Republic of the Philippines

**SOUTHERN LEYTE STATE UNIVERSITY-BONTOC CAMPUS**

San Ramon, Bontoc, Southern Leyte

Telefax Nos. (053) 382-3121 / (053)382 260

Email:slsu@yahoo.com

**IT 301: Advanced Database Systems**

**Name: Jemaline T. Temario Year & Section: 3-B**

**Instructor: Mr. Rexal Toledo Date: 12-18-20**

***MODULE 3 PRE-TEST***

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

* For the most part, structured data refers to information with a high degree of organization, such that inclusion in a relational database is seamless and readily searchable by simple, straightforward search engine algorithms or other search operations; whereas unstructured data is essentially the opposite.

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

* Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text (or other content-based) indexing.

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

* **Types of Users**. The user may be an*expert user*(for example, a curator or alibrarian), who is searching for specific information that is clear in his/her mind and forms relevant queries for the task, or a *layperson user* with a generic information need. The latter cannot create highly relevant queries for search (for example, students trying to find information about a new topic, researchers trying to assimilate different points of view about a historical issue, a scientist verifying a claim by another scientist, or a person trying to shop for clothing).
* **Types of Data**. Search systems can be tailored to specific types of data. Forexample, the problem of retrieving information about a specific topic may be handled more efficiently by customized search systems that are built to collect and retrieve only information related to that specific topic. The information repository could be hierarchically organized based on a concept or topic hierarchy. These topical *domain-specific* or *vertical IR systems* are not as large as or as diverse as the generic World Wide Web, which contains information on all kinds of topics. Given that these domain-specific collections exist and may have been acquired through a specific process, they can be exploited much more efficiently by a specialized system.

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

* Informational Web Search Queries Informational searches are utilized to educate the searcher, they’re looking for answers or more details on a subject. For example, a search on “Internet Marketing” will yield many results with a very broad approach to the keyword. These searches can be purely academic, where the searcher is looking for knowledge on a particular topic which will lead to no monetary exchange or they can be commercially based, where a prospective client would be looking for details on a particular product, and who offers the product.

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

* There are two main modes of interaction with IR systems: **retrieval and browsing** which, although similar in goal, are accomplished through different interaction tasks. **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. **Hyperlinks** are used to interconnect Web pages and are mainly used for browsing. **Anchor texts** are text phrases within documents used to label hyperlinks and are very relevant to browsing.
* **Web search**combines both aspects—browsing and retrieval—and is one of themain applications of information retrieval today. Web pages are analogous to documents. Web search engines maintain an indexed repository of Web pages, usually using the technique of inverted indexing (see Section 27.5). They retrieve the most relevant Web pages for the user in response to the user’s search request with a possible ranking in descending order of relevance. The **rank of a Webpage** in a retrieved set is the measure of its relevance to the query that generated the result set.

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

Database:

* Structured data
* Schema driven
* Relational (or object, hierarchical, and network) model is predominant
* Structured query model
* Rich metadata operations
* Query returns data
* Results are based on exact matching (always correct)

IR systems:

* Unstructured data
* No fixed schema; various data models(e.g., vector space model)
* Free-form query models
* Rich data operations
* Search request returns list or pointers to documents
* Results are based on approximate matching and measures of effectiveness (may be imprecise and ranked.

7. Describe the main components of the IR system.

Components of a traditional information retrieval system experiment include the:

* *indexing system* – indexing and searching methods and procedures (an indexing system can be human or automated);
* *collection of documents* – text, image or multimedia documents, or document surrogates (for example bibliographical records);
* *defined set of queries* – which are input into the system, with or without the involvement of a human searcher; and
* *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?

* A digital library is where the information is digitalized and stored in the form of electronic portal that provides access to all kind of database such as bibliographies, full-text resources, catalogues, search engines, internet resources, reference works, E-journals etc. Earlier its only paper-based works only available in libraries. However, technologies have overcome the traditional techniques and ways of gathering information from libraries. . Now in present life we have modest facilities which are called a digital library. Types of data found in digital libraries includes text, visual material, audio material, video material, stored as electronic media formats (as opposed to print, microform, or other media), along with means for organizing, storing, and retrieving the files and media contained in the library collection.

