PUNE VIDYARTHI GRIHA'S

COLLEGE OF SCIENCE AND TECHNOLOGY

Affiliated to University of Mumbai

Question Bank

Class: T.Y.B. Sc.CS **Semester: VI**

Subject: Information Retrieval

Unit I

1. Information retrieval is querying of _____ textual data.
a) structured
b) unstructured
c) formatted
d) None

2. _____are indexed units in the incidence matrix.
a) Terms
b) Collection
c) Information
d) Data

3. The number of documents in the collection that contain a term t is called as ____
a) Document Index di_t
b) Document frequency df_t
c) Document Inverse din_t
d) Document Incidence Matrix dim_t

4. The standard way of quantifying the similarity between two documents d1 and d2 is to compute the _____of their vector representations.
a) sine similarity
b) cot similarity
c) cosine similarity
d) None

5. PM stands for _____
a) Cost per migrating
b) Cost per making

c) Cost per manage
d) Cost per mil

6. _____fraction of tChe returned results are relevant to the information need.
a) Proximity
b) Posting Merge
c) Precision
d) Posting list

7. A dictionary of terms is sometime also referred as_____
a) Corpus
b) Collection
c) Lexicon
d) none of the above

8. A model of information retrieval in which we can pose any query in which search terms are combined with the operators AND, OR, and NOT
a) Ad Hoc Retrieval
b) Ranked Retrieval Model
c) Boolean Information Model
d) D Proximity Query Model

9. The number of times that a word or term occurs in a document is called the:
a) Term Frequency
b) Vocabulary Lexicon
c) Proximity Operator
d) D Indexing Granularity

10. Which of the following is a technique for context sensitive spelling correction?
a) the Jaccard Coefficient
b) Soundex algorithms
c) k-gram indexes
d) Levenshtein distance

11. _____ is the fraction of the relevant documents in the collection returned by the system.
a) Reconnect
b) Recall
c) Reciprocal
d) Retrieved

12. The task of the _____ is to group words that are derived from a common stem.
a) Parser
b) Stopper
c) Transformer
d) Stemmer

13. A model of information retrieval in which we can pose any query in which search terms are combined with operators AND, OR and NOT
a) Ad Hoc retrieval

b) Ranked Retrieval model
c) Boolean Information model
d) Proximity Query model

14. A web server communicates with a client (browser) using which protocol:
a) HTML
b) HTTP
c) FTP
d) Telnet

15. The purpose of the inverse document frequency is to increase the weight of terms with high collection frequency
a) True
b) False

16. The basic operation of a web browser is to pass a request to the web server. This request is an address for a web page and is known as the
a) UAL: Universal Address Locator
b) HTML: Hypertext Markup Language
c) URL: Universal Resource Locator
d) HTTP: Hypertext transfer protocol

17. Information retrieval systems have much in common with
a)Filing systems
b)Transaction systems
c)Database systems
d)Management systems

18.A deadlock can be broken down by
a)Committing one or more transactions
b)Aborting one or more transactions
c)Rolling back one or more transactions
d)Terminating one or more transactions

19. Querying of unstructured textual data is referred to as
a)Information access
b)Information updation
c)Information manipulation
d)Information retrieval

20. Information is
a)Data
b)Processed Data
c)Manipulated input
d)Computer output

21. Online transaction processing is used because
a)disk is used for storing files
b)it is efficient
c)it can handle random queries.

d)Transactions occur in batches

22. The quality of information which is based on understanding user needs
    a)Complete
    b)Trustworthy
    c)Relevant
    d)None of the above

23.The primary storage medium for storing archival data is
    a)floppy disk
    b)magnetic disk
    c)magnetic tape
    d)CD- ROM

24.Organizations have hierarchical structures because
    a)it is convenient to do so
    b)it is done by every organization
    c)specific responsibilities can be assigned for each level
    d)it provides opportunities for promotions

