UNIT V: INFORMATION RETRIEVAL AND WEB SEARCH

SYLLABUS

IR concepts – Retrieval Models – Queries in IR system – Text Preprocessing – Inverted Indexing – Evaluation Measures – Web Search and Analytics – Ontology based Search - Current trends.

SNAPSHOT

1) IR concepts
2) Retrieval Models
3) Queries in IR system
4) Text Preprocessing
5) Inverted Indexing
6) Evaluation Measures
7) Web Search and Analytics
8) Ontology based Search
9) Current trends

TABLE OF CONTACTS

IR CONCEPTS

Information retrieval is the process of retrieving documents from a collection in response to a query (or a search request) by a user.

- Historically, information retrieval is "the discipline that deals with the structure, Analysis, organization, storage, searching, and retrieval of information"
- We can enhance the definition slightly to say that
  It applies in the context of unstructured documents to satisfy a user's information Needs.
- IR systems go beyond database systems in that they do not limit the user to a specific query language, nor do they expect the user to know the structure (schema) or Content of a particular database.
- IR systems use a user's information need expressed
  As a free-form search request (sometimes called a keyword search query, or just Query) for interpretation by the system.

An IR system can be characterized at different levels: Different IR systems are designed to address specific Problems that require a combination of different characteristics. These characteristics can be briefly described as follows:

- **Types of Users.** The user may be an expert user (for example, a curator or a Librarian), 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.
- **Types of Data**. Search systems can be tailored to specific types of data. For Example, 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.
- **Types of Information Need.** In the context of Web search, users' information Needs may be defined as navigational, informational, or transactional.
  - Navigational search refers to finding a particular piece of information that a user needs quickly.
  - The purpose of Informational search is to find current information about a topic.
  - The goal of transactional search is to reach a site where further interaction happens.

MODES OF INTERACTION IN IR SYSTEMS

- We defined information retrieval as the process of
  Retrieving documents from a collection in response to a query (or a search request) By a user.

- Other kinds of documents include images, audio recordings, video strips, and Maps.
- Typically the collection is made up of documents containing unstructured Data.
- Data may be scattered non uniformly in these documents with no definitive Structure.
- A query is a set of terms (also referred to as keywords) used by the Searcher to specify an information need.
- There are two main modes of interaction with IR systems
    - Retrieval is concerned with the extraction of relevant information from A repository of documents through an IR query
    - Browsing signifies the activity of a user visiting or navigating through similar or related documents based on the user's assessment of relevance
- 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.

RESPONSE TO A QUERY (OR A SEARCH REQUEST) BY A USER.

- Response to a query, or a search request, in information retrieval refers to the process of retrieving and presenting relevant information to the user based on their query.
- Information retrieval systems are designed to efficiently search and retrieve relevant documents or data from large collections of information.

DETAILED EXPLANATION OF THE RESPONSE TO A QUERY IN INFORMATION RETRIEVAL:

1. **Query Input:**
    - The user submits a query to the information retrieval system.
    - The query can be a single word, a phrase, or a complex question.
    - It represents the user's information need and what they are looking for.
2. **Query Processing:**
    - The information retrieval system processes the query to understand its structure and meaning.
    - This typically involves several steps, such as tokenization, stop word removal, stemming, and parsing.
    - The goal is to transform the query into a format that can be used for matching against the documents in the collection.
3. **Indexing:**
    - Prior to the query, the information retrieval system builds an index of the collection.
    - The index is a data structure that allows efficient lookup of documents based on terms or keywords.

- It typically includes an inverted index, which maps terms to the documents that contain them.
- During query processing, the system uses the index to identify potentially relevant documents.