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

* **Digital libraries** can be broadly defined as collections of electronic resources and services for the delivery of materials in a variety of formats. These collections may include a university’s library catalog, catalogs from a group of participating universities as in the State of Florida University System, or a compilation of multiple external resources on the World Wide Web such as Google Scholar or the IEEE/ACM index. These interfaces provide universal access to different types of content—such as books, articles, audio, and video—situated in different database systems and remote repositories. Similar to real libraries, these digital collections are maintained via a catalog and organized in categories for online reference. Digital libraries “include personal, distributed, and centralized collections such as online public access catalogs (OPACs) and bibliographic databases, distributed document databases, scholarly and professional discussion lists and electronic journals, other online databases, forums, and bulletin boards.”

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

* Information retrieval has been a common task since the times of ancient civilizations, which devised ways to organize, store, and catalog documents and records. Media such as papyrus scrolls and stone tablets were used to record documented information in ancient times. These efforts allowed knowledge to be retained and transferred among generations. With the emergence of public libraries and the printing press, large-scale methods for producing, collecting, archiving, and distributing documents and books evolved. As computers and automatic storage systems emerged, the need to apply these methods to computerized systems arose. Several techniques emerged in the 1950s, such as the seminal work of H. P. Luhn, who pro-posed using words and their frequency counts as indexing units for documents, and using measures of word overlap between queries and documents as the retrieval criterion. It was soon realized that storing large amounts of text was not difficult. The harder task was to search for and retrieve that information selectively for users with specific information needs. Methods that explored word distribution statistics gave rise to the choice of keywords based on their distribution properties and keyword-based weighting schemes.
* The earlier experiments with document retrieval systems such as SMART in the 1960s adopted the *inverted file organization* based on keywords and their weights as the method of indexing (see Section 27.5). Serial (or sequential) organization proved inadequate if queries required fast, near real-time response times. Proper organization of these files became an important area of study; document classification and clustering schemes ensued. The scale of retrieval experiments remained a challenge due to lack of availability of large text collections. This soon changed with the World Wide Web. Also, the Text Retrieval Conference (TREC) was launched by NIST (National Institute of Standards and Technology) in 1992 as a part of the TIPSTER program with the goal of providing a platform for evaluating information retrieval methodologies and facilitating technology transfer to develop IR products.

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

* In this model, documents are represented as a set of *terms*. Queries are formulated as a combination of terms using the standard Boolean logic set-theoretic operators such as AND, OR and NOT. Retrieval and relevance are considered as binary concepts in this model, so the retrieved elements are an “exact match” retrieval of relevant documents. There is no notion of ranking of resulting documents. All retrieved documents are considered equally important—a major simplification that does not consider frequencies of document terms or their proximity to other terms com-pared against the query terms.
* Boolean retrieval models lack sophisticated ranking algorithms and are among the earliest and simplest information retrieval models. These models make it easy to associate metadata information and write queries that match the contents of the documents as well as other properties of documents, such as date of creation, author, and type of document.

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

* The vector space model provides a framework in which term weighting, ranking of retrieved documents, and relevance feedback are possible. Documents are represented as *features* and *weights* of term features in an *n*-dimensional vector space of terms. **Features** are a subset of the terms in a *set of documents* that are deemed most relevant to an IR search for this particular set of documents. The process of selecting these important terms (features) and their properties as a sparse (limited) list out of the very large number of available terms (the vocabulary can contain hundreds of thousands of terms) is independent of the model specification. The query is also specified as a terms vector (vector of features), and this is compared to the document vectors for similarity/relevance assessment.
* The similarity assessment function that compares two vectors is not inherent to the model—different similarity functions can be used. However, the cosine of the angle between the query and document vector is a commonly used function for similarity assessment. As the angle between the vectors decreases, the cosine of the angle approaches one, meaning that the similarity of the query with a document vector increases. Terms (features) are weighted proportional to their frequency counts to reflect the importance of terms in the calculation of relevance measure. This is different from the Boolean model, which does not take into account the frequency of words in the document for relevance match.

