CS6007-INFORMATION RETRIEVAL

Two Marks Question with Answers

Unit-I

UNIT I – INTRODUCTION

1 Define information retrieval.

Information Retrieval is finding material of an unstructured nature that satisfies an information need from within large collections.

2 Explain difference between data retrieval and information retrieval.

Parameters	Data Retrieval	Information retrieval
Example	Data Base Query	WWW Search
Matching	Exact	Partial Match, Best Match
Inference	Deduction	Induction
Model	Deterministic	Probabilistic

3. List and explain components of IR block diagram.

Input – Store Only a representation of the document

A document representative – Could be list of extracted words considered to be significant.

Processor – Involve in performance of actual retrieval function

Feedback – Improve

Output – A set document numbers.

4. What is objective term and nonobjective term?

Objective Terms – Are extrinsic to semantic content, and there is generally no disagreement about how to assign them.

Nonobjective Terms – Are intended to reflect the information manifested in the document, and there is no agreement about the choice or degree of applicability of these terms.

5. Explain the type of natural language technology used in information retrieval.

Two types

I. Natural language interface make the task of communicating with the information source easier, allowing a system to respond to a range of inputs.

II. Natural Language text processing allows a system to scan the source texts, either to retrieve particular information or to derive knowledge structures that may be used in accessing information from the texts.

6. What is search engine?

A search engine is a document retrieval system design to help find information stored in a computer system, such as on the WWW. The search engine allows one to ask for content meeting specific criteria and retrieves a list of items that match those criteria.

7. What is conflation?

Stemming is the process for reducing inflected words to their stem, base or root form, generally a written word form. The process of stemming if often called conflation.

8. What is an invisible web?

Many dynamically generated sites are not index able by search engines; This phenomenon is known as the invisible web.

9. Define Zipf's law.

An empirical rule that describes the frequency of the text words. It state that the i^{th} most frequent word appears as many times as the most frequent one divided by i^{α}, for some $\alpha>1$.

10. What is open source software?

Open source software is software whose source code is available for modification or enhancement by anyone.

"Source code" is the part of software that most computer users don't ever see; it's the code computer programmers can manipulate to change how a piece of software—a "program" or "application"—works.

11. What is proprietary software?

Proprietary software is computer software which is the legal property of one party. The term of use for other parties is defined by contracts or licensing agreements. These terms may include various privileges to share, alter , dissemble, and use the software and its code.

12. What is closed software?

Closed software is a term for software whose license does not allow for the release or distribution of the software's source code. Generally it means only the binaries of a computer program are distributed and the license provides no access to the programs source code. The source code of such programs is usually regarded as a

trade secret of the company. Access to source code by third parties commonly requires the party to sign a non-disclosure agreement.

13. List the advantage of open source.

The right to use the software in any way.
There is usually no license cost and free of cost. The
source code is open and can be modified freely. Open
standards.
It provides higher flexibility.

14. List the disadvantage of open source.

There is no guarantee that development will happen.
It is sometimes difficult to know that a project exist, and its current status.
No secured follow-up development strategy.

15. What are the reasons for selecting open software?

Development and maintenance of open source software is a community based activity.
Open source software licenses are copyright protected they strictly ensure the user freedom to use, modify and distribute the programs.
Is interoperable customizable according to the needs and fulfills the software industry standards.
Open source software allows everyone to use, study, modify and distribute the software.
Allows a broader perspective when comes to its support.

16. What do you mean by Apache License?

The Apache License is a free software license written by the Apache Software Foundation (ASF). The name Apache is a registered trademark and may only be used with the trademark holders express permission.
Apache license is a high performance, Full-featured text search engine library
written entirely in Java.

17. Explain features of GPL version2.

It gives permission to copy and distribute the programs unmodified source code. It allows modifying the programs source code and distributing the modified source code.

User distributes compiled versions of the program, both modified and unmodified.

All modified copies are distributed under the GPL v2.

All compiled versions of the program are accompanied by the relevant source code.

UNIT II – INFORMATION RETRIEVAL

1. What do you mean information retrieval models?

A retrieval model can be a description of either the computational process or the human process of retrieval: The process of choosing documents for retrieval; the process by which information needs are first articulated and then refined.

2. What is cosine similarity?

This metric is frequently used when trying to determine similarity between two documents. Since there are more words that are in common between two documents, it is useless to use the other methods of calculating similarities.

