c .vector space d. neural net. Ans: a

1. \_\_\_\_\_\_\_\_ approaches store information used in generating a relative confedence level of an item relevance to a query.
   1. Bayesian b. vector space c. both d. none Ans: c

5. \_\_\_\_\_\_ are dynamic learning structures that are discussed under concept indexing where they are used to determine concept classes.

* 1. Probalistic b. Bayesian c .vector space d. neural net. Ans: d

6. \_\_\_\_\_\_ indexing uses words with in an item to correlate to concepts discussed in the item.

a. Statistical b. Natural Language c. Concept d. Hypertext Linkages Ans: c

7. \_\_\_\_\_\_ approach ia based upon direct application of the theory of probability to IRS.

a. Probabilistic b. Natural Language c. Concept d. Hypertext Linkages Ans: a

8. \_\_\_\_\_\_\_ produces the efficient results when data is retrieving from multiple databases.

a. Probabilistic b. Natural Language c. Concept d. Hypertext Linkages Ans: a

9. \_\_\_\_\_\_\_\_\_ processing is used to semantic information in addition to statistical information to enhance the indexing of the item.

a. Probabilistic b. Natural Language c. Concept d. Hypertext Linkages Ans: b

10. Tagged Text Parser structure allows for identification of potential term phrases based upon \_\_\_\_\_ identification.

a. verb b. noun c. adjective d. all Ans: b

11. \_\_\_\_\_\_ processing will use DR-LINK System.

a. Probabilistic b. Natural Language c. Concept d. Hypertext Linkages Ans: b

12. \_\_\_\_\_ system attempts to introduce a higher level of abstraction indexing on top of the statistical processes.

a. Probabilistic b. Natural Language c. Concept d. Hypertext Linkages Ans: b

13. \_\_\_\_\_\_ indexing is a statistical technique whose goal is to determine a canonical representation of concept.

a. Probabilistic b. Natural Language c. Concept d. Hypertext Linkages Ans:c

14. \_\_\_\_\_\_\_\_ techniques have very powerful representation.

a. Binary b. Vector c. Both c. None Ans : b

15. \_\_\_\_\_\_\_\_\_ pages at each Internet site are indexed automatically.

a. Automatically generated b. manually generated c. Crawlers d. All Ans: a

16. In \_\_\_\_\_\_\_\_\_\_\_\_ users define search terms, and it goes to various sites searching for the desired information.

a. Automatically generated b. manually generated c. Crawlers d. All Ans: c

17. \_\_\_\_\_ is the example for WebCrawler’s

a. WebCrawler’s b. Open Text c. Path Finder d. All Ans: d

**Fill in the blanks**

* 1. **Term Frequency TFij is** the frequency of occurrence of a term *Ti* in a document *Dj*.
  2. **Total Term Frequency TTFi** is the frequency of occurrence of a term *Ti* in the entire collection.
  3. **Document Frequency DFi** is the number of unique documents in the collection that contain a term *Ti*..
  4. Tagged Text Parser structure allows for identification of potential term phrases based upon **Noun** identification.
  5. **Automatic Indexing** is the process of analyzing an item to extract the information to be kept permanently in an index.
  6. **Manually generated (e.g. Yahoo!)** pages are indexed manually into a linked hierarchy(an “index”). Users browse in the hierarchy by following links.
  7. **Automatically generated (e.g. Alta Vista)** pages at each Internet site are indexed automatically (creating a “searchable data structure”).
  8. **Automatically generated** structures are used for querying, rather than browsing.
  9. **Crawlers (e.g. WebCrawler)** No a *priori* indexing.
  10. **Crawlers (e.g. WebCrawler)** Users define search terms, and the crawler goes to various sites searching for the desired information.
  11. **Hypertext Linkages** Provides virtual threads of concepts between items versus directly defining the concepts with in an item.
  12. The **SMART** systemuses Vector Model.

**Objective Questions**

1. \_\_\_\_\_\_\_\_ provide a grouping of similar objects into a class

a. Grouping b. Clustering c. Mapping d. Characterizing Ans: b

2. \_\_\_\_\_\_\_\_ allows linkage between clusters to be specified.

a. Grouping b. Clustering c. Mapping d. Characterizing Ans: b

3. \_\_\_\_\_\_\_\_\_Increase recall by expanding searches with related terms.

a. Term Clustering b. Document Clustering c. Both d. None Ans: a

4.\_\_\_\_\_\_ is used to create document clusters.

a. Term Clustering b. Document Clustering c. Both d. None Ans: b

5. Increase recall by expanding searches with related terms

a. Term Clustering b. Document Clustering c. Both d. None Ans: a

6. The similarity between every term pair is calculated as a basis for determining the\_\_\_\_\_

a. Grouping b. Clustering c. Mapping d. Characterizing Ans: b

7. \_\_\_\_\_\_\_ model is used for thesaurus generation.

a. Binary b. Vector c. Static d. Dynamic Ans: b

8. Cliques, single link, stars, and connected components algorithms are used for\_\_\_\_

a. Clustring Classes b. Applying Stemming c. Data Structure d. All Ans: a

9. \_\_\_\_\_\_\_ Ward’s method typically took the longest to implement Similarity Measure between Clusters.

a. Single link clustering b. complete linkage c. Group average d. Ward’s method Ans: d

10. Minimum computation on the order of One Pass Assignment is \_\_\_\_\_

a. O (n) b. O (log (n)) c. O (n2) d. O (nlog (n)) Ans: a

11\_\_\_\_\_\_ displays the keywords followed by their context.

a. KWOC B. KWIC C. KWAC D.ALL Ans: c

12. \_\_\_\_\_\_ displays the possible terms in its phrase context.

a. KWOC B. KWIC C. KWAC D.ALL Ans: b

13. \_\_\_\_\_\_ character is used in KWIC to indicate the end of the phrase.

a. “” b. / c. @ d. % Ans: b

14. Which one is the Software Text Search Algorithm?

a. Knuth-Morris-Pratt b. Boyer Moore c. Rabin-Karp d. All Ans:d

15. \_\_\_\_\_\_\_\_\_technique is used to complete a query by searching for query terms that could not be satisfied by the index.

a. Text Search b. Term Search c. Index Search d. All Ans:a

16. The evaluation criteria of Information Retrieval System is based on

a. Effectiveness b. Efficiency c. Usability d. All Ans:d

1. **Word Coordination Approach** Specifies if phrases as well as individual terms are to be clustered.

2. **KWOC** is another name for concordance.

3. **KWIC** AND **KWAC** is useful to determine the meaning of homographs.

4. WKOC **Key Word Out of Context**

5. KWIC **Key Word In Context**

6. KWAC **Key Word And Context**

7. **Vector Model** is used for Automatic Term Clustering

8. **User** and **System** perspectives are used to made Measurements

9. **Fallout** is another measure that is directly related to retrieving non-relevant items can be used in defining how effective an information system is operating.

10. TREC stands for **Text Retrieval Evaluation Conference**