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?

Term Frequency-Inverse Document Frequency Scheme

TF-IDF Scheme:

* TF-IDF (Term Frequency-Inverse Document Frequency) is used to determine the document term weight based on some variation of the term frequency.
* It is a statistical measure of the weights which can be used to determine the importance of the word in a document.
* TF-IDF is used mainly while retrieving the information and also used while mining the text.

14. What are probabilistic and semantic models of IR?

Probabilistic and Semantic Models of IR

Probabilistic model of IR (Information retrieval):

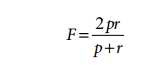
* The probabilistic model is definitive and ranking of documents is based on the estimated probability of relevance with each query and the document.
* The documents ranking is done based on the probability ranking principle in probabilistic model.
* In the probabilistic framework, Information retrieval system will verify whether the documents belong to the relevant set of documents (or) the non-relevant set for a query.
* Calculate the probabilities of the relevant and non-relevant set of documents and compare them. So, this is one the important model of information retrieval systems.

15. Define recall and precision in IR systems.

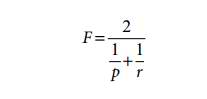
* [Precision and recall](https://doi.org/10.1007/978-0-387-30164-8_652) are the measures used in the information retrieval domain to measure how well an information retrieval system retrieves the relevant documents requested by a user. The measures are defined as follows:
* Precision - Total number of documents retrieved that are relevant/Total number of documents that are retrieved.
* Recall - Total number of documents retrieved that are relevant/Total number of relevant documents in the database.
* We can use the same terminology used in a [confusion matrix](https://doi.org/10.1007/978-0-387-30164-8_157) to define these two measures. Let relevant documents be positive examples and irrelevant documents, negative examples. The two measures can be redefined with reference to a special case of the confusion matrix, with two classes, one designated the *positive* class, and the other the *negative* class.

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

* F-score (*F*) is the harmonic mean of the precision (*p*) and recall (*r*) values. High precision is achieved almost always at the expense of recall and vice versa. It is a matter of the application’s context whether to tune the system for high precision or high recall. F-score is a single measure that combines precision and recall to compare different result sets:



* One of the properties of harmonic mean is that the harmonic mean of two numbers tends to be closer to the smaller of the two. Thus, *F* is automatically biased toward the smaller of the precision and recall values. Therefore, for a high F-score, both precision and recall must be high.



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

Types of queries used in IR systems are:

**Keyword Queries:**

* Simplest and most common queries.
* The user enters just keyword combinations to retrieve documents.
* These keywords are connected by logical AND operator.
* All retrieval models provide support for keyword queries.

**Boolean Queries:**

* Some IR systems allow using +, -, AND, OR, NOT, ( ), Boolean operators in combination of keyword formulations.
* No ranking is involved because a document either satisfies such a query or does not satisfy it.
* A document is retrieved for boolean query if it is logically true as exact match in document.

**Phase Queries:**

* When documents are represented using an inverted keyword index for searching, the relative order of items in document is lost.
* To perform exact phase retrieval, these phases are encoded in inverted index or implemented differently.
* This query consists of a sequence of words that make up a phase.
* It is generally enclosed within double quotes.

**Proximity Queries:**

* Proximity refers ti search that accounts for how close within a record multiple items should be to each other.
* Most commonly used proximity search option is a phase search that requires terms to be in exact order.
* Other proximity operators can specify how close terms should be to each other. Some will specify the order of search terms.
* Search engines use various operators’ names such as NEAR, ADJ (adjacent), or AFTER.
* However, providing support for complex proximity operators becomes expensive as it requires time-consuming pre-processing of documents and so it is suitable for smaller document collections rather than for web.

**Wildcard Queries:**

* It supports regular expressions and pattern matching-based searching in text.
* Retrieval models do not directly support for this query type.
* In IR systems, certain kinds of wildcard search support may be implemented. Example: usually words ending with trailing characters.