4. **Retrieval Models:**
   - Information retrieval systems employ different retrieval models to rank the documents based on their relevance to the query.
   - Commonly used models include the vector space model, probabilistic models (such as the Okapi BM25), and language models.
   - These models assign a relevance score to each document, reflecting its estimated relevance to the query.
5. **Ranking:**
   - The information retrieval system ranks the retrieved documents based on their relevance scores.
   - The ranking determines the order in which the documents will be presented to the user.
   - Typically, the most relevant documents are displayed at the top of the list.
6. **Presentation:**
   - The system presents the retrieved documents to the user in a user-friendly format.
   - This can vary depending on the type of information being retrieved.
   - For textual documents, the system might show snippets or summaries of the documents, along with the document titles and URLs.
   - In some cases, additional features like highlighting relevant terms or clustering related documents may be provided.
7. **Result Evaluation:**
   - The user evaluates the presented results and determines if they meet their information needs.
   - The user may choose to click on a document to access its full content or refine their query based on the initial results.
8. **Query Refinement:**
   - Based on the initial results, the user may refine their query by modifying or expanding it to better capture their information needs.
   - This iterative process helps the user to gradually narrow down the search and retrieve more relevant information.
9. **Feedback:**
   - The user's interactions with the search results, such as clicks on documents or explicit feedback, can be utilized to improve the performance of the information retrieval system.
   - This feedback is often used to re-rank the documents or adapt the system's retrieval models to better align with the user's preferences.

RETRIEVAL MODELS

BOOLEAN MODEL

- 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.
- Boolean retrieval models lack sophisticated ranking algorithms and are among the Earliest and simplest information retrieval models.

VECTOR SPACE MODEL

- 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.
- 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 process of selecting these important terms (features) and their properties is independent of the model specification.
- 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.

PROBABILISTIC MODEL

- In the probabilistic model, a more concrete and
  Definitive approach is taken: ranking documents by their estimated probability of Relevance with respect to the query and the document.
- This is the basis of the Probability Ranking Principle.
- In the probabilistic framework, the IR system has to decide whether the documents Belong to the relevant set or the non relevant set for a query.

- To make this decision, it is assumed that a predefined relevant set and non relevant set exist for the query, And the task is to calculate the probability that the document belongs to the relevant Set and compare that with the probability that the document belongs to the non relevant set.
- Given the document representation D of a document, estimating the relevance R
  And nonrelevance NR of that document involves computation of conditional probability P(R | D) and P(NR | D).
- These conditional probabilities can be calculated using
  Bayes' Rule:12
  P(R | D) = P(D | R) × P®/P(D)
  P(NR | D) = P(D | NR) × P(NR)/P(D)
- A document D is classified as relevant if P(R | D) > P(NR | D). Discarding the constant
- P(D), this is equivalent to saying that a document is relevant if:
  P(D | R) × (P) > P(D | NR) × P(NR)

SEMANTIC MODEL

- In semantic models, the process of Matching documents to a given query is based on concept level and semantic Matching instead of index term (keyword) matching.
- This allows retrieval of relevant documents even when these associations are not inherently observed or statistically captured.
- Semantic approaches include different levels of analysis, such as
  - **In Morphological analysis**, roots and affixes are analysed to determine the parts of Speech (nouns, verbs, adjectives, and so on) of the words.
  - Following morphological Analysis, **syntactic analysis** follows to parse and analyse complete phrases in documents.
  - Finally, the semantic methods have to resolve word ambiguities and/or generate relevant synonym.
- The development of a sophisticated semantic system requires complex knowledge bases.
- These systems often Require techniques from artificial intelligence and expert systems.
- Knowledge bases Like Cyc15 and WordNet16 have been developed for use in knowledge-based IR systems based on semantic models.

QUERIES IN IR SYSTEM

KEYWORD QUERIES

- Keyword-based queries are the simplest and most commonly used forms of IR Queries: the user just enters keyword combinations to retrieve documents.
- The Query keyword terms are implicitly connected by a logical AND operator.
- A query Such as 'database concepts' retrieves documents that contain both the words 'data-
  Base' and 'concepts' at the top of the retrieved results.
- In addition, most systems also Retrieve documents that contain only 'database' or only 'concepts' in their text.
- Some Systems remove most commonly occurring words (such as a, the, of, and so on, called stop words) as a pre-processing step before sending the filtered query keywords to the IR engine.

BOOLEAN QUERIES

- Some IR systems allow using the AND, OR, NOT, ( ), + , and – Boolean operators in Combinations of keyword formulations.
- AND requires that both terms be found.
- OR lets either term be found.
- NOT means any record containing the second term
  Will be excluded.
- '( )' means the Boolean operators can be nested using parentheses.
- Complex Boolean queries can be built out of these operators and their combinations, and They are evaluated according to the classical rules of Boolean algebra.
- No ranking is Possible, because a document either satisfies such a query (is "relevant") or does not
  Satisfy it (is "nonrelevant").
- A document is retrieved for a Boolean query if the
  Query is logically true as an exact match in the document.

PHRASE QUERIES

- A Phrase query consists of a sequence of words that makes up a phrase.
- The phrase is Generally enclosed within double quotes.
- Each retrieved document must contain at
  Least one instance of the exact phrase.

PROXIMITY QUERIES

- The most commonly used proximity search Option is a phrase search that requires terms to be in the exact order.
- Other proximity operators can specify how close terms should be to each other.
- Some will also Specify the order of the search terms.

- Each search engine can define proximity operators differently, and the search engines use various operator names such as NEAR,ADJ(adjacent), or AFTER.
- In some cases, a sequence of single words is given,
  Together with a maximum allowed distance between them.

WILDCARD QUERIES

- Wildcard searching is generally meant to support regular expressions and pattern Matching-based searching in text.
- In IR systems, certain kinds of wildcard search
  Support may be implemented—usually words with any trailing characters (for example, 'data*' would retrieve data, database, datapoint, dataset, and so on).
- Retrieval models do not directly provide support for this query type.

NATURAL LANGUAGE QUERIES

- There are a few natural language search engines that aim to understand the Structure and meaning of queries written in natural language text, generally as a question or narrative.
- The system tries to formulate answers for such queries from retrieved Results. Some search systems are starting to provide natural language interfaces to Provide answers to specific types of questions.
- Such questions are usually easier to answer Because there are strong linguistic patterns giving clues to specific types of sentences—for example, 'defined as' or 'refers to.' Semantic models can provide support for this query type.

TEXT PREPROCESSING

- Stop words are very commonly used words in a language that play a major role in the formation of a sentence but which seldom contribute to the meaning of that Sentence.
- Words that are expected to occur in 80 percent or more of the documents in a collection are typically referred to as stop words, and they are rendered potentially useless.
- Because of the commonness and function of these words, they do not Contribute much to the relevance of a document for a query search.
- Examples Include words such as the, of, to, a, and, in, said, for, that, was, on, he, is, with, at, by, And it.
- Removal of stop words from a document must be performed before indexing.
- Queries must also be pre-processed for stop word removal before the Actual retrieval process.

- Removal of stop words results in elimination of possible Spurious indexes, thereby reducing the size of an index structure by about 40 percent or more.
- However, doing so could impact the recall if the stop word is an Integral part of a query (for example, a search for the phrase 'To be or not to be,' Where removal of stop words makes the query inappropriate, as all the words in the Phrase are stop words). Many search engines do not employ query stop word Removal for this reason.

STEMMING

- A stem of a word is defined as the word obtained after trimming the suffix and prefix of an original word.
- For example, 'compute' is the stem word for computer, computing, and computation.
- These suffixes and prefixes are very common in the English language for supporting the notion of verbs, tenses, and plural forms.
- Stemming reduces the different forms of the word formed by inflection (due to plurals or tenses) and derivation to a common stem.

UTILIZING A THESAURUS

- A thesaurus comprises a precompiled list of important concepts and the main word That describes each concept for a particular domain of knowledge.
- For each concept in this list, a set of synonyms and related words is also compiled.
- Thus, a synonym Can be converted to its matching concept during preprocessing.
- Usage of A thesaurus, also known as a collection of synonyms, has a substantial impact on the Recall of information systems.
- This process can be complicated because many words Have different meanings in different contexts.

OTHER PREPROCESSING STEPS: DIGITS, HYPHENS, PUNCTUATION MARKS, CASES

- Digits, dates, phone numbers, e-mail addresses, URLs, and other standard types of Text may or may not be removed during preprocessing. Web search engines, However, index them in order to use this type of information in the document
- Hyphens and punctuation marks may be handled in different ways. Either the entire Phrase with the hyphens/punctuation marks may be used, or they may be eliminated. In some systems, the character representing the hyphen/punctuation mark May be removed, or may be replaced with a space.
- Most information retrieval systems perform case-insensitive search, converting all the letters of the text to uppercase or lowercase.

INFORMATION EXTRACTION

- Information extraction (IE) is a generic term used for extracting structured content from text.
- Text analytic tasks such as identifying noun phrases, facts, events, People, places, and relationships are examples of IE tasks.
- IE technologies are mostly used to identify contextually relevant features that involve text analysis, matching, and categorization for improving the relevance of search systems.

INVERTED INDEXING

- An inverted index Structure comprises vocabulary and document information.
- Vocabulary is a set of Distinct query terms in the document set.
- Each term in a vocabulary set has an associated collection of information about the documents that contain the term, such as Document id, occurrence count, and offsets within the document where the term Occurs.
- Weights are assigned to document terms to represent an estimate of the usefulness of the given term.
- These Weights are normalized to account for varying document lengths, further ensuring That longer documents with proportionately more occurrences of a word are not Favoured for retrieval over shorter documents with proportionately fewer occurrences.

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

- **Pre-process**
  - Break the documents into vocabulary terms by tokenizing, cleansing, Stopword removal, stemming, and/or use of an additional thesaurus as Vocabulary.
- **Collecting document statistics**
  - 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.
- **Conversion**
  - 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.

- **Vocabulary search.** If the query comprises multiple terms, they are separated and treated as independent terms. Each term is searched in the vocabulary
- **Document information retrieval.** The document information for each term Is retrieved.
- **Manipulation of retrieved information**. The document information vector for each term obtained in step 2 is now processed further to incorporate various forms of query logic. Various kinds of queries like prefix, range, context, and proximity queries are processed in this step to construct the final result Based on the document collections returned in step.

EVALUATION MEASURES

- 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

- Topical relevance measures the extent to which the topic of a result Matches 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.

USER RELEVANCE

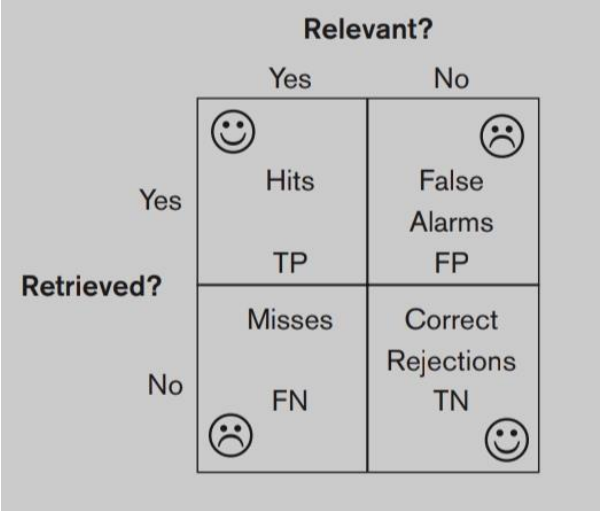
- User relevance is a term used to describe the
  "goodness" of a retrieved result with regard to the user's information need.
- User relevance includes other implicit factors, such as user perception, context, timeliness, The user's environment, and current task needs.

In Web information retrieval, no binary classification decision is made on whether a Document is relevant or non relevant to a query. Instead, a ranking of the documents is produced for the user.

Therefore, some evaluation measures Focus on comparing different rankings produced by IR systems.

RECALL AND PRECISION

- Recall and precision metrics are based on the binary relevance assumption (whether Each document is relevant or nonrelevant to the query).



The notation for Figure 27.5 is as follows:

- TP: true positive
- FP: false positive
- FN: false negative
- TN: true negative

- Using the term hits for the documents that Truly or "correctly" match the user request
- Recall is defined as the Number of relevant documents retrieved by a search divided by the total number of Existing relevant documents.

  Recall = |Hits| / |Relevant|
- Precision is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by That search.

  Precision = |Hits| / |Retrieved|
- Recall and precision can also be defined in a ranked retrieval setting.

  Recall r(i) = |Si| / |Dq|

  Precision p(i) = |Si| /i

AVERAGE PRECISION

- Average precision is computed based on the precision at each relevant document in the ranking.
- This measure is useful for computing a single precision value to compare different retrieval algorithms on a query q.

RECALL/PRECISION CURVE

- A recall/precision curve can be drawn based on the recall and precision values at Each rank position, where the x-axis is the recall and the y-axis is the precision.

- Instead of using the precision and recall at each rank position, the curve is commonly plotted using recall levels r(i) at 0 percent, 10 percent, 20 percent…100 percent.

F-SCORE

- 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:

$$F=(2pr)/(p+r)$$

WEB SEARCH AND ANALYSIS

- The emergence of the Web has brought millions of users to search for information.
- To make this information accessible, search engines such as Google and Yahoo! Have to crawl and index these sites and document collections in their index databases.
- These search engines are developed with the help of computer-Assisted systems to aid the curators with the process of assigning indexes.
- They consist of manually created specialized Web directories that are hierarchically organized Indexes to guide user navigation to different resources on the Web.
- Vertical search Engines are customized topic-specific search engines that crawl and index a specific Collection of documents on the Web and provide search results from that specific Collection.
- Metasearch engines are built on top of search engines: they query different search engines simultaneously and aggregate and provide search results from These sources.
- Another source of searchable Web documents is digital libraries. 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 catalogue, catalogues 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.

WEB ANALYSIS AND ITS RELATIONSHIP TO INFORMATION RETRIEVAL

- Analyse or mine information on the Web for new information of interest is an important task in information retrieval.
- Application of data analysis techniques for discovery and analysis of Useful information from the Web is known as Web analysis.
- Over the past few years The World Wide Web has emerged as an important repository of information for Many day-to-day applications for individual consumers, as well as a significant platform for e-commerce and for social networking. These properties make it an interesting target for data analysis applications.

The goals of Web analysis are to **Finding relevant information**.

- People usually search for specific information on the Web by entering keywords.
- Search services are constrained by search Relevance problems.
    - Low precision Ensues due to results that are nonrelevant to the user.
    - High recall is impossible to determine due to the inability to index all the pages on the Web. Also, measuring recall does not make sense Since the user is concerned with only the top few documents.

PERSONALIZATION OF THE INFORMATION.

- Different people have different content and presentation preferences. By collecting personal information and then Generating user-specific dynamic Web pages, the pages are personalized for
  The user.
- A personalization Engine typically has algorithms that make use of the user's personalization Information—collected by various tools—to generate user-specific search Results.

FINDING INFORMATION OF COMMERCIAL VALUE.

- This problem deals with finding interesting patterns in users' interests, behaviours, and their use of products and services, which may be of commercial value.
- For example, businesses Such as the automobile industry, clothing, shoes, and cosmetics may improve Their services by identifying patterns such as usage trends and user preferences using various Web analysis techniques.

Based on the above goals, we can classify Web analysis into three categories:

- **Web Content analysis,** which deals with extracting useful information/knowledge from Web page contents.
- **Web structure analysis**, which discovers knowledge from Hyperlinks representing the structure of the Web.

- **Web usage analysis,** which Mines user access patterns from usage logs that record the activity of every user.

SEARCHING THE WEB

- The World Wide Web is a huge corpus of information, but locating resources that Are both high quality and relevant to the needs of the user is very difficult.
- Index-based search engines have been one of the prime tools
    - By which users search for information on the Web.
    - Web search engines crawl the Web and create an index to the Web for searching purposes.
    - When a user specifies His need for information by supplying keywords, these Web search engines query Their repository of indexes and produce links or URLs with abbreviated content as Search results.
- Web pages, unlike standard text collections, contain connections to other Web pages or documents (via the use of hyperlinks), allowing users to browse from page to Page.
- A hyperlink has two components: a destination page and an anchor text Describing the link.
- A hub is a Web page or a website that links to a collection of prominent sites (authorities) on a common topic.
- A good authority is a page that is pointed to by Many good hubs, while a good hub is a page that points to many good authorities.

ONTOLOGY BASED SEARCH

- Ontologies—formal models of representation with explicitly defined concepts and Named relationships linking them—are used to address the issues of semantic heterogeneity in data sources.
- Using ontologies to effectively combine information from multiple heterogeneous sources.

Different classes of approaches are used for information Integration using ontologies.

- **Single ontology approaches**
    - Single ontology approaches use one global ontology that provides a shared Vocabulary for the specification of the semantics.
    - They work if all information sources to be integrated provide nearly the same view on a domain of Knowledge.
    - It Can serve as a Common ontology for biomedical applications.
- **Multiple ontology approach**

- In a multiple ontology approach, each information source is described by Its own ontology.
- In principle, the "source ontology" can be a combination of Several other ontologies but it cannot be assumed that the different "source Ontologies" share the same vocabulary.
- Dealing with multiple, partially over-
  Lapping, and potentially conflicting ontologies is a very difficult problem Faced by many applications, including those in bioinformatics and other Complex area of knowledge.

- **Hybrid ontology approaches**
  - Hybrid ontology approaches are similar to multiple ontology approaches:
  - The semantics of each source is described by its own ontology.
  - But in order to Make the source ontologies comparable to each other, they are built upon
    One global shared vocabulary.
  - The shared vocabulary contains basic terms
    (The primitives) of a domain of knowledge.
  - The advantage of a hybrid Approach is that new sources can be easily added without the need to modify the mappings or the shared vocabulary.

ONTOLOGY-BASED SEARCH

- Ontology-based search refers to a search approach that utilizes ontologies to enhance the retrieval of information from a given dataset or knowledge base.
- An ontology is a formal representation of knowledge that describes the concepts, relationships, and properties within a specific domain.
- It provides a structured and organized framework for understanding and reasoning about the information contained within the domain.
- In ontology-based search, the ontology acts as a semantic model that enables a more intelligent and context-aware search process.
- It goes beyond simple keyword matching and considers the meaning and relationships of the concepts involved.

the key components and steps involved in ontology-based search:

1. **Ontology Creation:**
   - The first step is to create or select an appropriate ontology that represents the domain of interest.
   - This involves defining the concepts, their hierarchical relationships, properties, and any other relevant information.
   - The ontology can be created manually or automatically generated using various methods such as text mining or machine learning techniques.

2. **Ontology Integration:**
    - If the search involves multiple datasets or knowledge bases, the ontologies representing each of them need to be integrated.
    - This step ensures that the concepts and relationships from different sources are aligned and can be used together during the search process.
3. **Query Formulation:**
    - In ontology-based search, the user formulates a query using concepts and relationships defined in the ontology.
    - The query may include keywords, specific concepts, or complex relationships between concepts.
    - The ontology provides a standardized vocabulary for expressing the query, ensuring that it is consistent and interpretable.
4. **Semantic Search:**
    - The search engine uses the ontology to perform a semantic search.
    - It examines the query and its relationships with the concepts in the ontology to identify relevant information.
    - The search process may involve techniques such as ontology reasoning, semantic matching, or inference to expand the search scope and retrieve more accurate results.
5. **Result Ranking:**
    - Once the search engine retrieves a set of results, it ranks them based on their relevance to the query.
    - The ranking may consider factors such as the proximity of the concepts mentioned in the query, the specificity of the relationships, or the popularity of the retrieved information.
6. **Result Presentation:**
    - The final step is to present the search results to the user in a meaningful and user-friendly manner.
    - The results may include relevant documents, data entries, or any other form of information that matches the query.
    - The ontology can also assist in organizing and categorizing the results based on the concepts and relationships present in the ontology.

