CS8080 – INFORMATION RETRIEVAL TECHNIQUES

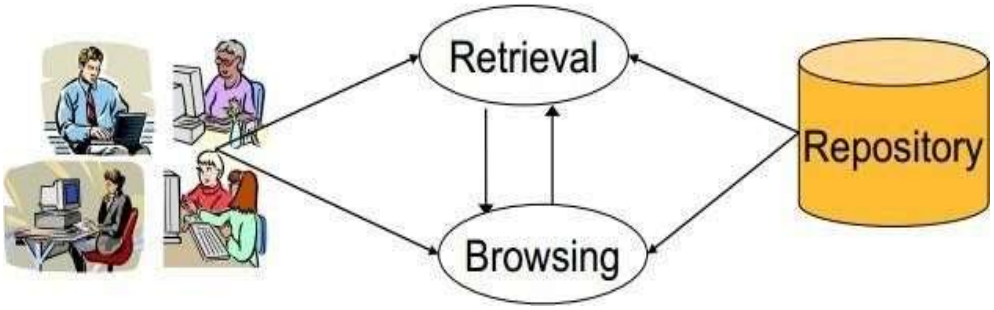
UNIT I – INTRODUCTION

Information Retrieval – Early Developments – The IR Problem – The Users Task – Information versus Data Retrieval – The IR System – The Software Architecture of the IR System – The Retrieval and Ranking Processes – The Web – The e-Publishing Era – How the web changed Search – Practical Issues on the Web – How People Search – Search Interfaces Today – Visualization in Search Interfaces.

1.1 INTRODUCTION:

- Information Retrieval (IR) is finding material (usually documents) of an unstructured nature(usually text) that satisfies an information need from within large collections.
- The role of an IR system is to retrieve all the documents, which are relevant to a query while retrieving as few non - relevant documents as possible. IR allows access to whole documents, whereas, search engines do not.
- There is a huge quantity of text, audio, video and other documents available on the Internet,on about any subject. Users need to be able to find relevant information to satisfy their particular information needs.
- There are two ways of searching for information: to use a search engine or to browse directories organized by categories. There is still a large part of the Internet that is not accessible (for example private databases and intranets).
- Information retrieval is the task of representing, storing, organizing, and offering access to information items. IR is different from data retrieval, which is about finding precise data in databases with a given structure.
- In IR systems, the information is not structured. It is contained in free form in text (WebPages or other documents) or in multimedia content. The first IR systems implemented in 1970's were designed to work with small collections of text. Some of these techniquesare now used in search engines.
- The information retrieval techniques focusing on the challenges faced by search engine. One particular challenge is the large scale, given by the huge number of WebPages availableon the Internet.
- Another challenge is inherent to any information retrieval system that deals with text: the ambiguity of the natural language (English or other languages) that makes it difficult to have perfect matches between documents and user queries.
- Information retrieval is never an easy task. The problem with IR is that document representation, either by index terms or texts cannot satisfy user need representation, which is dynamic and complicated.

- Moreover, traditional IR systems are designed to support only one type of information-seeking strategy that users engage in query formulation.
- Information retrieval (IR) is concerned with representing, searching, and manipulating large collections of electronic text and other human-language data.
- Web search engines — Google, Bing, and others — are by far the most popular and heavily used IR services, providing access to up-to-date technical information, locating people and organizations, summarizing news and events, and simplifying comparison shopping.



Web Search:

- Regular users of Web search engines casually expect to receive accurate and near- instantaneous answers to questions and requests merely by entering a short query — a few words — into a text box and clicking on a search button. Underlying this simple and intuitive interface are clusters of computers, comprising thousands of machines, working cooperatively to generate a ranked list of those Web pages that are likely to satisfy the information need embodied in the query.
- These machines identify a set of Web pages containing the terms in the query, compute a score for each page, eliminate duplicate and redundant pages, generate summaries of the remaining pages, and finally return the summaries and links back to the user for browsing.

Consider a simple example.

- If you have a computer connected to the Internet nearby, pause for a minute to launch a browser and try the query “information retrieval” on one of the major commercial Web search engines.
- It is likely that the search engine responded in well under a second. Take some time to review the top ten results. Each result lists the URL for a Web page and usually provides a title and a short snippet of text extracted from the body of the page.
- Overall, the results are drawn from a variety of different Web sites and include sites associated with leading textbooks, journals, conferences, and researchers. As is common for informational queries such as this one, the Wikipedia article may be present.

Other Search Applications:

- Desktop and file system search provides another example of a widely used IR application. A desktop search engine provides search and browsing facilities for files stored on a local hard disk and possibly on disks connected over a local network. In contrast to Web search engines, these systems require greater awareness of file formats and creation times.
- For example, a user may wish to search only within their e-mail or may know the general time frame in which a file was created or downloaded. Since files may change rapidly, these systems must interface directly with the file system layer of the operating system and must be engineered to handle a heavy update load.

Other IR Applications:

1) Document routing, filtering, and selective distribution reverse the typical IR process.
2) Summarization systems reduce documents to a few key paragraphs, sentences, or phrases describing their content. The snippets of text displayed with Web search results represent one example.
3) Information extraction systems identify named entities, such as places and dates, and combine this information into structured records that describe relationships between these entities — for example, creating lists of books and their authors from Web data.

1.2 History of IR(Early Developments) :

The idea of using computers to search for relevant pieces of information was popularized in thearticle “As We May Think” by VannevarBush in 1945. It would appear that Bushwas inspiredby patents for a 'statistical machine' - filed by Emanuel Goldberg in the 1920s and '30s - that searched for documents stored on film. The first description of a computer searching for information was described by Holmstrom in 1948, detailing an early mention of the Univac computer.

Automated information retrieval systems were introduced in the 1950s: one even featured in the 1957 romantic comedy, Desk Set. In the 1960s, the first large information retrieval research group was formed by Gerard Salton at Cornell. By the 1970s several different retrieval techniques had been shown to perform well on small text corpora such as the Cranfield collection(several thousand documents). Large-scale retrieval systems,such as the Lockheed Dialog system, came into use early in the 1970s.

In 1992, the US Department of Defense along with the National Institute of Standards and Technology (NIST), cosponsored the Text Retrieval Conference (TREC) as part of the TIPSTER text program.

The aim of this was to look into the information retrieval community by supplying the infrastructure that was needed for evaluation of text retrieval methodologies on a very large text

collection.This catalyzed research on methods that scale to huge corpora. The introduction of web search engines has boosted the need for very large scale retrieval systems even further.

1.2.1 Timeline :

1950s:

- **1950:** The term **"information retrieval"** was coined by **Calvin Mooers**.
- **1951:** Philip Bagley conducted the earliest experiment in computerized document retrieval in a master thesis at MIT.
- **1955:** Allen Kent joined from Western Reserve University published a paper in American Documentation describing the **precision and recall** measures as well as detailing a proposed **"framework" for evaluating an IR system** whichincluded statistical sampling methods for determiningthenumberof relevant documents notretrieved.
- **1959:**HansPeter Luhnpublished"Auto-encodingofdocumentsforinformationretrieval."

1960s:

early 1960s: Gerard Salton began work on IR at Harvard, later moved to **Cornell.**

- **1963:**Joseph Beckerand Robert M. Hayespublishedtextoninformationretrieval. Becker, Joseph; Hayes, Robert Mayo. Information storage and retrieval: tools, elements, theories. New York, Wiley(1963).
- **1964:**
- Karen Spärck Jones finished her thesis at Cambridge, Synonymy and Semantic Classification, and continued work on computation all inguisticsas it applies to IR.
- The National Bureau of Standards sponsored a symposium titled "Statistical Association Methods for Mechanized Documentation." Several highly significant papers, including G. Salton's firstpublishedreference (we believe) to the SMART system.

mid-1960s:

- National Library of Medicine developed MEDLARS Medical Literature Analysis and Retrieval System, the first major machine-readable database and **batch-retrieval system.**
- Project Intrex at MIT.
- **1965:** J. C. R. Licklider published Libraries of the Future.
- **late 1960s:** F. Wilfrid Lancastercompletedevaluationstudiesofthe MEDLARSsystemand publishedthe first edition of his **text on information retrieval.**
- **1968:** Gerard Salton published Automatic Information Organization and Retrieval. John W. Sammon, Jr.'s RADCTechreport"Some Mathematicsof Information Storageand Retrieval..." outlinedthevector model.
- **1969:** Sammon's "A nonlinear mapping for data structure analysis" (IEEE Transactions on Computers) was the first proposalfor visualization interface to an IR system.