**Natural Language Queries:**

* There are only a few natural language search engines that aim to understand the structure and meaning of queries written in natural language text, generally as question or narrative.
* The system tries to formulate answers for these queries from retrieved results.
* Semantic models can provide support for this query type.

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

**Processing Phrase and Proximity Queries**

The following are approaches of processing phrase and proximity queries.

Phrase Queries:

* The phrase queries are the sequence of the words which are used to make a phrase.
* We have to enclose the phrase in double quotes.
* We should encode the phrases in the inverted index to make sure that the documents are not lost and for performing exact phrase retrieval by placing the terms in the exact order.
* It is based on the conceptual database design and this type of searching is very much restricted for using.

Proximity Queries:

* It looks for documents where two or more separately matching term occurrences are within a specified [distance](https://en.wikipedia.org/wiki/String_distance), where distance is the number of intermediate words or characters.
* In addition to proximity, some implementations may also impose a constraint on the word order, in that the order in the searched text must be identical to the order of the search query.
* Proximity searching goes beyond the simple matching of words by adding the constraint of proximity and is generally regarded as a form of advanced search.

19. Describe the detailed IR process

**Information Retrieval (IR)** is the activity of obtaining information from large collections of Information sources in response to a need.

The working of Information Retrieval process is explained below

* The Process of Information Retrieval starts when a user creates any query into the system through some graphical interface provided.
* These user-defined queries are the statements of needed information. for example, queries fork by users in search engines.
* In IR single query does not match to the right data object instead it matches with the several collections of data objects from which the most relevant document is taken into consideration for further evaluation.
* The ranking of relevant documents is done to find out the most related document to the given query.
* This is the key difference between the Database searching and Information Retrieval.
* After the query is sent to the core of the system. This part has the access to the content management module which is directly linked with the back-end i.e. the large collections of data objects.
* Once results R are generated by the core system then it is returned to the user by some graphical user interfaces.
* The process repeats and results are modified until the user satisfied for what he is actually looking for.

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

* Stop word elimination and stemming are commonly used method in indexing. Stop words are high frequency words that have little semantic weight and are thus unlikely to help the retrieval process. Usual practice in IR is to drop them from index. Stemming conflates morphological variants of words in its root or stem. It frees user from worrying about the truncation and inflection while framing queries and helps in reducing index size. Stemming does help in improving the retrieval performance. Particularly, recall is expected to improve after stemming. When using stemming as a means to improve retrieval effectiveness one should be careful about under stemming and over stemming in choosing stemmer. Under stemming occurs when related words are not reduced to same stem. This may result in missing relevant document. Over stemming occurs when unrelated words are reduced to same stem thereby causing a match between query and irrelevant documents.

**Stemming:** reducing related words to a common stem.

**Removal of stop words:** removal of commonly used words unlikely to be useful for learning.

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

In the context of [information retrieval](https://en.wikipedia.org/wiki/Information_retrieval), a thesaurus (plural: "thesauri") is a form of [controlled vocabulary](https://en.wikipedia.org/wiki/Controlled_vocabulary) that seeks to dictate semantic manifestations of [metadata](https://en.wikipedia.org/wiki/Metadata) in the indexing of content objects. A thesaurus serves to minimize semantic ambiguity by ensuring uniformity and consistency in the storage and retrieval of the manifestations of content objects.

* The prime function of a thesaurus is to support information retrieval by guiding the choice of terms for indexing and searching.
* The traditional aim of a thesaurus is to guide the indexer and the searcher to choose the same term for the same concept, a thesaurus should first list all the concepts that might be useful for retrieval purposes in a given domain. The concepts are represented by terms, and for each concept, one of the possible representations is selected as the preferred term. Secondly, a thesaurus should present the preferred terms in such a way that people will easily identify the one(s) they need. This is achieved by establishing relationships between terms — and/or between concepts — and using the relationships to present the terms in a structured display.