25.Operational information is
    a)Haphazard
    b)Well organized
    c)Unstructured
    d)Partly structured

26. A computer based information system is needed because
(i) The size of organization have become large and data is massive
(ii) Timely decisions are to be taken based on available data
(iii) Computers are available
(iv) Difficult to get clerks to process data

    a)(ii) and (iii)
    b)(i) and (ii)
    c)(i) and (iv)
    d)(iii) and (iv)

27. Operational information is needed for
    a)Day to day operations
    b)Meet government requirements
    c)Long range planning
    d)Short range planning

28. Data by itself is not useful unless
    a)It is massive
    b)It is processed to obtain information
    c)It is collected from diverse sources
    d)It is properly stated

29. For taking decisions data must be
a)Very accurate
b)Massive
c)Processed correctly
d)Collected from diverse sources

30. Measures of Similarity are as Follows :
i. The lengths of the Documents.
ii. The number of terms in common.
iii. Whether the terms are common or unusual.
iv. How many times each term appears.

a)i) & ii)
b)ii) & iii)
c)iii) & iv)
d)i),ii),iii) & iv)

31. one of the application of Personalized Search is,
a)Google
b)Yahoo
c)IBM
d)Alpha Search Engine
32. Structure of Web has following entities:
i. Web Graph
ii. Static and Dynamic Pages
iii. Hidden web pages
iv. Size of web page

a)i) & ii)
b)i) & ii)
c)iii) & iv)
d)i),ii),iii) & iv)

33. Goal of IR is to find documents _____to an information need from a large document set
a) applicable
b) Relevant
c) knowledgeable
d) none of the above

34. Static collection of documents is known as _____.
a) Information
b) corpus
c) data
d) None of the above

35. Information retrieval systems have much in common with_____.

a) Filing systems
b) Transaction system

c) Database systems
d) Management systems

36. A data structure that maps terms back to the parts of a document in which they occur is called as___.
a) Postings list
b) Incidence Matrix
c) Dictionary
d) Inverted Index

37. The number of times that a word or term occurs in a document is called the:
a) Proximity Operator
b) Vocabulary Lexicon
c) Term Frequency
d) Indexing Granularity

38. for ad hoc information ret_____ is/are the test collectionsrieval system evaluation.
a) Cranfield
b) TREC
c) only a
d) both a and b

39. Every web page is assigned _____score(s).
a) 1
b) 2
c) 4
d) 3

40. _____ maintains the file system tree and the metadata for all the files and directories present in the system.
a) Namenode
b) Datanode
c) Mapper
d) Tracker

41. A query such as mon* is known as a _____
a) trailing wildcard query
b) leading wildcard query
c) both a and b
d) Mixed wildcard query

42. Precision (P) is the fraction of _____
a) P(retrieved/relevant)
b) P(relevant/retrieved)
c) P(relevant/retrieved)
d) P(retrieved/true)

43. ____ is not the Basic Ranking Models of information retrieval system.
a) Boolean Retreival
b) Vector Space model

c) Probabilistic model
d) Data model

44. ____ is the number of documents contains the term.
a) Term
b) Df
c) Idf
d) Inverse df

45. In information retrieval, extremely common words which would appear to be of little value in helping select documents that are excluded from the index vocabulary are called:
a) Stop Words
b) Tokens
c) Lemmatized Words
d) Stemmed Terms

46. Boolean retrieval model does not provide provision for
a) Ranked search
b) Proximity search
c) Phrase search
d) Both proximity and ranked search

47. Variable size postings lists is used when
a) More seek time is desired and the corpus is dynamic
b) Less seek time is desired and the corpus is dynamic
c) Less seek time is desired and the corpus is static
d) More seek time is desired and the corpus is dynamic

48. Structured data allows for
a) Does not depend on data complexity
b) Less complex queries
c) No relationship
d) More complex queries

49. Boolean queries often result in
a) Too many or too few results
b) None of the above
c) Too few results
d) Too many results