1970s:

- **Early 1970s:** First online systems—NLM'sAIM-TWX,MEDLINE;Lockheed'sDialog;SDC'sORBIT. 1971: Nicholas Jardine and Cornelis J. van Rijsbergen published "The use of hierarchic clustering in information retrieval", which articulated the "cluster hypothesis."
- **1975:** Three highly influential publications by Salton fully articulated his vector processing framework and term discrimination model: ATheoryof Indexing(Societyfor IndustrialandAppliedMathematics) 1979: C. J. van Rijsbergen published Information Retrieval (Butterworths). Heavy emphasis on probabilistic models.
- **1979:** Tamas Doszkocs implemented the CITE natural language user interface for MEDLINE at the National Library of Medicine. The CITE system supported free form query input, ranked output and relevance feedback.

1980s

- **1982:** Nicholas J. Belkin, Robert N. Oddy, and Helen M. Brooks proposed the ASK (Anomalous State of Knowledge) viewpoint for information retrieval. This was an important concept, though their automated analysis toolproved ultimatelydisappointing.
- **1983:** Salton(and Michael J. McGill) published Introduction to Modern Information Retrieval (McGraw- Hill), with heavy emphasis on vector space models.
- **Mid-1980s:** Effortstodevelopend-userversionsofcommercial IR systems. 1989: First World Wide Web proposalsbyTimBerners-Lee at CERN.

1990s

- **1992:** First TREC conference.
- **1997:** Publication of Korfhage's Information Storage and Retrieval with emphasis on visualization and multi-reference point systems.late 1990s: Web search engines implementation of many features formerly found only in experimental IR systems. Search engines become the most common and maybe best instantiation of IRmodels.

2000s-present:

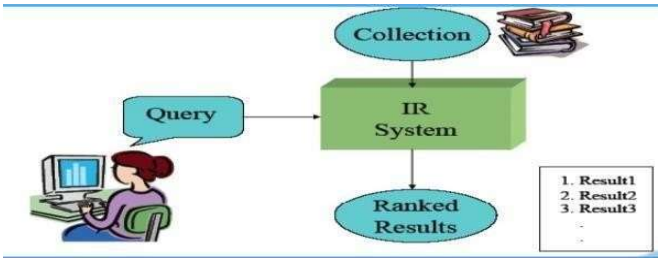
More applications, especially Web search and interactions with other fields like Learning to rank, Scalability (e.g., MapReduce), Real-time search

- Information Retrieval (IR) is about the process of providing answers to client's information needs. It is thus concerned with the collection, representation, storage, organization, accessing, manipulation and display, of the information items necessary to satisfying those needs.
- **Definition:** Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections.
- There is a huge quantity of text, audio, video and other documents available on the Internet, on about any subject. Users need to be able to find relevant information to satisfy their particular information needs.

- There are two ways of searching for information: To use a search engines or to browse directories organized by categories.
- IR is the task of representing, storing, organizing, and offering access to information items. IR is different from data retrieval, which is about finding precise data in databases with a given structure.
- In IR systems, the information is not structured. It is contained in free form in text (web pages or other documents) or in multimedia content.
- The first IR systems implemented in 1970's were designed to work with small collections of text. Some of these techniques are now used in search engines.
- The information retrieval techniques focusing on the challenges faced by search engine.
    1. One particular challenge is the large scale, given by the huge number of web-pages available on the Internet.
    2. The ambiguity of the natural language (English or other languages) that makes it difficult to have perfect matches between documents and user queries.
- Information retrieval is never an easy task. The problem with IR is that document representation, either by index terms or texts cannot satisfy user need representation, which is dynamic and complicated.
- Traditional IR systems are designed to support only one type of information-seeking strategy that users engage in: Query formulation.

1.3 IR PROBLEMS

Users of modern IR systems, such as search engine users, have information needs of varying complexity, An example of complex information need is as follows:Find all documents that address the role of the Federal Government in financing the operation of the National Railroad Transportation Corporation (AMTRAK)This full description of the user information need is not necessarily a good query to be submitted to the IR system Instead, the user might want to first translate this information need into a queryThis translation process yields a set of keywords, or index terms, which summarize the user information need . Given the user query, the key goal of the IR system is to retrieve information that is useful or relevant to the user.



1. **Document and Query indexing**
   How to best represent their contents?

2. **Query evaluation(or retrieval process)**
   To what extent does a document correspond to a query?

3. **System evaluation**
   How good is a system?

Three Big Issues in IR :

1.Relevance

- It is the fundamental concept in IR.
- A relevant document contains the information that a person was looking for when she submitted a query to the searchengine.
- There are many factors that go into a person's decision as to whether a document is relevant.
- These factors must be taken into account when designing algorithms for comparing text and ranking documents.
- Simply comparing the text of a query with the text of a document and looking for an exact match, as might be done in a database system produces very poor results in terms ofrelevance.
- To address the issue of relevance, **retrieval models** are used.
- A retrieval model is a formal representation of the process of matching
- a query and a document. It is the basis of the ranking algorithm that is used in a search engine to produce the ranked list of documents.
- A good retrieval model will find documents that are likely to be considered relevant by the person who submitted the query.
- The retrieval models used in IR typically model the statistical properties of text rather than the linguistic structure. For example, the ranking algorithms are concerned with the counts of word occurrences than whether the word is a noun or anadjective.

2.Evaluation

- Two of the evaluation measures are precision andrecall.
    - Precision is the proportion of retrieved documents that are relevant. Recall is the proportion of relevant documents that are retrieved.

Precision = $\frac{\text{Relevant documents} \cap \text{Retrieved documents}}{\text{Retrieved documents}}$

- Recall = $\frac{\text{Relevant documents} \cap \text{Retrieved documents}}{}$

Relevant documents

- When the recall measure is used, there is an assumption that all the relevant documents for a given query are known. Such an assumption is clearly problematic in a web search environment, but with smaller test collection of documents, this measure can be useful. It is not suitable for large volumes of log data.

3. Emphasis on users and their information needs

- The users of a search engine are the ultimate judges of quality. This has led to numerous studies on how people interact with search engines and in particular, to the development of techniques to help people express their information needs.
- Text queries are often poor descriptions of what the user actually wants compared to the request to a database system, such as for the balance of a bank account.
- Despite their lack of specificity, one-word queries are very common in web search. A one-word query such as "cats" could be a request for information on where to buy cats or for a description of the Cats (musical).
- Techniques such as query suggestion, query expansion and relevance feedback use interaction and context to refine the initial query in order to produce better ranked results.
- The figure summarizes the major issues involved in search engine design

1.4 USER TASK

The User Task.- The user of a retrieval system has to translate his information need into a query in the language provided by the system. With an information retrieval system, this normally implies specifying a set of words which convey the semantics of the information need.

1. Consider a user who seeks information on a topic of their interest :This user first translates their information need into a query, which requires specifying the words that compose the query In this case, we say that the user is searching or querying for information of their interest
2. Consider now a user who has an interest that is either poorly defined or inherently broad.

For instance, the user has an interest in car racing and wants to browse documents on Formula 1 and Formula Indy, In this case, we say that the user is browsing or navigating the documents of the collection

- The user of a retrieval system has to translate his information need into a query in the language provided by the system. With an information retrieval system, this normally implies specifying a set of words which convey the Semantics of the information need.
- With a data retrieval system, a query expression is used to convey the constraints that must be satisfied by objects in the answer set. In both cases, we say that the user searches for useful information executing a retrieval task. Fig. 1.2.1 shows Interaction of the user with

the retrieval system.

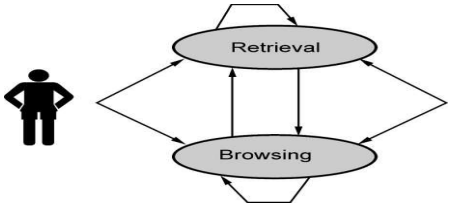
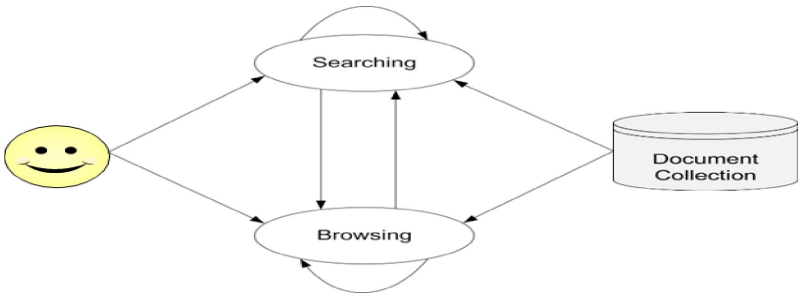


Fig. Interaction of the user with the retrieval system

- Suppose the user may be interested in web site about healthcare product. In this situation, the user might use an interactive interface to simply look around in the collection for documents related to healthcare product.
- User may be interested in new beauty product, weight loss or gain product. Here user is browsing the documents in the collection, not searching. It is still a process of retrieving information, but one whose main objectives are not clearly defined in the beginning and whose purpose might change during the interaction with the system.
- Pull technology: User requests information in an interactive manner. It performs three retrieval tasks, i.e. browsing (hypertext), Retrieval (classical IR systems) and Browsing and retrieval (modern digital libraries and web systems).
- Push technology: Automatic and permanent pushing of information to user. It acts like a software agents.