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

* **Information extraction is the process of extracting information from unstructured textual sources to enable finding entities as well as classifying and storing them in a database. Semantically enhanced information extraction (also known as semantic annotation) couples those entities with their semantic descriptions and connections from a knowledge graph. By adding metadata to the extracted concepts, this technology solves many challenges in enterprise content management and knowledge discovery.**

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

Information Retrieval Systems

Vocabularies in IR (Information retrieval) systems:

* IR systems is refers to set of unique query terms in the document set.
* Every term in a vocabulary set will have all the associated or combined collection of information of the document like the document id and counts of the occurrence.

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

* An **inverted index** of a document collection is a data structure that attaches distinct terms with a list of all documents that contains the term. The process of inverted index construction involves the extraction and processing steps. Acquired text is first preprocessed and the documents are represented with the vocabulary terms. Documents’ statistics are collected in document lookup tables. Statistics generally include counts of vocabulary terms in individual documents as well as different collections, their positions of occurrence within the documents, and the lengths of the documents. The vocabulary terms are weighted at indexing time according to different criteria for collections. For example, in some cases terms in the titles of the documents may be weighted more heavily than terms that occur in other parts of the documents.
* One of the most popular weighting schemes is the TF-IDF (term frequency-inverse document frequency) metric. For a given term this weighting scheme distinguishes to some extent the documents in which the term occurs more often from those in which the term occurs very little or never. These weights are normalized to account for varying document lengths, further ensuring that longer documents with proportionately more occurrences of a word are not favored for retrieval over shorter documents with proportionately fewer occurrences. These processed document-term streams (matrices) are then inverted into term-document streams (matrices) for further IR steps.

The different steps involved in inverted index construction can be summarized as follows:

**1.**    Break the documents into vocabulary terms by tokenizing, cleansing, stop-word removal, stemming, and/or use of an additional thesaurus as vocabulary.

**2.**    Collect document statistics and store the statistics in a document lookup table.

**3.**    Invert the document-term stream into a term-document stream along with additional information such as term frequencies, term positions, and term weights.

 Searching for relevant documents from the inverted index, given a set of query terms, is generally a three-step process.

25. Define relevance feedback.

* Without proper evaluation techniques, one cannot compare and measure the relevance of different retrieval models and IR systems in order to make improvements.

Evaluation techniques of IR systems measure the topical relevance and user relevance.

* **Topical relevance** measures the extent to which the topic of a resultmatches the topic of the query. Mapping one’s information need with “perfect” queries is a cognitive task, and many users are not able to effectively form queries that would retrieve results more suited to their information need. Also, since a major chunk of user queries are informational in nature, there is no fixed set of right answers to show to the user.
* **User relevance** is a term used to describe the “goodness” of a retrieved result with regard to the user’s information need. User relevance includes other implicit factors, such as user perception, context, timeliness, the user’s environment, and current task needs. Evaluating user relevance may also involve subjective analysis and study of user retrieval tasks to capture some of the properties of implicit factors involved in accounting for users’ bias for judging performance.

26. Describe the three types of Web analyses.

**Web Analyses**

The following are the three different types of Web Analyses are

1) Web analytics

2) Web spamming

3) Web security

**Web Analytics:**

* The main objective of the web analytics is to understand and also to improve the performance of the web usage.
* They will collect the data, analyze the data and monitor the performance to see the effectiveness of the website.
* The data which they collect is compared to improve the performance.
* It is used to improve the marketing strategies.

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

* **The agent-based approach**involves the development of sophisticated artificialintelligence systems that can act autonomously or semi-autonomously on behalf of a particular user, to discover and process Web-based information. Generally, the agent-based Web analysis systems can be placed into the following three categories:
* **Intelligent Web agents**are software agents that search for relevant information using characteristics of a particular application domain (and possibly a user profile) to organize and interpret the discovered information. For example, an intelligent agent that retrieves product information from a variety of vendor sites using only general information about the product domain.
* **Information Filtering/Categorization**is another technique that utilizesWeb agents for categorizing Web documents. These Web agents use methods from information retrieval, and semantic information based on the links among various documents to organize documents into a concept hierarchy.
* **Personalized Web agents**are another type of Web agents that utilize the personal preferences of users to organize search results, or to discover information and documents that could be of value for a particular user. User preferences could be learned from previous user choices, or from other individuals who are considered to have similar preferences to the user.