3. What is language model based IR?

A language model is a probabilistic mechanism for generating text. Language models estimate the probability distribution of various natural language phenomena.

4. Define unigram language.

A unigram (1-gram) language model makes the strong independence assumption that words are generated independently from a multinomial distribution

5. What are the characteristics of relevance feedback?

It shields the user from the details of the query reformulation process.

It breaks down the whole searching task into a sequence of small steps which are easier to grasp.

Provide a controlled process designed to emphasize some terms and de-emphasize others.

6. What are the assumptions of vector space model?

Assumption of vector space model:

The degree of matching can be used to rank-order documents;

This rank-ordering corresponds to how well a document satisfying a users information needs.

7. What are the disadvantages of Boolean model?

It is not simple to translate an information need into a Boolean expression
Exact matching may lead to retrieval of too many documents.

The retrieved documents are not ranked.

The model does not use term weights.

8. Define term frequency.

Term frequency: Frequency of occurrence of query keyword in document.

9. Explain Luhn's ideas

Luhn's basic idea to use various properties of texts, including statistical ones, was critical in opening handling of input by computers for IR. Automatic input joined the already automated output.

10. Define stemming.

Conflation algorithms are used in information retrieval systems for matching the morphological variants of terms for efficient indexing and faster retrieval operations. The Conflation process can be done either manually or automatically. The automatic conflation operation is also called stemming.

11. What is Recall?

Recall is the ratio of the number of relevant documents retrieved to the total number of relevant documents retrieved.

12. What is precision?

Precision is the ratio of the number of relevant documents retrieved to the total number of documents retrieved.

13. Explain Latent semantic Indexing.

Latent Semantic Indexing is a technique that projects queries and documents into a space with "latent" Semantic dimensions. It is statistical method for automatic indexing and retrieval that attempts to solve the major problems of the current technology. It is intended to uncover latent semantic structure in the data that is hidden. It creates a semantic space where in terms and documents that are associated are placed near one another.

UNIT III – WEB SEARCH ENGINE – INTRODUCTION AND CRAWLING

1. Define web server.

Web server is a computer connected to the internet that runs a program that takes responsibility for storing, retrieving and distributing some of the web files.

2. What is web Browsers?

A web browser is a program. Web browser is used to communicate with web servers on the Internet, Which enables it to download and display the web pages. Netscape Navigator and Microsoft Internet Explorer are the most popular browser software's available in market.

3. Explain paid submission of search service.

In paid submission user submit website for review by a search service for a preset fee with the expectation that the site will be accepted and include d in that company's search engine, provided it meets the stated guidelines for submission. Yahoo! is the major search engine that accepts this type of submission. While paid submissions guarantee a timely review of the submitted site and notice of acceptance or rejection, you're not guaranteed inclusion or a particular placement order in the listings.

4. Explain paid inclusion programs of search services.

Paid inclusion programs allow you to submit your website for guaranteed inclusion in a search engines database of listings for a set period of time. While paid inclusion guarantees indexing of submitted pages or sites in a search database, you're not guaranteed that the pages will rank well for particular queries.

5. Explain in pay-for-placement of search services.

In pay-for-placement, you can guarantee a ranking in a search listing for the terms of your choice. Also known as paid placement, paid listing, or sponsored listings, this program guarantees placement in search results. The leaders in pay-for-placement are Google, Yahoo! and Bing.

6. Define Search Engine Optimization.

Search Engine Optimization is the act of modifying a website to increase its ranking in organic, crawler-based listing of search engines. There are several ways to increase the visibility of your website through the major search engines on the internet today. The two most common forms of internet marketing paid placement and natural placement.

7. Describe benefit of SEO.

Increase your search engine visibility

Generate more traffic from the major search engines.

Make sure your website and business get NOTICED and VISITED.

Grow your client base and increase business revenue.

8. Explain the difference between SEO and Pay-per-click

SEO	Pay-Per-click
SEO results take 2 weeks to 4 months	It results in 1-2 days
It is very difficult to control flow of traffic	It has ability to turn on and at any moment
Requires ongoing learning and experience to reap results	Easier for a novice
It is more difficult to target local markets	Ability to target “local” markets
Better for long-term and lower margin campaigns	Better for short-term and high-margin campaigns.
Generally more cost-effective , does not penalize for more traffic	Generally more costly per visitor and per conversion

9. What is web crawler?

A web crawler is a program which browses the world web in a methodical, automated manner. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to p[provide fast searches.