50. Ranked retrieval models take as input
a) None of the above
b) Boolean queries
c) Logical queries
d) Free text queries

UNIT II

1. Given two strings s1 and s2, the edit distance between them is sometimes known as the:

a) Levenshtein distance
b) isolated-term distance
c) porter stemmer algorithm

2. _____nodes that can be reached from the giant SCC but cannot reach it.

a) In
b) Out
c) Gcc
d) in-out

3. Postings list should be sorted by:

a) Document Frequency
b) DocID
c) TermID
d) Term frequency

4. TREC stands for_____

a) a.Text Retrieval Conference
b) b. Text Retrieval Context
c) c.Text Retrieval Congestion
d) d.All the above

5. CLEF stands for_____

a) Cross Language Evaluation Forum
b) Cross lingual evaluating field
c) Cross Language Evaluating Field

d) Cross Language Evaluating Forum

6. A good_____page for a topic links to many authority pages for that topic.
   a) Crawler
   b) SEO
   c) Web
   d) Hub

7. A large repository of documents in IR is called as
   a) Corpus
   b) Database
   c) Dictionary
   d) Collection

8. Term document incidence matrix is
   a) Sparse
   b) Depends upon the data
   c) Dense
   d) Cannot predict

9. Document frequency of a term is the
   a) Number of documents that contain the term
   b) None of the above
   c) Number of times the term appears in the document
   d) Number of times the term appears in the collection

10. What is contiguity hypothesis in vector space classification
   a) Documents from different classes dont overlap
   b) Documents in the same class form a contiguous region of space
   c) All of the above.
   d) Intra cluster similarity is higher than inter-cluster similarity

11. Strategic information is needed for
   a) Day to day operations
   b) Meet government requirements
   c) Long range planning
   d) Short range planning

12. Strategic information is required by
   a) Middle managers
   b) Line managers
   c) Top managers
   d) All workers

13. Tactical information is needed for
   a) Day to day operations
   b) Meet government requirements
   c) Long range planning
   d) Short range planning

14. The Search tool CANNOT be used on which major Access object
    a) Forms
    b) Queries
    c) Reports
    d) Tables

15. The _____ is a wild card that represents one or more characters
    a) question mark
    b) asterisk
    c) exclamation mark
    d) dollar sign

16. Which is not an option for Filter on a text field
    a) Begins With
    b) Between
    c) Contains
    d) End With

17. The Search tool is best used when searching for which kind of data.
    a) simple
    b) multiple
    c) unique
    d) formatted

18. _____ filtering recommends products which are similar to the ones that a user has liked in the past.
    a) Collaborative based
    b) Context based
    c) Collection based
    d) Content based

19. _____is a way of measuring the importance of website pages.
    a) A Querying
    b) B Page Rank
    c) C Link Analysis
    d) D HITS

20. Given two strings s1 and s2, the edit distance between them is sometimes known as the
    a) A Levenshtein distance
    b) B isolated-term distance
    c) C k-gram overlap
    d) D Jaccard Coefficient

21. Hadoop is a framework that works with a variety of related tools. Common cohorts include _____
    a) A MapReduce, Hive and HBase
    b) B MapReduce, MySQL and Google Apps
    c) C MapReduce, Hummer and Iguana
    d) D MapReduce, Heron and Trumpet

22. _____ can best be described as a programming model used to develop Hadoop-based applications that can process massive amounts of data.

a) A MapReduce
b) B Mahout
c) C Oozie
d) D All of the mentioned

23. Collaborative Filtering has following problems

a) Cold Start
b)Scalability
c)Sparsity
d)All of the above

24. Input, Purpose and Output are the factors of _____ .

a)Summarization
b)Question Answering
c)Page Rank
d)Personalized Search

25. Which one of the following is not Test Collection and Evaluation Series

a)Text Retrieval Conference (TREC)
b)NII Test Collections for IR Systems (NTCIR)
c)Cross Language Evaluation Forum(CLEF)
d)Collaborative Filtering