BENEFITS OF ONTOLOGY-BASED SEARCH:

1. **Semantic Understanding:**
    - Ontology-based search goes beyond keyword matching and considers the meaning and relationships between concepts.
    - It enables a more accurate and context-aware understanding of the user's query.

2. **Improved Precision and Recall:**
   - By utilizing the rich knowledge encoded in the ontology, ontology-based search can improve the precision and recall of search results.
   - It can retrieve information that might have been missed in traditional keyword-based approaches.
3. **Structured Information Retrieval:**
   - Ontologies provide a structured framework for representing knowledge, allowing for more structured and organized retrieval of information.
   - Users can explore and navigate the domain-specific concepts and relationships to refine their search.
4. **Domain-specific Knowledge:**
   - Ontologies capture domain-specific knowledge, enabling domain experts to model and customize the search process according to their specific needs.
   - This makes ontology-based search highly adaptable and flexible.
5. **Integration of Heterogeneous Data:**
   - Ontology-based search facilitates the integration of diverse and heterogeneous data sources.
   - By mapping different ontologies, it enables the search across multiple datasets or knowledge bases, providing a comprehensive view of the domain.

CURRENT TRENDS

FACETED SEARCH

- Faceted Search is a technique that allows for integrated search and navigation experience by allowing users to explore by filtering available information.
- Facets are generally used for handling three or more dimensions of classification.
- This allows the faceted classification scheme to classify an object in various ways based on different taxonomical Criteria.
- For example, a Web page may be classified in various ways:
  by content (air-Lines, music, news, …);
  by use (sales, information, registration, …);
  by location;
  by Language used (HTML, XML, …) and

in other ways or facets. Hence, the object can Be classified in multiple ways based on multiple taxonomies.

- This search Technique is used often in ecommerce Websites and applications enabling users to
  Navigate a multi-dimensional information space.

- A facet defines properties or characteristics of a class of objects. The properties Should be mutually exclusive and exhaustive.
- For example, a collection of art objects Might be classified using
    - an artist facet (name of artist),
    - an era facet (when the art was created),
    - a type facet (painting, sculpture, mural, …),
    - a country-of-origin facet,
    - A media facet (oil, watercolour, stone, metal, mixed media, …),
    - a collection facet (where the art resides), and so on.
- Faceted search uses faceted classification that enables a user to navigate information Along multiple paths corresponding to different orderings of the facets.

SOCIAL SEARCH

- Socially enabled online information search (social search) is a new phenomenon Facilitated by recent Web technologies.
- Collaborative social search involves different
  Ways for active involvement in search-related activities such as
    - Co-located search,
    - Remote collaboration on search tasks,
    - use of social network for search,
    - use of expertise networks,
    - involving social data mining or collective intelligence to improve the Search process and
    - even social interactions to facilitate information seeking and sense Making.
- People In social groups can provide solutions (answers to questions), pointers to databases or to other people (meta-knowledge), validation and legitimization of ideas, and can Serve as memory aids and help with problem reformulation.
- Guided participation is A process in which people co-construct knowledge in concert with peers in their community .

CONVERSATIONAL SEARCH

- Conversational Search (CS) is an interactive and collaborative information finding interaction.
- The participants engage in a conversation and perform a social search activity that is aided by intelligent agents.
- The search agent Performs multiple tasks of finding relevant information and connecting the users Together; participants provide feedback to the agent during the conversations that Allows the agent to perform better.