10. Define focused crawler.

A focused crawler or topical crawler is a web crawler that attempts to download only pages that are relevant to a pre-defined topic or set of topic.

11. What is hard and soft focused crawling?

In **hard focused crawling** the classifier is invoked on a newly crawled document in a standard manner. When it returns the best matching category path, the out-neighbors of the page are checked into the database if and only if some node on the best matching category path is marked as good.

12. What is the Near-duplicate detection?

Near-duplicate is the task of identifying documents with almost identical content. Near- duplicate web documents are abundant. Two such documents differ from each other in a very small portion that displays advertisements, for example. Such differences are irrelevant and for web search.

13. What are requirements of XML information retrieval systems?

Query language that allows users to specify the nature of relevant components, in particular with respect to their structure.

Representation strategies providing a description not only of the content of XML documents, but also their structure.

Ranking strategies that determine the most relevant elements and rank these appropriately for a given query.

UNIT IV – WEB SEARCH – LINK ANALYSIS AND SPECIALIZED SEARCH

1. What is link analysis?

The goal of information retrieval is to find all documents relevance for a user query in a collection of documents. With the advent of the web new source of information became available, one of them being the hyperlink between documents and records of user behavior. Collections of documents connected by hyperlinks. Hyperlinks provide a valuable source of information for web information retrieval. This area of information retrieval is commonly link analysis.

2. What is in query independent ranking?

In query-independent ranking a score is assigned to each page without a specific user query with the goal of measuring the intrinsic quality of a page. At query time this score is used with or without some query-dependent criteria to rank all documents matching the query.

3. What is query dependent ranking?

In query-dependent ranking a score measuring the quality and the relevance of a page to a given user query is assigned to some of the pages.

4. Define authorities?

Authorities are pages that are recognized as providing significant, trustworthy and useful information on a topic. In-degree is one simple measure of authority. However in-degree treats all links as equal.

5. Define hubs.

Hubs are index pages that provide lots of useful links to relevant content pages. Hub pages for IR are included in the home page.

6. What is Hadoop?

At Goggle MapReduce operation are run on a special file system called Google File System that is highly optimized for this purpose. GFS is not open source. Doug Cutting and Yahoo! reverse engineered the GFS and called it Hadoop Distributed File System. The software framework that supports HDFS, MapReduce and other related entities is called the project Hadoop or simply Hadoop.

7. What are the Hadoop Distributed File System?

The Hadoop Distributed File System is designed to store very large data sets reliably, and to stream those data sets at high bandwidth to user application. HDFS stores file

system metadata and application data separately. The HDFS namespace is a hierarchy of files and directories. Files and directories are represented on the NameNode by inodes, Which record attributes like permissions, modification and access times, namespace and disk space quatas.

.

8. Define MapReduce.

MapReduce is a programming model and software framework first developed by Google. Intended to facilitate and simplify the processing of vast amounts of data in parallel on large clusters of commodity hardware in a reliable, fault-tolerant manner.

9. List the characteristics of MapReduce?

Very large scale data: peta, exa bytes
Write once and read many data. It allows for parallelism without mutexes Map and Reduce are the main operations: Simple code
All the map should be completed before reduce operation starts.
Map and reduce operations are typically performed by the same physical processor. Number of map tasks and reduce tasks are configurable.

10. What are the limitation of Hadoop/Map Reduce?

Cannot control the order in which the maps or reductions are run.
For maximum parallelism, you need Maps and Reduces to not depend on data generated in the same Map Reduce job.
A database with an index will always be faster than a Map Reduce job on unindexed data.
Reduce operations do not take place until all Maps are complete.
General assumption that the output of Reduce is smaller than the input to Map large data source used to generate smaller final values.

11. What is Cross-Lingual Retrieval?

Cross – Lingual Retrieval refers to the retrieval of documents that are in a language different from the one in which the query is expressed. This allows users to search document collections in multiple language and retrieve relevant information in a form that is useful to them, even when they have little or no linguistic competence in the target languages.

12. Define Snippets.

Snippets are short fragments of text extracted from the document content or its metadata. They may be static or query based. In static snippet, it always shows the first 50 words of the document, or the content of its description metadata, or a description taken from a directory site such as dmoz.org.

13. List advantages of invisible web content.

Specialized content focus – large amounts of information focused on an exact subject. Contains information than might not be available on the visible web.
Allows a user to find a precise answer to a specific question Allow a user to find WebPages from a specific date or time.