The general objective of an **Information Retrieval System** is to minimize the time it takes for a user to locate the **information** they need. The goal is to provide the **information** needed to satisfy the user's question. Satisfaction does not necessarily mean finding all **information** on a particular issue.



Information versus Data Retrieval

Information Retrieval Vs information Extraction

Information Retrieval:

Given a set of terms and a set of document terms select only most relevant document (precision) ,and preferably all the relevant ones(recall)

Infromation Extraction:

Extract from the text what the document means.Data retrieval**:** the task of determining which documents of a collection contain the keywords in the user query

Data retrieval system

Ex: relational databases

Deals with data that has a well defined structure and semantics.

Data retrieval does not solve the problem of retrieving information about a subject or topic.

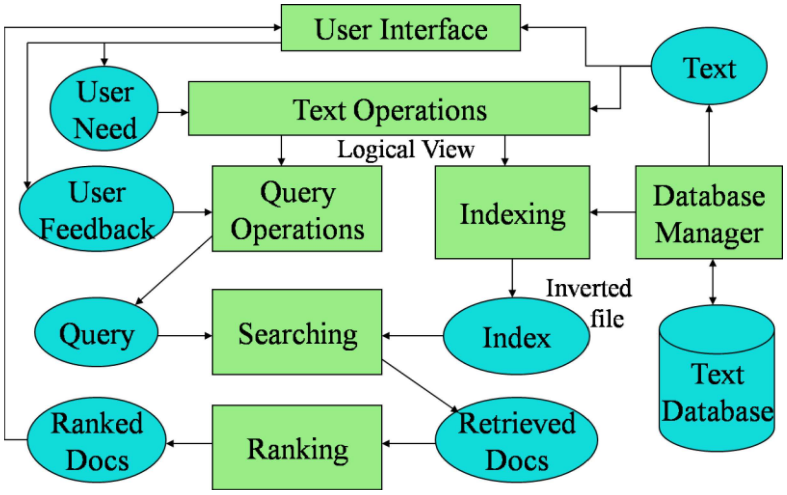
- An information retrieval system is software that has the features and functions required to manipulate "information" items versus a DBMS that is optimized to handle "structured" data.
- Information retrieval and Data Retrieval (DR) are often viewed as two mutually exclusive means to perform different tasks, IR being used for finding relevant documents among a collection of unstructured/semi-structured documents.
- Data retrieval being used for finding exact matches using stringent queries on structured data, often in a Relational Database Management System (RDBMS).
- IR is used for assessing human interests, i.e., IR selects and ranks documents based on the likelihood of relevance to the user's needs. DR is different; answers to users' queries are exact matches which do not impose any ranking.
- Data retrieval involves the selection of a fixed set of data based on a well-defined query. Information retrieval involves the retrieval of documents of natural language.
- IR systems do not support transactional updates whereas database systems support structured data, with schemas that define the data organization. IR systems deal with some querying issues not generally addressed by database systems and approximate searching by keywords.

Difference between Data Retrieval and Information Retrieval

Parameters	Data retrieval	Information retrieval
Example	Data base query	WWW search
Matching	Exact	Partial match Best match
Inference	Deduction	Induction
Model	Deterministic	Probabilistic
Query language	Artificial	Natural
Query specification	Complete	Incomplete

Items wanted	Matching	Relevant
Error response	Sensitive	Insensitive
Classification	Monotonic	Polytechnic

1.5 The IR System



COMPONENTS OF IR:

The above figure shows the architecture of IR System with the Specified Components

Components:

1 Text operations
2 Indexing
3 Searching
4 Ranking
5 User Interface
6 Query operations

Text operation:
Text Operations forms index words (tokens).
Stop word removal , Stemming

Indexing:
Indexing constructs an inverted index of word to document pointers.

Searching:
Searchingretrievesdocumentsthat containagivenquerytokenfromtheinverted index.

Ranking :
Ranking scores all retrieved documents according to a relevance metric.

User Interface:

User Interface manages interaction with the user:

- Query input and documentoutput.
- Relevance feedback.
- Visualization ofresults.

Query Operations:

Query Operations transform the query to improve retrieval:

- Query expansion using a thesaurus.
- Querytransformation using relevance feedback.

- An information retrieval system is an information system, which is used to store items of information that need to be processed, searched, retrieved, and disseminated to various user populations.
- Information retrieval is the process of searching some collection of documents, in order to identify those documents which deal with a particular subject. Any system that is designed to facilitate this literature searching may legitimately be called an information retrieval system.
- Conceptually, IR is the study of finding needed information. It helps users to find information that matches their information needs. Historically, IR is about document retrieval, emphasizing document as the basic unit.
- Information retrieval locates relevant documents, on the basis of user input such as keywords or example documents, for example: Find documents containing the words "database systems".
- Fig. shows information retrieval system block diagram. It consists of three components: **Input, processor and output**.

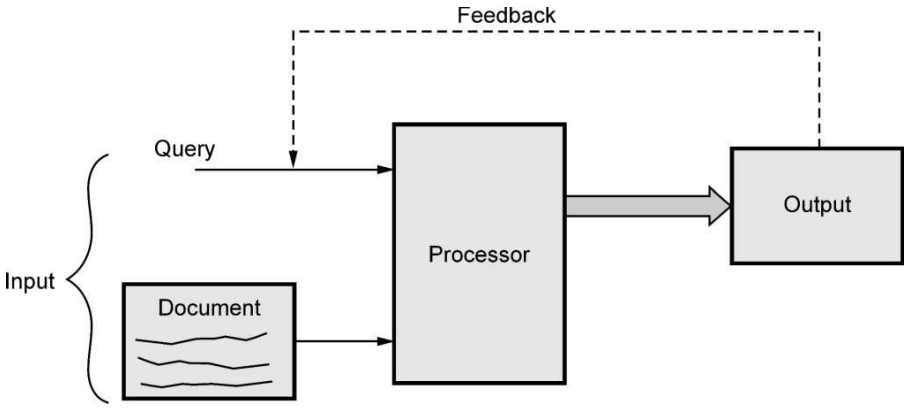


Fig : IR block diagram

a) **Input :** Store only a representation of the document or query which means that the text of a document is lost once it has been processed for the purpose of generating its representation.

b) A **document representative** could be a list of extracted words considered to be significant.

c) **Processor:** Involve in performing actual retrieval function, executing the search strategy

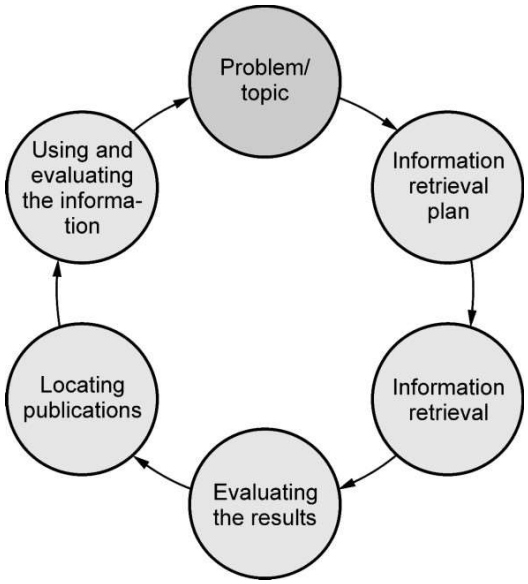
in response to a query.

d) **Feedback:** Improving the subsequent run after sample retrieval.

e) **Output:** A set of document numbers.

Process of Information Retrieval

- Information retrieval is often a continuous process during which you will consider, reconsider and refine your research problem, use various different information resources, information retrieval techniques and library services and evaluate the information you find.
- Fig. shows that the stages follow each other during the process, but in reality they are often active simultaneously and you usually will repeat some stages during the same information retrieval process.



Stages of IR process

- The different stages of the information retrieval process are :

1. **Problem / Topic :** An information need occurs when more information is required to solve a problem.
2. **Information retrieval plan :** Define your information need and choose your informationresources, retrieval techniques and search terms
3. **Information retrieval :** Perform your planned information retrieval (information retrievaltechniques)
4. **Evaluating the results :** Evaluate the results of your information retrieval (number andrelevance of search results)
5. **Locating publications :** Find out where and how the required publication, e.g. article,can be acquired
6. **Using and evaluating the information :** Evaluate the final results of the process (criticaland ethical evaluation of the information and information resources)

1.6 The Software Architecture of the IR System

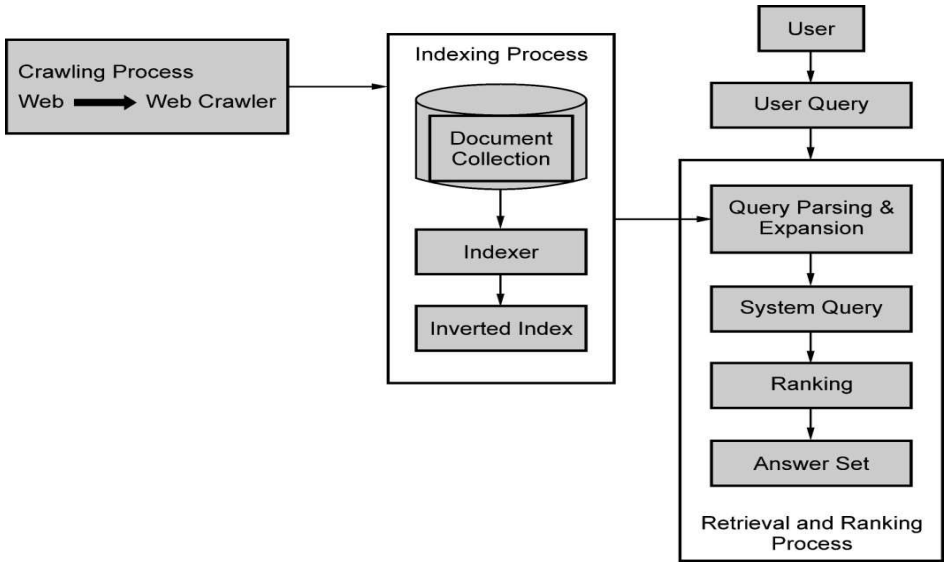


Fig. Architecture of IR system

The Process of retrieving information :

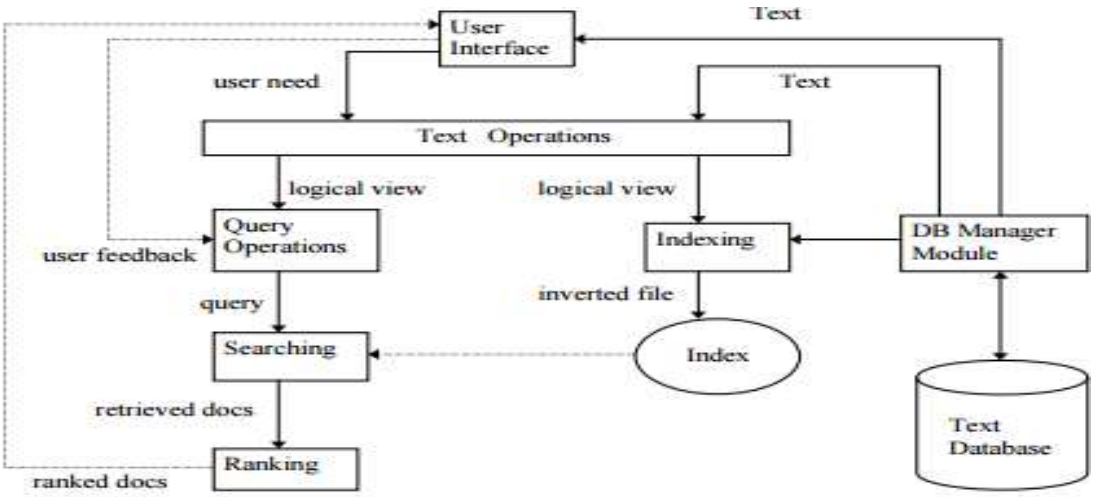


Figure . The process of retrieving information

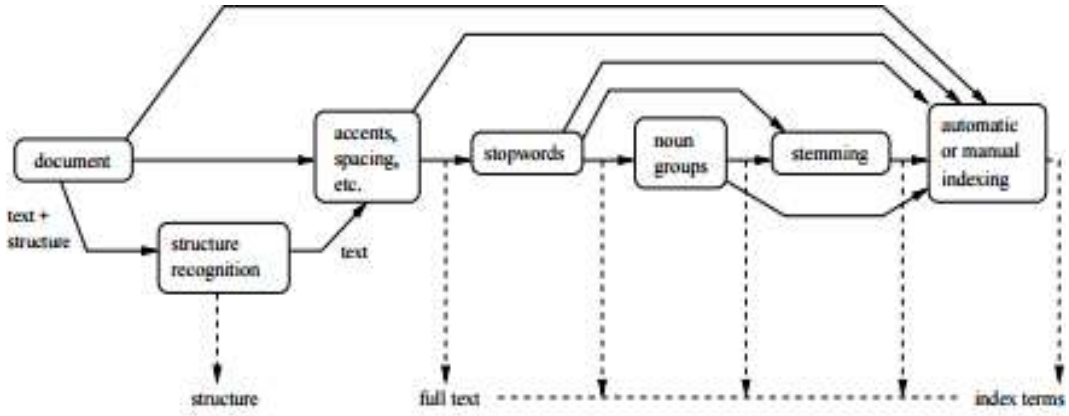
- The user's query is processed by a search engine, which may be running on the user's local machine, on a large cluster of machines in a remote geographic location, or anywhere in between.
- A major task of a search engine is to maintain and manipulate an inverted index for a document collection. This index forms the principal data structure used by the engine for searching and relevance ranking.

- As its basic function, an inverted index provides a mapping between terms and the locationsin the collection in which they occur.
- To support relevance ranking algorithms, the search engine maintains collection statistics associated with the index, such as the number of documents containing each term and the length of each document.
- In addition the search engine usually has access to the original content of the documents in order to report meaningful results back to the user.
- Using the inverted index, collection statistics, and other data, the search engine accepts queries from its users, processes these queries, and returns ranked lists of results.
- To perform relevance ranking, the search engine computes a score, sometimes called a Retrieval Status Value (RSV), for each document. After sorting documents according to their scores, the result list must be subjected to further processing, such as the removal of duplicate or redundant results.
- For example, a web search engine might report only one or results from a single host or domain, eliminating the others in favor of pages from different sources.

Logical View of the Documents:

Due to historical reasons, documents in a collection are frequently represented through a set of index terms or keywords. Such keywords might be extracted directly from the text of the document or might be specified by a human subject (as frequently done in the information sciences arena). No matter whether these representative keywords are derived automatically

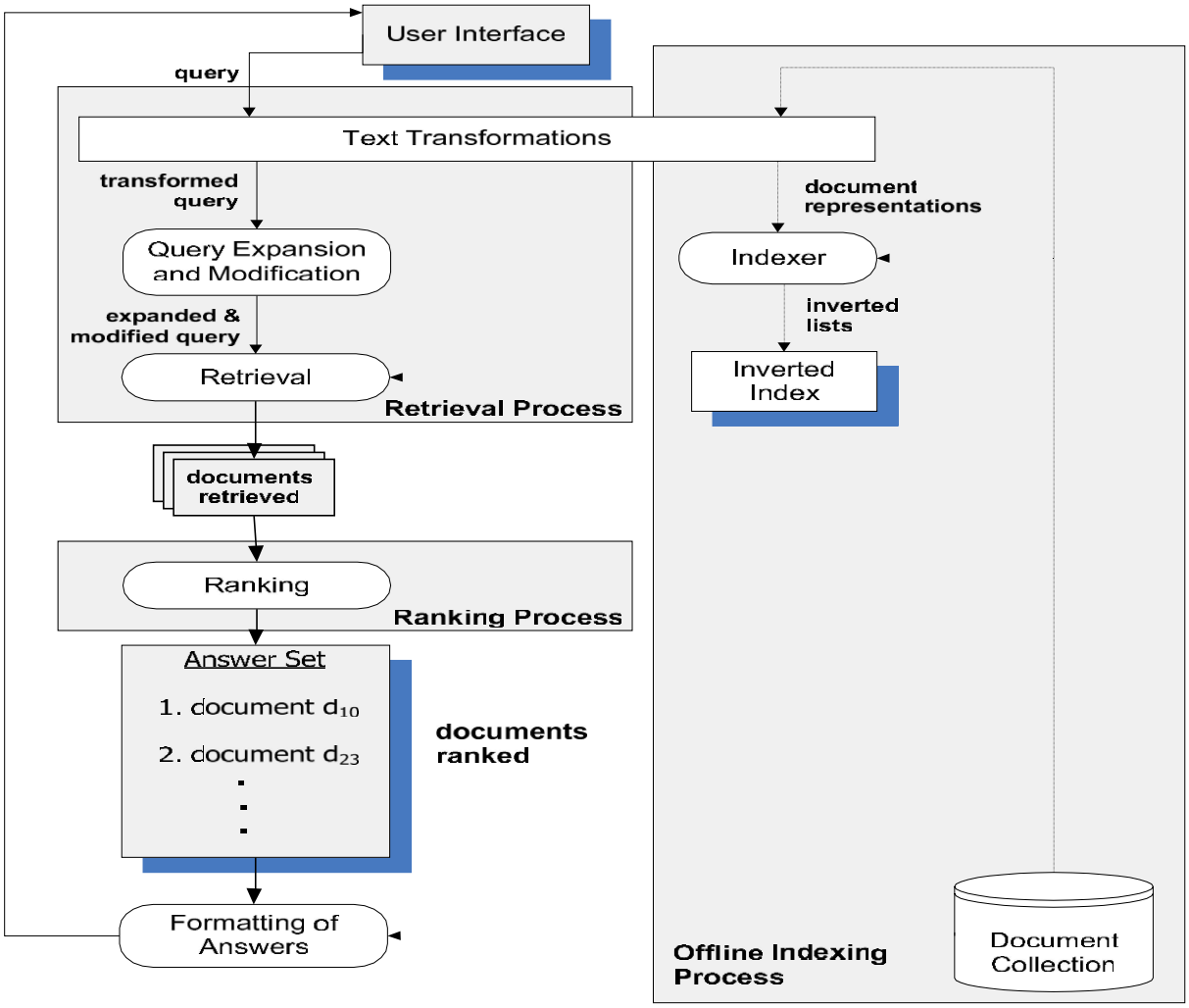
The processes of indexing, retrieval, and ranking



Logical view of a document: from full text to a set of index terms.

- A good retrieval model will find documents that are likely to be considered relevant by the person who submitted the query. Some retrieval models focus on topical relevance, but a search engine deployed in a real environment must use ranking algorithms that incorporates user relevance.

- Relevancy ranking is the method that is used to order the results list in such a way that the records most likely to be of interest to a user will be at the front. This makes searching easier for users as they will not have to spend as much time looking through records for the information that interests them.
- Each relevancy ranking algorithm slightly biases one type of data over another. While most any of the relevancy ranking algorithms will make a large difference, it is sometimes worthwhile trying several of the ranking methods. This way, you will be able to find the algorithm which most closely reflects the needs of your application as well as you and your user's expectations.
- There are a number of ways of calculating how a given record ranks and the factors that are taken into consideration vary with each technique.
  a) The number of times the search term occurs within a given record.
  b) The number of times the search term occurs across the collection of records.
  c) The number of words within a record.
  d) The frequencies of words within a record.
  e) The number of records in the index.
- Typically, relevancy ranking algorithms rank records in relation to each other. The weight assigned to a given record is a weight that reflects the weight of the record in relation to other records within the same database and for the same query.



The Retrieval Process :

To describe the retrieval process, we use a simple and generic software architecture as shown in Figure First of all, before the retrieval process can even be initiated, it is necessary to dene the text database. This is usually done by the manager of the database, which species the following: (a) the documents to be used, (b) the operations to be performed on the text, and (c) the text model (i.e., the text structure and what elements can be retrieved). The text operations transform the original documents and generate a logical view of them.

Once the logical view of the documents is defined, the database manager (using the DB Manager Module) builds an index of the text. An index is a critical data structure because it allows fast searching over large volumes of data. Different index structures might be used, but the most popular one is the inverted index as indicated in Figure. The resources (time and storage space) spent on defining the text database and building the index are amortized by querying the retrieval system many times. Given that the document database is indexed, the retrieval process can be initiated. The user need which is then parsed and transformed by the same text operations applied to the text. Then, query operations might be applied before the actual query, which provides a

system representation for the user need, is generated. The query is then processed to obtain the retrieved documents. Fast query processing is made possible by the index structure previously built. Before been sent to the user, the retrieved documents are ranked according to a likelihood of relevance. The user then examines the set of ranked documents in the search for useful information. At this point, he might pinpoint a subset of the documents seen as definitely of interest and initiate a user feedback cycle. In such a cycle, the system uses the documents selected by the user to change the query formulation. Hopefully, this modified query is a better representation of the real user need.

1.7 THE WEB

- World Wide Web is collection of millions of files stored on thousands of servers all over the world. These files represent documents, pictures, video, sounds, programs, interactive environments.
- A web page is an HTML document that is stored on a web server. A web site is a collectionof web pages belonging to a particular organization.
- URL of these pages share a common prefix, which is the address of the home page of the size. Search engines are a bottom-up approach for finding your way around the web. Some search engines search only the titles of web pages. While other search every word. Keywords can be combined with Boolean operations, such as AND, OR and NOT, to produce rather complicated queries.

With the rapid growth of the Internet, more information is available on the Web and Web information retrieval presents additional technical challenges when compared to classic information retrievaldue to the heterogeneity and sizeof the web.

1. Web information retrieval is unique due to the dynamism, variety of languages used, duplication, high linkage,illformed query and wide variance in the nature of users.
2. Another issue is the rising number of inexperienced users, they the majority of Web users are not very sophisticated searchers.
3. Many software tools are available for web information retrieval such as search engines (Google and Alta vista), hierarchical directories (Yahoo), many other software agents and collaborative filtering systems. The commonly cited problems in search engines are the slow speed of retrieval, communication delays, and poor quality of retrieved results.

In addition, in the current scenario, multimedia information is increasingly becoming available on the Web and modem IR systems must be capable of tackling not only the text but also multimedia information like sound, image and video

IR on the web Vs.IR

Traditional IR systems normally index a closed collection of documents, which are mainly text-based and usually offer little linkage between documents. Traditional IR systems are often

referred to as full-text retrieval systems. Libraries were among the first to adopt IR to index their catalogsand later, to search through information which was typically imprinted onto CD- ROMs. The main aim of traditional IR was to return relevant documents that satisfy the user's information need. Although the main goal of satisfying the user's need is still the central issue in web IR (or web search), there are some very specific challenges that web search poses that have required new and innovative solutions.

- The first important difference is the scale of web search, as we have seen that the current size of the web is approximately 600 billion pages. This is well beyond the size of traditional documentcollections.
- The Web is dynamic in a way that was unimaginable to traditional IR in terms of its rate of change and the different types of web pages ranging from static types (HTML, portable document format (PDF), DOC, Postscript, XLS) to a growing number dynamic pages written in scripting languages such a JSP, PHP or Flash. We also mention that a large number of images, videos, and a growing number of programs are delivered through the Web to our browsers.
- The Web also contains an enormous amount of duplication, estimated at about 30%. Such redundancy is not present in traditional corpora and makes the search engine's task even moredifficult.
- The quality of web pages vary dramatically; for example, some web sites create web pages with the sole intention of manipulating the search engine's ranking, documents may contain misleading information, the information on some pages is just out of date, and the overall quality of a web page may be poor in terms of its use of language and the amount of useful information it contains. The issue of quality is of prime importance to web search engines as they would very quickly lose their audience if, in the top- ranked positions, they presented to users poor qualitypages.
- The range of topics covered on the Web is completely open, as opposed to the closed collections indexed by traditional IR systems, where the topics such as in library catalogues, are much better defined and constrained.
- Another aspect of the Web is that it is globally distributed. This poses serious logistic problems to search engines in building their indexes, and moreover, in delivering a service that is being used from all over the globe. The sheer size of the problem is daunting, considering that users will not tolerate anything but an immediate response to their query. Users also vary in their level of expertise, interests, information- seeking tasks, the language(s) they understand, and in many other ways.

	Classical IR	**Web IR**
Volume	Large	Huge
Data Quality	Clean,No Duplicates	Noisy, duplicates Available
Data change rate	Infrequent	In flux
Data accessibility	Accessible	Partially accessible

Format diversity	Homogeneous	Widely Diverse
Documents	Text	HTML
No.of Matches	Small	Large
IR techniques	Content based	Link based

COMPARISION :

S. No	Differentiator	Web Search	IR
1	Languages	Documents in many different languages. Usually search engines use full text indexing; no additional subject analysis.	Databases usually cover only one language or indexing of documents written in different languages with the same vocabulary.
2	File types	Several file types, some hard to index because of a lack of textual information.	Usually all indexed documents have the same format (e.g. PDF) or only bibliographic information is provided.
3	Document length	Wide range from very short to very long. Longer documents are often divided into parts.	Document length varies, but not to such a high degree as with the Web documents
4	Document structure	HTML documents are semi structures.	Structured documents allow complex field searching
5	Spam	Search engines have to decide which documents are suitable for indexing.	Suitable document typesare defined in the process of database design.
6	Amount of data, size of databases	The actual size of the Web is unknown. Complete indexing of the whole Web is impossible.	Exact amount of data can be determined when using formal criteria.

7	Type of queries	Users have little knowledge how to search; very short queries (2-3 words).	Users know the retrieval language; longer, exact queries.
8	User interface	Easy to use interfaces suitable for laypersons.	Normally complex interfaces; practice needed to conduct searches.
9	Ranking	Due to the large amount of hits relevance ranking is the norm.	Relevance ranking is often not needed because the users know how to constrain the amount of hits.
10	Search functions	Limited possibilities.	Complex query languages allow narrowing searches.

1.8 E-PUBLISHING-ERA

1. E-publishing refers to a publishing process where the manuscript are submitted in E-format, edited, printed and even distributed to users in E-form by computer and communication technology, which may be online, CD-ROM, Networks etc. It involves the storage of information in electronic or digital form. It also refers to a type of publishing that does not include printed books.
2. E-publishing has been defining as any non-print media material that is published in digitized form to an identifiable public. The media in electronic publishing can be text, numeric, graphic, still or motion pictures, video, sound or as infrequently the case a combination of any or all of these.

- There are four main reasons for the development of e-publishing,
  a) Rapid development and wide use of computer technology.
  b) The tremendous growth of computer networks.
  c) Merging of computer and telecommunication technology.
  d) Development of information industry.

Since its inception, the Web became a huge success - Well over 20 billion pages are now available and accessible in the Web More than one fourth of humanity now access the Web on a regular basis.

Why is the Web such a success? What is the single most important characteristic of the Web that makes it so revolutionary?

In search for an answer, let us dwell into the life of a writer who lived at the end of the 18th Century.

I. She finished the first draft of her novel in 1796 The first attempt of publication was refused without a reading. The novel was only published 15 years later! She got a flat fee of $110, which meant that she was not paid anything for the many subsequent editions Further, her authorship was anonymized under the reference “By a Lady”

II. Pride and Prejudice is the second or third best loved novel in the UK ever, after The Lord of the Rings and Harry Potter. It has been the subject of six TV series and five film versions The last of these, starring Keira Knightley and Matthew Macfadyen, grossed over 100 million dollars

III. Jane Austen published anonymously her entire life Throughout the 20th century, her novels have never been out of print, Jane Austen was discriminated because there was no freedom to publish in the beginning of the 19th century .

IV. The Web, unleashed by the inventiveness of Tim Berners-Lee, changed this once and for all It did so by universalizing freedom to publish - The Web moved mankind into a new era, into a new time, into The e-Publishing Era.

The term "**electronic publishing**" is primarily used in the 2010s to refer to online and web-based **publishers**, the term has a history of being used to describe the development of new forms of production, distribution, and user interaction in regard to computer-based production of text and other interactive media.

The first digitization projects were transferring physical content into digital content. Electronic publishing is aiming to integrate the whole process of editing and publishing (production, layout, publication) in the digital world.

The traditional publishing, and especially the creation part, were first revolutionized by new desktop publishing softwares appearing in the 1980s, and by the text databases created for the encyclopedias and directories. At the same time the multimedia was developing quickly, combining book, audiovisual and computer science characteristics. CDs and DVDs appear, permitting the visualization of these dictionaries and encyclopedias on computers.

The arrival and democratization of Internet is slowly giving small publishing houses the opportunity to publish their books directly online. Some websites, like Amazon, let their users buy eBooks; Internet users can also find many educative platforms (free or not), encyclopedic websites like Wikipedia, and even digital magazines platforms. The eBook then becomes more and more accessible through many different supports, like the e-reader and even smartphones. The digital book had, and still has, an important impact on publishing houses and their economical models; it is still a moving domain, and they yet have to master the new ways of publishing in a digital era.

1.9 HOW THE WEB CHANGED SEARCH

- The web has introduced millions of people to search. The information retrieval community stands ready to suggest helpful strategies for finding information on the Web.
- Let us consider the impact of web on search engine:

1. Characteristics of the document collection itself

2. Size of the collection and volume of user queries
3. Vast size of the document collection
4. Web adverting

- Search has changed dramatically over the past year and semantic technology has been at the centre of it all. Consumers increasingly expect search engines to understand natural language and perceive the intent behind the words they type in, and search engine algorithms are rising to this challenge.

Web search is today the most prominent application of IR and its techniques—the ranking and indexing components of any search engine are fundamentally IR pieces of technology.

The first major impact of the Web on search is related to the characteristics of the document collection itself

- The Web is composed of pages distributed over millions of sites and connected through hyperlinks
- This requires collecting all documents and storing copies of them in a central repository, prior to indexing
- This new phase in the IR process, introduced by the Web, is called crawling

The second major impact of the Web on search is related to:

- The size of the collection
- The volume of user queries submitted on a daily basis
- As a consequence, performance and scalability have become critical characteristics of the IR system

The third major impact: in a very large collection, predicting relevance is much harder than before

- Fortunately, the Web also includes new sources of evidence
- Ex: hyperlinks and user clicks in documents in the answer set

The fourth major impact derives from the fact that the Web is also a medium to do business

- Search problem has been extended beyond the seeking of text information to also encompass other user needs
- Ex: the price of a book, the phone number of a hotel, the link for downloading a software

The fifth major impact of the Web on search is Web spam

- Web spam: abusive availability ofcommercial information disguised in the form of informational content
- This difficulty is so large that today we talk of Adversarial Web Retrieval

1.10 PRACTICAL ISSUES IN THE WEB

- **Security:**

Commercial transations over the Internet are not yet a completely safe procedure

- **Privacy:**

Frequently, people are willing to exchange information as long as it does not become

public

- **Copyright and patent rights:**

It is far from clear how the wide spread of data on the Web affects copyright and patent laws in the various countries

- **Log In Issue:**

One of the most common problems faced by online businesses is the inability to log in to the control panel. You need easy access to the control panel for additions and deletions of content and for other purposes.

- **Frequent Technical Breakdown:**

Running a website business effectively is only possible when all the functional parameters respond to your input quickly and smoothly. Unfortunately, most of the times, this does not happen.

- **Slow Performance of Web Server:**

Slow web server is one of the biggest headaches that businesses have to deal with. When your customers encounter pages that load slowly, they tend to abandon their search and look for other alternatives.

- **Server Limitations:**

A few hosting companies follow the undesirable business practice of not disclosing their limit in terms of space and bandwidth. They try to serve more customers with their limited resources which can result in major performance issues in the long term

1.11HOW PEOPLE SEARCH

User interaction with search interfaces differs depending on

- The type of task
- The domain expertise of the information seeker
- The amount of time and effort available to invest in the process

Marchionini makes a distinction between information lookup and exploratory search

Information lookup tasks

1. Are a kin to fact retrieval or question answering
2. Can be satisfied by discrete pieces of information: numbers, dates, names, or Web sites
3. Can work well for standard Web search interactions

Exploratory search is divided into learning and investigating tasks Learning search

1. Requires more than single query-response pairs
2. Requires the searcher to spend time
3. Scanning and reading multiple information items
4. Synthesizing content to form new understanding

Investigating refers to a longer-term process which

- Involves multiple iterations that take place over perhaps very long periods of time
- May return results that are critically assessed before being integrated into personaland professional knowledge bases
- May be concerned with finding a large proportion of the relevant information available

Classic × Dynamic Model

Classic notion of the information seeking process:

1. **Problem identification**
2. **Articulation of information need(s)**
3. **Query formulation**
4. **Results evaluation**

More recent models emphasize the dynamic nature of the search process

- The users learn as they search
- Their information needs adjust as they see retrieval results and other document surrogates This dynamic process is sometimes referred to as theberry picking model of search.

Navigation × Search

Navigation: the searcher looks at an information structure and browses among the available information

This browsing strategy is preferrable when the information structure is well-matched to the user's information need

- It is mentally less taxing to recognize a piece of information than it is to recall it
- It works well only so long as appropriate links are available

If the links are not avaliable, then the browsing experience might be frustrating

Search Process

- Numerous studies have been made of people engaged in the search process
- The results of these studies can help guide the design of search interfaces

- One common observation is that users often reformulate their queries with slight modifications
- Another is that searchers often search for information that they have previously accessed The users' search strategies differ when searching over previously seen materials
- Researchers have developed search interfaces support both query history and revisitation
- Studies also show that it is difficult for people to determine whether or not a document is relevant to a topic,Other studies found that searchers tend to look at only the top-ranked retrieved results.Further, they are biased towards thinking the top one or two results are better than those beneath them.
- Studies also show that people are poor at estimating how much of the relevant material they have found ,Other studies have assessed the effects of knowledge of the search process itself.
- These studies have observed that experts use different strategies than novices searchers.

Information Lookup versus Exploratory Search

- Search activities are commonly divided into two broad categories: lookup and exploratory. Exploratory search is an increasingly important activity yet challenging for users.
- Lookup search is by far the better understood and assumed to have precise search goals. The predominant design goal in information retrieval systems has been fast and accurate completion of lookup searches.
- Exploratory search is presently thought to center around the acquisition of new knowledge and considered to be challenging for the user.
- Lookup is the most basic kind of search task and has been the focus of development for database management systems and much of what Web search engines support.
- Lookup tasks return discrete and well-structured objects such as numbers, names, short statements, or specific files of text or other media.
- Database management systems support fast and accurate data lookups in business and industry; in journalism, lookups are related to questions of who, when, and where as opposed to what, how, and why questions.
- In libraries, lookups have been called "known item" searches to distinguish them from subject or topical searches.
- A typical example would be a user wanting to make a reservation to a restaurant and looking for the phone number on the Web.
- On the other hand, exploratory search is described as open-ended, with an unclear information need, an ill-structured problem of search with multiple targets. This search activity is evolving and can occur over time.
- For example, a user wants to know more about Senegal, she doesn't really know what kind of information she wants or what she will discover in this search session; she only knowsshe wants to learn more about that topic.

1.12 SEARCH INTERFACES TODAY

- The job of the search user interface is to aid users in the expression of their information needs, in the formulation of their queries, in the understanding of their search results, and inkeeping track of the progress of their information seeking efforts.
- The typical search interface today is of the form: type-keywords-in-entry-form, view-results-in-a-vertical-list.
- Some important reasons for the relative simplicity and unchanging nature of the standard Web search interface are :

  a) Search is a means towards some other end, rather than a goal in itself. When a person is looking for information, they are usually engaged in some larger task, and do not want their flow of thought interrupted by an intrusive interface.

  b) Search is a mentally intensive task. When a person reads text, they are focused on that task; it is not possible to read and to think about something else at the same time. Thus, the fewer distractions while reading, the more usable the interface.

  c) Since nearly everyone who uses the Web uses search, the interface design must be understandable and appealing to a wide variety of users of all ages, cultures and backgrounds, applied to an enormous variety of information needs.

How does an information seeking session begin in online information systems?

- The most common way is to use a Web search engine
- Another method is to select a Web site from a personal collection of already-visited sites
- Online bookmark systems are popular among a smaller segment of users Ex: Delicious.com
- Web directories are also used as a common starting point, but have been largely replaced by search engines

The primary methods for a searcher to express their information need are either

a) entering words into a search entry form
b) selecting links from a directory or other information organization display

For Web search engines, the query is specified in textual form. Typically, Web queries today are very short consisting of one to three words

Query Specification

- The query specification process is :

1. The kind of information the searcher supplies. Query specification input spans a spectrum from full natural language sentences, to keywords and key phrases, to syntax- heavy command language-based queries.
2. The interface mechanism the user interacts with to supply this information. These

include command line interfaces, graphical entry form-based interfaces, and interfaces for navigating links.

- Queries over collections of textual information usually take on a textual form. Keyword queries consist of a list of one or more words or phrases -- rather than full natural language statements.
- **Example:** English keyword queries include flip cam, fresh chilli paste recipes, and video game addiction. Some keyword queries consist of lists of different words and phrases, which together suggest a topic.
- Many others are noun compounds and proper nouns. Less frequently, keyword queries contain syntactic fragments including prepositions and verbs and in some cases, full syntactic phrases.
- Dynamic query term suggestions can be provided as the user types in a term before they view the results or it can be presented following the result display stage.
- It is interesting to note that the performance of query term suggestions across the three search engines is varied in terms of the number of suggestions and how they handle single word and multi-word queries. Following table provides a comparative overview of the number of suggested query terms for TREC topics.

Search engine	Average number of words suggested	Median number of words suggested
Google	4	4
Yahoo!	6.46	10
Bing	6.18	8

- All the three search engines offer spelling error correction features and around 80% of the time they provide 4 or more dynamic query suggestions.

Short queries reflect the standard usage scenario in which the user *tests the waters*:

- If the results do not look relevant, then the user reformulates their query
- If the results are promising, then the user navigates to the most relevant-looking web site

Query Specification Interface

The standard interface for a textual query is a **search box entry form**

Studies suggest a relationship between query length and the width of the entry form

- Results found that either small forms discourage long queries or wide forms encourage longer queries

- Some entry forms are followed by a form that filters the query in some way , For instance, at yelp.com, the user can refine the search by location using a second form

- Notice that the yelp.com form also shows the user's home location, if it has been specified previously
- Some search forms show hints on what kind of information should be entered into each form, For instance, in zvents.com search, the first box is labeled "what are you looking for"?

- Some interfaces show a list of query suggestions as the user types the query - this is referred to as **auto-complete, auto-suggest**, or **dynamic query suggestions**

- Dynamic query suggestions, from Netflix.com

- Dynamic query suggestions, grouped by type, from NextBio.com

RETRIVAL RESULTS DISPLAY :

- When displaying search results, either the documents must be shown in full or else the searcher must be presented with some kind of representation of the content of those documents.
- The documents surrogate refers to the information that summarizes the document.
- The appearance of search engine results pages is constantly in flux due to experiments conducted by Google, Bing, and other search engine providers to offer their users a more intuitive, responsive experience.
- The quality of the surrogate can greatly effect the perceived relevance of the search results listing. In Web search, the page title is usually shown prominently along with the URL and sometimes other metadata.
- The user enters their search query, upon which the search engine presents them with a SERP. Every SERP is unique, even for search queries performed on the same search engineusing the same keywords or search queries.
- This is because virtually all search engines customize the experience for their users by presenting results based on a wide range of factors beyond their search terms, such as the user's physical location, browsing history, and social settings. Two SERPs may appear identical, and contain many of the same results, but will often feature subtle differences.
- A deep link is a hypertext link to a page on a website other than its homepage. Deep links are often used to link directly to products of an online store to or appropriate content.

- Google itself uses deep links in the form of rich snippets or site links. A hyperlink thatpoints to a deeper level of a domain can also be useful for link hubs, lists of topics or in citations. Again, the user's interest is in the foreground.
- Price comparison portals also work with deep links. In this case, this type of link is necessary because the potential customer would want to find and buy the exact product heis comparing.
- For example, a query on a term like "rainbow" may return sample images as one entry in the results listing
- A query on the name of a sports team might retrieve the latest game scores and a link to buy tickets

QUERY REFORMULATION :

- After a query is specified and results have been produced, a number of tools exist to help the user reformulate their query.
- Query formulation is an essential part of successful information retrieval. The challenges in formulating effective queries are emphasized in web information search, because the web isused by a diverse population varying in their levels of expertise.
- Query formulation is the stage of the interactive information access process in which user translates an information need into a query and submits the query to an information access system such as a search engine.
- The system performs some computation to match the query with the documents most likely to be relevant to the query and returns a ranked list of relevant documents to the user.

There are tools to help users reformulate their query

- One technique consists of showing terms related to the query or to the documents retrieved in response to the query
- A special case of this is spelling corrections or suggestions
- Usually only one suggested alternative is shown: clicking on that alternative re-executes the query
- In years back, the search results were shown using the purportedly incorrect spelling
- **Relevance feedback** is another method whose goal is to aid in query reformulation
- The main idea is to have the user indicate which documents are relevant to their query

- In some variations, users also indicate which terms extracted from those documents are relevant
- The system then computes a new query from this information and shows a new retrieval set

ORGANISING SEARCH RESULTS

Organizing results into meaningful groups can help users understand the results and decide what to do next

Popular methods for grouping search results: category systems and clustering

Category system: meaningful labels organized in such a way as to reflect the concepts relevant to a domain

The most commonly used category structures are **flat**, **hierarchical**, and **faceted** categories.Most Web sites organize their information into general categories

Clustering refers to the grouping of items according to some measure of similarity

It groups together documents that are similar to one another but different from the rest of the collection.

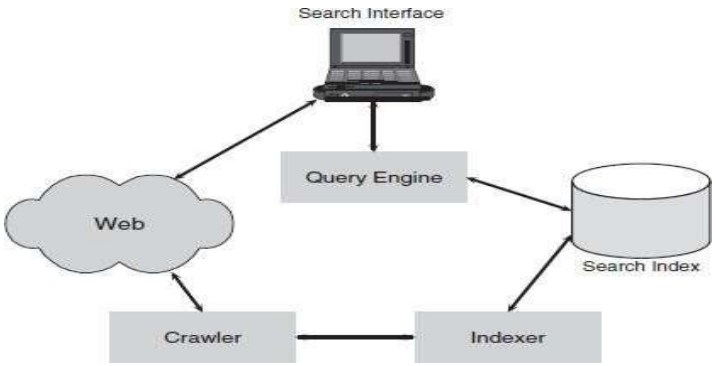
The greatest advantage of clustering- is that it is fully automatable

The disadvantages of clustering include-an unpredictability in the form and quality of results , the difficulty of labeling the groups

COMPONENTS OF SEARCH ENGINE

The main components of a search engine are

1 Crawler
2 Indexer
3 Search index
4 Query engine
5 Search interface.



- **Crawler:** A *web crawler* is a software program that traverses web pages, downloads

them for indexing, and follows the hyperlinks that are referenced on the downloaded pages; a web crawler is also known as a *spider*, a *wanderer* or a *software robot*.

- **Indexer:**The second component is the *indexer* which is responsible for creating the search index from the web pages it receives from the crawler
- **Search Index:**The *search index* is a data repository containing all the information the search engine needs to match and retrieve web pages. The type of data structure used to organize the index is known as an *inverted file*.
- **Query Engine:** The *query engine* is the algorithmic heart of the search engine. The inner working of a commercial query engine is a well-guarded secret, since search engines are rightly paranoid, fearing web sites who wish to increase their ranking by unfairly taking advantage of the algorithms the search engine uses to rank result pages.
- **Search Interface:** Once the query is processed, the query engine sends the results list to the *search interface*, which displays the results on the user's screen. The user interface provides the look and feel of the search engine, allowing the user to submit queries, browse the results list, and click on chosen web pages for further browsing.

1.13 VISUALIZATION IN SEARCH INTERFACES

Experimentation with visualization for search has been primarily applied in the following ways:

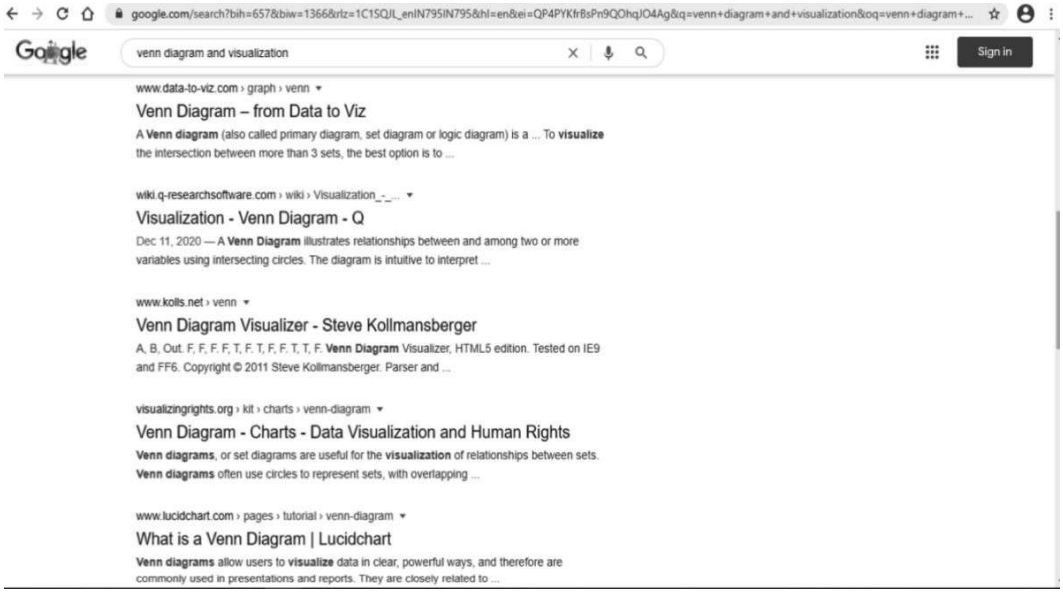
- Visualizing Boolean syntax
- Visualizing query terms within retrieval results
- Visualizing relationships among words and documents
- Visualization for text mining

Visualizing Boolean syntax

Boolean query syntax is difficult for most users and is rarely used in Web search ,For many years, researchers have experimented with how to visualize Boolean query specification. A common approach is to show Venn diagrams. A more flexible version of this idea was seen in the VQuery system, proposed by Steve Jones.

Visualizing Query Terms

Understanding the role of the query terms within the retrieved docs can help relevance assessment , Experimental visualizations have been designed that make this role more explicit. In the TileBars interface, for instance, documents are shown as horizontal glyphs ,the locations of the query term hits marked along the glyph. The user is encouraged to break the query into its different facets, with one concept per line. Then, the lines show the frequency of occurrence of query terms within each topic.



Words and Docs Relationships :

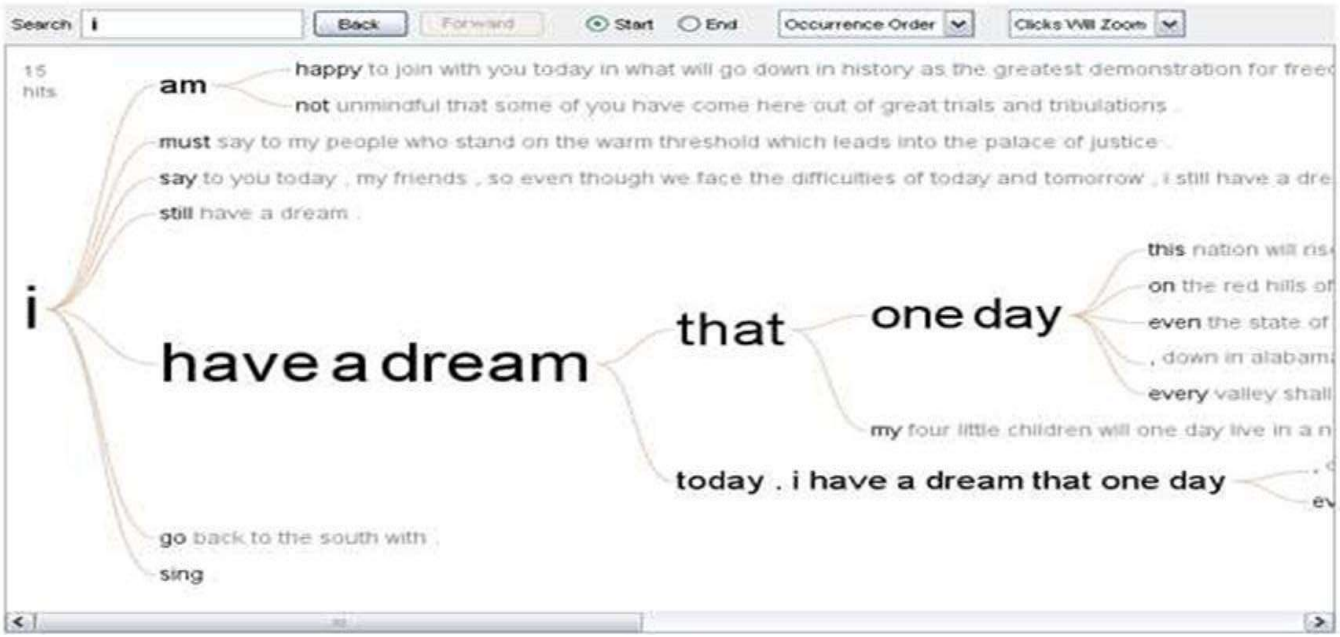
- Numerous works proposed variations on the idea of placing words and docs on a two-dimensional canvas
- In these works, proximity of glyphs represents semantic relationships among the terms or documents
- An early version of this idea is the VIBE interface
- Documents containing combinations of the query terms are placed midway between the icons representing those terms
- The Aduna Autofocus and the Lyberworld projects presented a 3D version of the ideas behind.
- Visualization developers suggest various idea of placing words and documents on a two-dimensional canvas, where proximity of glyphs represents semantic relationship among the terms or documents. Another method is to map documents or words from a very high-dimensional term space down into a two-dimensional plane and show where the documents or words fall within that plane using 2D or 3D.

Visualization for Text Mining

- Text mining is understood as a process of automatically extracting meaningful, useful, previously unknown and ultimately comprehensible information from textual document repositories. Text mining can be visualized as consisting of two phases: Text refining that transforms free-form text documents into a chosen intermediate form, and knowledge

distillation that deduces patterns or knowledge from the intermediate form.

- Visualization is also used for purposes of analysis and exploration of textual data
- Visualizations such as the Word Tree show a piece of a text concordance
- It allows the user to view which words and phrases commonly precede or follow a given word



- The Word Tree visualization, on Martin Luther King's , *I have a dream* speech, from Wattenberg *et al*
- Visualization is also used in search interfaces intended for analysts, an example is the TRIST information *triage* system, from Proulx *et al* ,In this system, search results is represented as document icons- thousands of documents can be viewed in one display.