26. A _____ node acts as the Slave and is responsible for executing a Task assigned to it by the JobTracker.

a) MapReduce
b) Mapper
c) TaskTracker
d) JobTracker

27. _____ part of the MapReduce is responsible for processing one or more chunks of data and producing the output results.

a) Maptask
b) Mapper
c) Task execution
d) All of the mentioned

28. _____ function is responsible for consolidating the results produced by each of the Map() functions/tasks.

a) Reduce
b) Map
c) Reducer
d) All of the mentioned

29. Although the Hadoop framework is implemented in Java, MapReduce applications need not be written in _____

a) Java
b) C

c) C#
d) None of the mentioned

30_____ is a utility which allows users to create and run jobs with any executables as the mapper and/or the reducer.
a) Hadoop Strdata
b) Hadoop Streaming
c) Hadoop Stream
d) None of the mentioned.

31. Which of the following algorithm is used by Google to determine the importance of a particular page?
a) SVD
b) PageRank
c) FastMap
d) All of the mentioned

32. Based on PageRank algorithm, Google returns _____ for a query that is parsed for its keywords.
a) SEP
b) SAP
c) SERP
d) Business Objects Build

33. Hadoop is a framework that works with a variety of related tools. Common cohorts include _____
a) MapReduce, Hive and HBase
b) MapReduce, MySQL and Google Apps
c) MapReduce, Hummer and Iguana
d) MapReduce, Heron and Trumpet

34._____ builds systems automatically answer questions posted by humans in a natural language.
a) Personalized Search
b) Question Answering
c) Specialized Search
d) Summarization

35._____ means to find a subset of data which contains the information of the entire set.

a) Personalized Search
b) Question Answering
c) Specialized Search
d) Summarization

36. _____ is a page that contains actual information on a topic.
a) A authority
b) B Hub

c) C Hyperlinks
d) D Image

UNIT III

1. SEO stands for _____
a) A Search engine order
b) B Search engine organizer
c) C Search engine option
d) D Search engine optimization

2. An advantage of a positional index is that it reduces the asymptotic complexity of a postings intersection operation.
a) True
b) False

3. A data structure that maps terms back to the parts of a document in which they occur is called an
a) Postings list
b) Incidence Matrix
c) Dictionary
d) Inverted Index

4. Who is Crawler ?
a)Web Spider
b)Pen Tester
c)Hacker
d)Ethical Hacker

5. The basic formula for paid placement is_____

a) Pay-per-click ($) = Advertising cost ($) ÷ Ads clicked (#)
b) Pay-per-click ($) = Advertising cost ($) * Ads clicked (#)
c) Pay-per-click ($) = Advertising cost ($) * Ads clicked (#)
d) Both a and b

6. The first special index for general wild card queries is the_____.
a) k-term index
b) Permuterm index
c) B-tree
d) Hashes

7. The _____ for finding terms based on a query consisting of k-grams.
a) Document index
b) k-gram index
c) Inverted index
d) Term index

8. _____ mainly encodes numerical and non-text attribute-value data.
a) Data centric XML
b) text centric XML
c) both a and b
d) User centric XML

9. Permuterm indexes are used for solving
a) Spelling Checking
b) b. Boolean queries
c) c. Phrase queries
d) d. Wildcard queries

10. Each node of the tree is an XML element and is written with an _____
a) Opening tag
b) closing tag
c) both a and b
d) only a

11. The _____consists of a dictionary of terms.
a) Bi-gram index
b) K-gram index
c) Inverted index
d) Incidence index

12. _____includes link building, increasing link popularity by submitting open directories, search engines, link exchange, etc.
a) Off Page SEO
b) In Page SEO
c) Middle Page SEO
d) Both a nd b