14. What is collaborative filtering?

Collaborative filtering is a method of making automatic predictions about the interests of a single user by collecting preferences or taste information from many users. It uses given rating data by many users for many items as the basic for predicting missing ratings and/or for creating a top-N recommendation list for a given user, called the active user.

15. What do you mean by item-based collaborative filtering?

Item-based CF is a model-based approach which produces recommendations based on the relationship between items inferred from the rating matrix. The assumption behind this approach is that users will prefer items that are similar to other items they like.

16. What are problem of user based CF?

The two main problems of user-based CF are that the whole user database has to be kept in memory and that expensive similarity computation between the active user and all other users in the database has to be performed.

17. Define user based collaborative Filtering.

User-based collaborative filtering algorithms work off the premise that if a user(A) has a similar profile to another user (B), then A is more likely to prefer things that B prefers when compared with a user chosen at random.

UNIT V – DOCUMENT TEXT MINING

1. What do you mean by information filtering?

An information filtering system is a system that removes redundant or unwanted information from an information stream using (semi)automated or computerized methods prior to presentation overload and increment of the semantic signal-to-noise ratio.

2. What are the characteristics of information filtering?

- □ Filtering system involve large amounts of data.
- □ Information filtering systems deal with textual information.
- □ It is applicable for unstructured or semi-structured data.

3. Explain difference between information filtering and information Retrieval.

Information Filter	Information Retrieval
IF is concerned with the removal of textual information from an incoming stream and its dissemination to groups or individuals.	IR systems are concerned with the collection and organization of texts so that users can then easily find a text in the collection.
Information filtering is concerned with repeated uses of the system by users with long-term, but changing interests and needs.	A query represents a one-time information need.
Filtering is based on descriptions of individual or group interests or needs that are usually called profiles.	Retrieval of information is instead based on user specified information needs in the form of a query.
IF systems deal with dynamic data.	IR systems deal with static databases.

4. What is text mining?

- □ Text mining is understood as a process of automatically extracting meaningful, useful, previously unknown and ultimately comprehensible information from textual document repositories.
- □ Text mining can be visualized as consisting of two phases: Text refining

that transforms free-form text documents into a chosen intermediate form, and knowledge distillation that deduces patterns or knowledge from the intermediate form.

5. What is classification?

Classification is a technique used to predict group membership for data instances. For example, you may wish to use classification to predict whether the weather on a particular day will be "sunny", "rainy" or "cloudy".

6. Explain clustering.

Clustering is a process of partitioning a set of data in a set of meaningful subclasses. Every data in the subclass shares a common trait. It helps a user to understand the natural grouping or structure in a data set.

7. What are the desirable properties of a clustering algorithm?

- □ Scalability
- □ Ability to deal with different data types
- □ Minimal requirements for domain knowledge to determine input parameters
- □ Interpretability and usability

8. What is decision tree?

- □ A decision tree is a simple representation for classifying examples. Decision tree learning is one of the most successful techniques for supervised classification learning. A decision tree or a classification tree is a tree in which each internal node is labeled with an input features.
- □ The arcs coming from a node labeled with a feature are labeled with each of the possible values of the feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes.

9. List the advantages of decision tree.

- □ Decision tree can handle both nominal and numeric input attributes.
- □ Decision tree representation is rich enough to represent any discrete value classifier.
- □ Decision trees are3 capable of handling database that may have errors.
- □ Decision trees are capable of handling datasets that may have missing values.
- □ It is self-explanatory and when compacted they are also easy to follow.

10. List the disadvantages of decision tree

- □ Most of the algorithms require that the target attribute will have only discrete values.
- □ Most decision-tree algorithms only examine a single field at a time.
- □ Decision trees are prone to errors in classification problems with much class.
- □ As decision tree use the "divide and conquer" method, they tend to perform well if a few highly relevant attribute exists, but less so if many complex interactions are present.

11. What is supervised learning?

In supervised learning, both the inputs and the outputs are provided. The network then processes the inputs and compares its resulting outputs against the desired outputs. Errors are then propagated back through the system, causing the system to adjust the weights which control the network.

12. What is unsupervised learning?

In an unsupervised learning, the network adapts purely in response to its inputs. Such networks can learn to pick out structure in their input.

13. What is dendrogram?

Decompose data objects into a several levels of nested partitioning called a dendrogram. A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster.