

**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.**

**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.*

**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?**

**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?**

The (standard) **Boolean model of information retrieval** (BIR) is a classical **information retrieval** (**IR**) **model** and, at the same time, the first and most-adopted one. It is used by many **IR** systems to this day. Retrieval is based on whether or not the documents contain the query terms.

**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?**

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?**

**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 referring 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.**

**25. Define relevance feedback.**

**26. Describe the three types of Web analyses.**

**Web Analyses**

The following are the three different types of Web Analyses are

* + Web analytics
  + Web spamming
  + 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?**

**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?**

**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.