13.Best implementation approach for dynamic indexing is
a) Periodic re indexing

b) Using Invalidation bit vector for deleted docs
c) None
d) Using logarithmic merge

14. Search engines use a of n _____ to automatically index sites
a) crawler
b) query
c) enterprise
d) sitebuilder

15. A search value can be an exact value or it can be
a) Logical operator
b) Relationship
c) Wild card character
d) Comparison operation

16. Major portion of web page contributes _____
a) image
b) text
c) video
d) audio

17. A piece of icon or image on a web page associated with another webpage is called _____
a) url
b) hyperlink
c) plugin
d) extension

18. Which on-page element carries the most weight for SEO?

a) The meta keywords tag
b) The title tag
c) The headers (H1, H2, H3, etc)
d) Other

19. What do the acronyms PA, DA, and PR stand for?

1. Personal authority, domain authority, parked rename
2. Page authority, domain age, page rank
3. Page authority, domain authority, page rank
4. Page authority, domain act, page range

20. The number of characters recommended for Title Tag?

a) 120
b) 250
c) 70
d) 100

21.Which query will give the list of web pages indexed by a particular search engine on given domain

1. list:http://www.websitename.com
2. link:http://www.websitename.com
3. webpage:http://www.websitename.com
4. site:http://www.websitename.com

22. If a website's search engines get saturated with respect to a particular search engine by 20%, what is it exactly?

a) 20% of the web pages of the website have been indexed by the search engine
b) 20% of the website's pages will never be indexed
c) Only 20% of the pages of the website will be indexed by the search engine
d) The website ranks in the first 20% of all websites indexed by the search engine for its most important search terms

23.What is anchor text?

a) It is the main body of text on a particular web page
b) The text within the left or top panel of a web page
c) It is the visible text that is hyperlinked to another page
d) It is the most prominent text on the page that the search engines use to assign a title to the page

24.What is the term for Optimization strategies that are in an unknown area of reputability?

a) Blue hat techniques
b) Orange hat techniques
c) Grey hat techniques
d) Shady hat techniques

25. Which of the following search engines offers a list of the top 50 most searched keywords?

1. AOL
2. Yahoo
3. Google
4. Lycos

26.How much time period is required to get a Google page ranking?

1. 2 week
2. 1 week
3. 2 months
4. More than 3 months

27. Which of the following is not a XML storage option?
a) Native storage as XML data type
b) Mapping between XML and relational storage
c) Small object storage
d) None of the Mentione

28. In which of the following scenario, using XML native storage would be inappropriate?
a) Fixed schema
b) You want to query and modify your XML data
c) You want to index the XML data for faster query processing
d) Your application needs system catalogue views to administer your XML data and XML schema

29. Which of the following part of the XML data stored in an XML column is very important for locking?
a) Granularity
b) Degree of Structure
c) Hierarchy
d) None of the mentioned

30.Search engine optimization is the process of _____ of a website or a web page in a search engine's search results.

None of these
Getting Meta Tags
Affecting the visibility
Generating Cached Files

31.Select the incorrect statement. The behaviour of a Web crawler is the outcome of a combination of policies:

a) A selection policy that states which pages to download,

b) A re-visit policy that states when to check for changes to the pages,
c) A politeness policy that states how to overload Web sites,
d) A parallelization policy that states how to coordinate distributed web crawlers.

32. The search results are generally presented in a line of results often referred to as _____.

a) Tag List
b) Search Engine Results Pages
c) Search Engine Pages
d) Category List

33. Arrange the search engines by their year of development.
1. Bing
2. Yahoo
3. Ask
4. Google

a) 2 4 1 3
b) 2 1 3 4
c) 2 1 4 3
d) 2 4 3 1

34. Web search engines stores information about many web pages by a _____.

a) Web Indexer
b) Web Organizer
c) Web Router
d) Web Crawler

35. Web Crawler is also called as _____.

a) Link Directory
b) Search Optimizer
c) Web Spider
d) Web Manager