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

Web Content and Web Query Systems

Database based approach for analyzing the web content:

* Web content is the process of transferring the website to a database to enhance the information management and also to apply queries on the web.
* For the database-based approach firstly, model and then integrate the data so more important queries will be performed.
* Object Exchange model is used by the database-based approach to represent the semi structured data in a labeled graph.

29. What algorithms are popular in ranking or determining the importance of Web pages? Which algorithm was proposed by the founders of Google?

* The structure and content of Web pages are often combined for information retrieval by Web search engines. Given a collection of interconnected Web documents, interesting and informative facts describing their connectivity in the Web subset can be discovered. Web structure analysis is also used to help with navigation and make it possible to compare/integrate different Web page schemes. This aspect of Web structure analysis facilitates Web document classification and clustering on the basis of structure.
* **The PageRank Ranking Algorithm.** Ranking algorithms are used to order search results based on relevance and authority. Google uses the well-known PageRank algorithm, 28 which is based on the “importance” of each page. Every Web page has a number of forward links (out-edges) and backlinks (in edges). It is very difficult to determine all the backlinks of a Web page, whereas it is relatively straightforward to determine its forward links. According to the PageRank algorithm, highly linked pages are more important (have greater authority) than pages with fewer links. However, not all backlinks are important. A backlink to a page from a credible source is more important than a link from some arbitrary page. Thus a page has a high rank if the sum of the ranks of its backlinks is high. PageRank was an attempt to see how good an approximation of the “importance” of a page can be obtained from the link structure.
* **The HITS Ranking Algorithm**. The HITS29 algorithm proposed by Jon Kleinberg is another type of ranking algorithm exploiting the link structure of the Web. The algorithm presumes that a good hub is a document that points to many hubs, and a good authority is a document that is pointed at by many other authorities. The algorithm contains two main steps: a sampling component and a weight-propagation component.

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

Web Usage Analysis

The following are the two things that we can learn from web usage analysis.

1) Web usage analysis is uses to know how to use application data analysis techniques to discover usage pattern from Web data.

2) We can know the pattern of the usage of the web pages like IP address, page references and time and date of access for a user. It will improve the searching experience for the users.

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

Mining Operations

The following are the mining operations are as follows:

* The Web usage data will be describes the pattern of web pages.

For example, preferences of the pages, history, date and time of the accessing user groups (or) users of the web pages and also the IP address.

* Mining operations can be used to detect threats and provide security.

For example, they can develop a security system which alerts if the user’s different from template which was analyzed by the regular usage of the web by the user.

32. What are the applications of Web usage mining?

Applications of Web usage Mining

Web Mining: It is defined as an analysis and detection of the information which is useful and available on the World Wide Web.

There are many application of web usage mining, some of them are mentioned below.

Market Segmentation:

* Market segmentation is used by the web marketers to target useful campaigns and messages to the groups which are interested.
* This can also be used to assist the user.

33. What is search relevance? How is it determined?

Search Relevance

Search relevance:

* Search relevance is one of the techniques which are used to compare the relevance of the different information retrieval systems for improving the system and making the system efficient.
* Search relevance is improves the user’s experience of searching the information that is essential for the user.

34. Define and explain conversational search.

* Conversational search is being consistently looked to as the technology that will power this search revolution.
* It allows users to submit queries, typically through voice, and receive answers in the form of a conversation. As opposed to a traditional keyword search, a conversational search system takes complex grammatical sentences and can use context from previous interactions to provide more useful and comprehensive results.
* Conversational search is distinct from voice search which allows users to submit spoken queries, but returns answers in text, voice, or other formats that don’t resemble a conversation.

35. Define topic modeling.

* Topic modeling is an unsupervised machine learning technique that’s capable of scanning a set of documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions that best characterize a set of documents.