UNIT 18 INFORMATION RETRIEVAL MODELS AND THEIR APPLICATIONS

Structure

18.0 OBJECTIVES

In the efforts for search for appropriate techniques of information retrieval, various models have been developed. This Unit deals with basics of information retrieval techniques and the different types of models developed from time to time. After reading this Unit, you will be able to:

- know the basics and types of factors involved in the information retrieval;
- know the shift from the conventional to modern information retrieval;
- understand the process of matching information need and retrieval of information from databases, knowledge bases, information systems and libraries;
- be conversant with the development of information retrieval models; and
- acquaint yourself with research areas in the field of information retrieval.

18.1 INTRODUCTION

Pattern of information retrieval indicates a knowledge seeking behaviour of individuals and groups of individuals. In this context, a searcher seeks some information from the vast store of a knowledge. An analysis and diagnosis of this state of mind provides guidelines for organisation of knowledge in libraries, information retrieval systems, databases, knowledge bases and similar environments. Such guidelines are aimed at providing conducive compatibility between searchers' approach and knowledge organisation in a database. The current human environment consisting of learning, problem-solving, and decision-making situation calls for flexibility in knowledge structures at each instance. The development in computer and communication technologies have made it possible to store vast amount of information in compact form. The variety of software developed have also given scope for quick retrieval of information from this store. While the speed of retrieval is valuable, it could be enriched further if the retrieved information is readily assimilable by the information seeker. It is in this context that modelling of retrieved information into user-friendly approaches calls for cognitive modelling of information retrieval.

Such development has given rise to a field 'Cognitive Science' which is an inter-disciplinary field drawing inputs from the fields of Psychology, Behavioural Studies, Computer Science, Artificial Intelligence and Information Science.

18.2 INFORMATION RETRIEVAL

Information Retrieval is a process of selecting information from a store. It connotes that search for information may be from documents, metadata which describes documents or searching interim databases. The databases may be standalone databases or hypertext networked databases like Intranet and Internet. It primarily helps a person who needs to get some information in his activities, be it research, problem solving, decision-making, production, service, etc. Broadly speaking, four kinds of information retrieval exist. *Document retrieval*, in which simple structured files are normally processed, using a small number of well-defined attributes to characterise each record, and a restricted set of pre-specified query types to access the database; *reference retrieval,* in which the records represent books, documents, and other library materials, and the number of different attributes available for the identification of the information items is effectively unlimited. In that case, the queries often refer to the information content of individual documents. In the most general case, a retrieval system might be designed to handle any kind of query and the system might furnish direct replies to such queries; in *fact retrieval,* a wide variety of different types of information identifiers may be needed, and the answers may have to be based not only on a deep analysis of each individual information item, but also on general world knowledge and other extraneous factors. In *text retrieval,* instead of retrieving reference or data or surrogates of documents, text on a particular topic are retrieved. Irrespective of any retrieval environment, the following four main system components must be taken into account in formulation of the retrieval problem.

a) **The objects, documents, or records themselves** (which in the aggregate constitute the information files to be processed);

b) **The information identifiers, terms, index terms, keywords, attributes**, etc. (which characterise the records or documents and represent the information content in each case);

c) **The information requests** (which enter into the system and are to be compared with the stored records for retrieval); and

d) **The relevance information** (often supplied by the users of the system connecting the information requests to the stored information items).

An Information Retrieval System is a system that is capable of storage, retrieval, and maintenance of information. Information in this context can be composed of text (including numeric and date data), images, audio, video and other multi-media objects. An Information Retrieval System thus consists of a software program that facilitates a user in finding the information of his needs. The system may use standard computer hardware or specialized hardware to support the search sub-function and to convert non-textual sources to a searchable media (e.g., transcription of audio to text). The success of an information system is gauged by how well it can minimize the overhead for a user to find the needed information. Overhead can be expressed as the time a user spends in all of the steps leading to reading an item containing the needed information (e.g., query generation, query execution, scanning results of query to select items to read, reading non-relevant items). The success of an information system is very subjective, based upon what information is needed and the willingness of a user to accept overhead. Under some circumstances, needed information can be defined as all information that is in the system that relates to a user's need. Thus, search composition, search execution, and reading non-relevant items are all aspects of information retrieval overhead.

The first information retrieval system originated with the need to organize information in libraries. Catalogues were created to facilitate the identification and retrieval of items. The term 'item' is used to represent the smallest complete unit that is processed and manipulated by the system. The definition of item varies by how a specific source treats information. A complete document, such as a book, newspaper or magazine could be an item. Each chapter, or article may also be defined as an item. As sources vary and systems include more complex processing, an item may address even lower levels of abstraction such as a contiguous passage of text or a paragraph.

With the advent of inexpensive powerful personal computer processing systems and high speed, large capacity secondary storage products, it has become commercially feasible to provide large textual information databases for the average user.

In information retrieval the term 'relevant' item is used to represent an item containing the needed information. From a user's perspective 'relevant' and 'needed' are synonymous. From a system perspective, information could be relevant to a search statement (i.e., matching the criteria of the search statement) even though it is not needed, relevant to user (e.g., the user already knew the information).

In addition to the complexities in generating a query, quite often the user is not an expert in the area that is being searched and lacks domain specific vocabulary that is unique to that particular subject area. The user starts the search process with a general concept of the information required, but does not have a focused definition of exactly what is needed. A limited knowledge of the vocabulary associated with a particular area along with lack of focus on exactly what information is needed leads to use of inaccurate and in some cases misleading search terms. Even when the user is an expert in the area being searched, the ability to select the proper search terms is constrained by lack of knowledge of the author's vocabulary.

There are natural obstacles to specification of the information, a user needs that come from ambiguities inherent in languages, limits to the user's ability to express what information is needed and differences between the user's vocabulary corpus and that of the authors of the items in the database. Languages suffer from word ambiguities such as homographs and use of acronyms that allow the same word

to have multiple meanings. Many users have trouble in generating a good search statement. The typical user does not have significant experience with, nor even the aptitude for Boolean logic statements. The use of Boolean logic is a legacy from the evolution of database management systems. Multi-media also adds an additional level of complexity in search specification.

Thus, an information retrieval system must provide tools to help to overcome the search specification problems. In particular, the search tools must assist the user automatically and through system interaction in developing a search specification that represents the need of the user and the writing style of diverse authors.

There are three basic components in a generalized modern computer-based information retrieval system. They are, the database; the information seeker; and the retrieval techniques, tools, models, and processes which attempt to bridge the gap between searcher and the database.

18.2.1 Database

It is designed to take-in information and information sources acquired for the specific purpose of serving user groups. The selection of information and other sources is based on this objective of serving them. A database has a logical and physical organisation. It is an arrangement based on the users' approach. The main purpose of a bibliographic database is to collect and collate standard bibliographic information, assign index terms and abstracts for technical papers and monographs, etc. In case of full text databases, the complete text of document is made available along with bibliographic details.

Information Retrieval Systems are increasingly using unformatted, free form means of storing information. Record-keeping files in a traditional paper based system might have, for example, a separate line for a person's name, a separate line for an address, a place for a seven-digit phone number, and so on. In computer systems, it is now more common to have information stored in a less specifically located form, making more difficult to locate specific features such as a name or an address. The increased use of word processing, the proliferation of desktop personal computers, and the introduction of optical scanners that can read and convert material into organizational databases, which store information in a way that can make it more difficult to recreate structurally, are making information storage systems more popular.

On the other hand, precisely because information can be stored in a free form, unformatted way, electronic means of storage also make it possible to retrieve and manipulate information in ways that are not possible in a paper-based system, as our increasing reliance on electronic means of storage and retrieval system makes it clear. The difference in power and flexibility in retrieving information can be illustrated by comparing searches done in print indexes and their electronic counterparts. A search in a print index restricts the user to chosen keywords or index terms that can be combined in relatively limited ways. A search in an electronic database generally allows freedom to the searcher and allows him to define more precisely the hierarchical and proximal relationships the terms should have.

18.2.2 Information Base

An Information Retrieval System locates and presents information to the user, based on a query presented to the system. A query may indicate precisely the characteristics of information to be retrieved, or it may express as approximation of information need, indicating merely an initial guess as to the characteristics of the information to be retrieved. A precise-query is most commonly found where an answer is needed to a single factual question, such as "which is the capital of Gujarat State". Question requiring more complex or more ambiguous answers may be expressed by queries that are less precise. Take for instance, an

organization that desires to hire internally to establish an information system that will store and retrieve documents in multiple languages. A secretary in a personnel department would pose a question, such as, "who has the experience with multilingual systems". This question is looking for a series of potential answers. Information Retrieval System must be able to receive different kinds of queries and answers in different ways, retrieving individual facts or groups of potentially relevant items.

Such flexibility is necessary because of the widely varying sources and reasons for needing information, which are rooted in what Belkin, Brooke and Oddy [1972] refer to as Anomalous State of Knowledge (ASK). These ASKs may represent the lack of a particular fact in the information seeker's knowledge base, they may represent a much larger area of missing information, or they may represent a lack of knowledge structure. A small need can often be answered with a single fact, while a much larger or more unstructured area within one's knowledge base may potentially be answered in a variety of ways be a number of facts or documents. Information base, whether databases or collections, must be structured and organised to meet the potential needs of the people who will use them, so that the means of information retrieval must also be selected or created with an understanding of what information needs will have to be met, and how people are likely to understand and use the system.

18.2.3 Structured Query Language (SQL)

Structured Query Language (SQL) is a query language used for accessing and modifying information in a database. Some common SQL commands include 'insert,' 'update,' and 'delete'. Queries take the form of a command language that lets a user select, insert, update, find out the location of data, and so forth. There is also a programming interface. The language was first created by IBM in 1975 and was called SEQUEL for "Structured English Query Language". Since then, it has undergone a number of changes, with a lot of influence from Oracle Corporation. Today, SQL is commonly used for Web database development and management. Though SQL is now considered to be a standard language, there are still a number of variations of it, such as mSQL and mySQL. Many database products such as MS-Access, SQL Server and Oracle support SQL with proprietary extensions to the standard language. Some Information Retrieval Systems are limited to finding those facts or documents containing characteristics specified by the query. Such systems are often referred to as database systems or database as SQL or forms variant of it recognized by the system under consideration. SQL allows precise specification of the value for attributes of terms to be retrieved.

An example of an SQL query might be:

Select from courses where Student's Name = "Anjali Kapoor" and Department = "Management Science"

Self Check Exercise

1) What do you understand by Information Retrieval?

Note: i) Write your answer in the space given below.
ii) Check your answer with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

18.3 INFORMATION RETRIEVAL TECHNIQUES

The organisation of information and the development of various techniques to retrieve information have been major areas of research. With the development in computer technology, interest in this area has been renewed through greater emphasis on the computerised information retrieval systems.

The required techniques can be broadly categorised into exact match, best match and partial match techniques and the latter can be divided into individual or network techniques. These can be further broken down to accommodate specific techniques, such as, cluster, probabilistic, vector-space and so on. The technique most widely used is the **exact match retrieval** one. It is implemented as Boolean, full-text or string searching. Its advantages are that it is easy to understand and use and is available on most systems. But its disadvantages are many. It misses many relevant texts, which match the query only partially; it neither ranks documents nor does it take into account the relative concepts either within the query or within the text. We may have to think about a researcher while looking for leads or at the beginning of the research problem, be satisfied with information retrieved using only this technique. In research of multi-dimensional and multi-disciplinary nature, and sometimes covering fringe areas, probably a **spreading activation technique** or the use of a citation strategy where the seed document is a highly cited one, will be more effective in this instance. The **best match** refers to comparison of two or more and need not be same as exact match. You may refer to Unit 17 for further details.

The issue of matching queries to retrieval has another aspect of study. The issue here is whether **specific retrieval techniques** for specific kinds of queries are on a professional basis of appropriateness and is preferable, or we can also opt for a single retrieval technique to many questions.

However, in the context of intelligent information systems, different types of retrieval techniques would be required for different situations of users. Cognitive mapping and other retrieval techniques provide an integrative approach to overall research process.

Self Check Exercise

2) What are the different Information Retrieval techniques?

Note: i) Write your answer in the space given below.
ii) Check your answer with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

18.4 MODELS BASED ON INPUT/OUTPUT

On the basis of input and the output, Information Retrieval Models can be grouped into three basic categories:

i) Data Retrieval Model

ii) Information Retrieval Model

iii) Knowledge Retrieval Model.

18.4.1 Data Retrieval Model

Data retrieval model essentially handles data. For the purpose of our understanding, data can be taken as unprocessed information or preliminary phase of information. Data is an unbiased fact which can be used to form an information. For example, we can say that the population of the city of Jaipur is eighty (80) lakhs. This is a data. Thus, a census system is a data retrieval system. Similarly, National Sample Survey Organization and Central Statistical Organisation can be taken to be numerical data systems. A data retrieval model calls for organisational structure based on various criteria such as properties, clusters and other different entities. There is a need for a taxonomic presentation of these aspects. Such a taxonomic presentation must also be accessible from other types of associations. A searcher of a data comes for a specific information retrieval. Therefore, the expression of information need should be very precise. Therefore, the data retrieval model is a simple model of information retrieval needing specific matching techniques viz., a taxonomic structure of the various entities involved and their properties.

18.4.2 Information Retrieval Model

Information is data oriented to a purpose. It actually combines several data into a relational structure. Information retrieval is, therefore, a more complex model. It has to generally comprehend multi-dimensional relationships. It is not amenable easily to a taxonomic structure. The representation of information is to be based on a relational data base structure using some associative mathematics. The expression of information need is also complex and time consuming. It draws out for a long conversational or browsing process and the information retrieval model must incorporate such facilities and interfaces.

18.4.3 Knowledge Retrieval Model

Knowledge is a kind of integration of general types of information. It normally occurs in the human mind. The human mind infers and integrates several coordinates with the information received by it. So, knowledge is assimilated information. In order to facilitate decision-making and problem solving, intelligent knowledge based information retrieval models are coming up. Such systems comprise three basic aspects:

a) The so-called **knowledge base** or a store of accumulated set of rules for converting information into knowledge. It also incorporates knowledge acquisition system.

b) The second aspect of the system is **inference engine**. An inference engine is capable of deriving appropriate information from the combination of rules for deriving a synthesized knowledge. This process of deriving is based on inferential logic using quantitative and non-quantitative techniques.

c) A **user interface**, i.e., conversational process in the model which is capable of receiving information in the conversation mode and converting it into database signals for interaction purposes. Thus, a knowledge retrieval model is a sophisticated model of information processing, organization and retrieval.

Self Check Exercise

3) Discuss briefly the various Retrieval Models.

Note: i) Write your answer in the space given below.
ii) Check your answer with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

18.5 MODELS BASED ON THEORIES AND TOOLS

Based on theories and methods/tools available in other disciplines, a number of models have been developed in order to find satisfactory solutions for information retrieval problems.

18.5.1 Boolean Retrieval Model

18.5.1.1 Standard Boolean Retrieval Model

Boolean logic allows a user to logically relate multiple concepts together to define what information is needed. Typically, the Boolean functions apply to processing tokens identified anywhere within an item. The typical Boolean operators are AND, OR, and NOT. These operations are implemented using set intersection, set union and set difference procedures. A few systems introduced the concept of 'Exclusive OR' but it is equivalent to a slightly more complex query using the other operators and is not generally useful to users since most users do not understand it.

The normal Boolean operations produce the following results:

"A AND B" retrieves those items that contain both terms A and B.

"A OR B" retrieves those items that contain the term A or the term B or both.

"A NOT B" retrieves those items that contain term A and not contain term B.

18.5.1.2 Weighted Boolean Retrieval Model

The two major approaches to generating queries are Boolean and natural language. Natural language queries are easily represented within statistical model and are usable by the similarity measures. Issues arise when Boolean queries are associated with weighted index systems. Some of the issues are associated with how the logic (AND, OR, NOT) operators function with weighted values and how weights are associated with the query terms. If the operators are interpreted in their normal interpretation, they act too restrictive or too general. Salton [1979] showed that using the strict definition of the operators would sub-optimise the retrieval expected by the user. Closely related to the strict definition problem is the ranking which is missing in pure Boolean process. Salton provided additional insight into the issues of merging the Boolean queries and weighted query terms under the assumption that there are no weights available in the indexes. The objective is to perform the normal Boolean operations and then refine the results using weighting techniques.

Weighting of index terms is not common in manual indexing systems. Weighting is the process of assigning an importance to an index term's use in an item. The weight should represent the degree to which the concept associated with the index term is represented in the item. The weight should help in discriminating the extent to which the concept is described in items of the database. The manual process of assigning weights adds additional overhead on the indexer and requires a more complex data structure to store the weights. In a weighted indexing system, an attempt is made to place a value on the index term's representation of its associated concept in the document. An index term's weight is based upon a function associated with the frequency of occurrence of the term in the item. Typically, values for the index terms are normalised between zero and one. The higher the weight, the more the term represents a concept discussed in the item. The weight can be adjusted to account for other information such as the number of items in the database that contain the same concept.

The query process uses the weights along with any weights assigned to terms in the query to determine a scalar value (rank value) used in predicting the likelihood that an item satisfies the query. The results are presented to the user in order of the rank value from highest number to lowest number.

If weights are assigned to the terms between the values 0.0 to 1.0, they may be interpreted as the significance that users are placing on each term. The value 1.0 is assumed to be the strict interpretation of a Boolean query. The value 0.0 is interpreted to mean that the user places little value on the term. Under these assumptions, a term assigned a value of 0.0 should have no effect on the retrieved set. Thus,

"A1 OR B0" should return the set of items that contain A as a term.

"A1 AND B0" will also return the set of items that contain term A.

"A1 NOT B0" also return set A.

This suggests that as the weight for term B goes from 0.0 to 1.0 the resultant set changes from the set of all items that contains term A to the set normally generated from the Boolean operation. The process can be visualised by use of the Venn diagram shown in Figure 18.1.

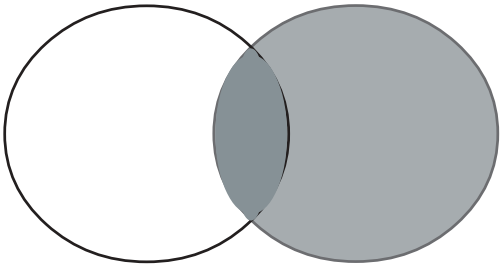

Fig. 18.1: Venn Diagram

Under the strict interpretation "A1 OR B1" would include all items that are in all the areas in the Venn diagram. "A1 OR B0" would be only those items in A (i.e., the white and dark shaded areas) which is everything except items in "B NOT A" (the grey area). Thus, as the value of query term B goes from 0.0 to 1.0, items from "B NOT A" are proportionally added until at 1.0 all of the items will be added.

Similarly, under the strict interpretation "A1 AND B1" would include all of the items that are in the black dotted area. "A1 AND B0" will be all of the items in A as described above. Thus, as the value of query term B goes from 1.0 to 0.0 items will be proportionally added from "A NOT B" (white area) until at 0.0 all of the items will be added.

Finally, the strict interpretation of "A1 NOT B1" is white area while "A1 NOT B0" is all of A. Thus as the value of B goes from 0.0 to 1.0, items are proportionally added from "A AND B" (dark shaded area) until at 1.0 all of the items have been added.

The final issue here is the determination of which items are to be added or dropped in interpreting the weighted values.

18.5.2 Fuzzy Logic Model

An Information Retrieval System has software component that has the features and functions required to manipulate 'information' items versus a DBMS that is optimized to handle 'structured' data. Here information is regarded as fuzzy text. The term 'fuzzy' is used to imply the results from the minimal standards or controls on the creators of the text items. The author presents concepts, ideas and abstractions along with supporting facts. As such, there is minimal consistency in the vocabulary and styles of items. The searcher has to be omniscient to specify all search term possibilities in the query.

Fuzzy logic supports values – true and false as well as other values in between.

The conceptual fuzzy logic was introduced by Professor Lotfi A Zadeah. The basic objective of the fuzzy logic is to develop a model that could be close to natural language process. It is an appropriate tool for modeling the kind of uncertainty associated with vagueness with imprecision.

Fuzzy retrieval provide the capability to locate spellings of words that are similar to the entered search term. This function is primarily used to compensate for errors in spelling of words. Fuzzy retrieval increases recall at the expense of decreasing precision. In the process of expanding a query term, fuzzy retrieval includes other terms that have similar spellings, giving more weight to words in the database that have similar word lengths and position of the characters as the entered term. A fuzzy search on the term 'computer' would automatically include the following words from the information database: 'computer', 'compiter', 'computter', 'compute'. An additional enhancement may lookup the proposed alternative spelling and if it is a valid word with a different meaning, include it in the search with a low ranking or not include it at all (e.g., 'commuter'). Systems allow the specification of the maximum number of few terms that the expansion includes in the query.

Fuzzy retrieval has its maximum utilisation in a system that accepts items that have been optical character recognised. In the OCR process, a hardcopy item is scanned into a binary image. The OCR process is a pattern recognition process that segments the scanned in image into meaningful sub-regions, often considering a segment – the area defining a single character. The OCR process will then determine the character and translate into an internal computer encoding. Based upon the original quality of the hardcopy this process introduces errors in recognising characters. With decent quality input, systems achieve in the 90-99 per cent range of accuracy. Since these are character errors throughout the text, fuzzy retrieval allows location of items of interest compensating for the erroneous characters.

18.5.3 Set Theoretic Model

The set theoretical view of information retrieval is based on the recognition that information requests are normally formulated by choosing collections or sets of item identifiers, or keywords. The keyword sets 'in turn' lead to the retrieval of record subsets chosen from among the stored collection of records. The fundamental data of retrieval theory are provided in this view by the relations which exist between the set of item descriptions and the corresponding record sets.

18.5.4 Vector Space Model

Often weighted systems are discussed as vectorised information systems. This association comes from the SMART system at Cornell University, created by Dr. Gerald Salton [1979]. The system emphasises weights as a foundation for information detection and stores these weights in a vector form. In systems, based upon a vector model, the semantics of every item are represented as a vector. A vector is a one-dimensional set of values, where the order/position of each value in the set is fixed and represents a particular domain. Each vector represents a document and each position in a vector represents a different unique word (processing token) in the database. There are two approaches to the domain of values in the vector – binary and weighted. Under the binary approach, the domain contains the value of one or zero, with one representing the existence of the processing token in the item. In the weighted approach, the domain is typically the set of all real positive numbers. The value of each processing token represents the relative importance of that processing token in

representing the semantics of the item. The value assigned to each position is the weight of that term in the document. A value of zero indicates that the word is not in the document. The system and its associated research results have been evolving for over 30 years. Queries can be translated into the vector form. Search is accomplished by calculating the distance between the query vector and the document vector. The use of weights also provides a basis for determining the rank of an item. The vector approach allows for a mathematical and a physical representation using a vector space model.

In addition to the general problem of dynamically changing databases and the effect on weighting factors, there are problems with the vector model on assignment of a weight for a particular processing token to an item. A major problem comes in the Vector Space Model when there are multiple topics being discussed in a particular item. There is no way to associate correlation factors between terms, since each dimension in a vector is independent of the other dimensions.

The Vector Space Model procedure can be divided into three stages. The first stage is the document indexing where the content bearing terms are extracted from the document text. It is obvious that many of the words in a document do not describe the content, like, the, is, are, in, to, of, etc. These are called non-significant words or stop words. In case of automatic indexing, these terms are removed from the document vector, so the document will only be represented by the content-bearing terms. In general, 40-50% of the total number of words, in a document, are stop words. These can be removed with the help of a stop word list. The second stage is the weighting of the indexed terms to enhance retrieval of document relevant to the user. The last stage ranks the document with respect to the query according to a similarity measure.

The VSM is contrary to the Boolean Retrieval Model in which retrieval is based on the hundred percent (exact) match. The VSM allows retrieval of the most similar to the query without the exact match. Thus, the VSM can be well explained in terms of *keyword-by-document matrix (A),* in which the rows correspond to keywords (W) in the database and the columns correspond to documents (D), then the matrix will be like:

		D1	D2	D3	D4	…..	Dn
	W1	A11	A12	A13	A14	…..	A1n
	W2	A21	A22	A23	A24	…..	A2n
A =	W3	A31	A32	A33	A34	…..	A3n
	W4	A41	A42	A43	A44	…..	A4n
	…..	….	….	….	….	…..	….
	Wm	Am1	Am2	Am3	Am4	…..	Amn

Let us take a hypothetical example, like, an information seeker searches information on *"Education information retrieval system"*. He uses four keywords W1, W2, W3, and W4. After searching the database, he gets six articles: A1, A2, A3, A4, A5, and A6. After analysis, it is found that the first article A1 talks only about W1; article A2 discusses one-third topic of W2 and two-third topic of W4; article A3 deals with 20% of W1, 30% of W3 and 50% of W4; article A4 deals with 60% of W1, 10% of W2 and 30% of W4; Article A5 talks 80% about W2 and 20% about W3; and article A6 discusses only about W4. Now this can be denoted in the form of a 4X6 matrix:

		A1	A2	A3	A4	A5	A6
	W1	1.00	0.00	0.20	0.60	0.00	0.00
	W2	0.00	0.33	0.00	0.10	0.80	0.00
A =	W3	0.00	0.00	0.30	0.00	0.20	0.00
	W4	0.00	0.67	0.50	0.30	0.00	1.00

The VSM is a retrieval model which constitutes a fairly large class of retrieval methods, each consisting of an indexing method and a retrieval function, The indexing method generates description vectors, and the retrieval function generates retrieval status values by comparing the query description vector with the document description vectors. A conceptual diagram of VSM is given at Figure 18.2. The information seeker is assumed to have information need, which he formulates as a query. The query ***q*** and the document ***dj*** are indexed in two steps. First appropriate indexing features are spotted in the query ***q*** and in the document ***dj***. Secondly, these features are assigned weights to obtain the query description and the document descriptions are sets of weighted indexing features. These are called document description vector and query vector. The query description and document descriptions are matched and a score is generated for every document pair. These scores are called Retrieval Status Values (RSVs). For every query, the documents are presented to the information seeker in descending order of these RSVs.

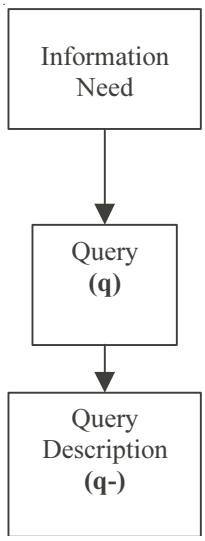

Fig. 18.2: Vector Space Model

The VSM relies on the premise that the meaning of a document can be derived from the document's constituent terms. Each keyword in a document collection forms document vector which represents the single or multiple occurrences of the term ***i*** in document ***d***. Similarly, a query is represented by a query vector which denotes the number of occurrences of terms in the query. Both the document vector and query vector provide the locations of the objects in the term-document space. There are two common one-dimensional measures that every vector has, length and angle with respect to a fixed point. The angle

between two vectors refers to the measure in degrees between those two vectors. The document vector whose angle is closest to the query vector's angle is the best choice, yielding the document most closely related to the query. It is measured in terms of cosine angle between the two vectors. If the cosine of the angle is 1, then the angle between the document vector and the query vector measures 0 degree, meaning the document vector and the query vector move in the same direction. A cosine measure of 0 would mean the document is unrelated to the query vector. Thus, a cosine measure close to 1 means that the document is closely related to the query.

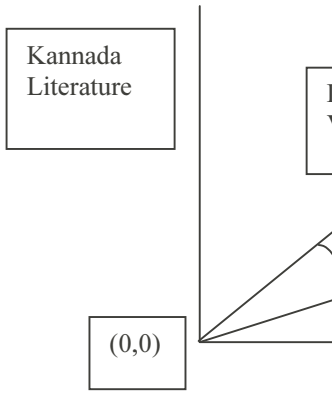

Fig. 18.3: Mapping of vector spaces

If a query ***(q)*** is considered to be a line in an imaginary space and the document ***(d)*** is also considered to be a line in the imaginary space, the geometrically determined angle between the two lines can be understood as measuring the degree to which the documents are similar to the query. While in the case of a large angle the document is presumed to be dissimilar to the query, in the case of a very small angle the document is presumed to be highly similar to the question. In this example, each Vector is plotted from (0,0) point to a point determined by the number of occurrences of the terms 'Kannada Literature' and 'Criticism' in the document or question. If there is one occurrence of 'Criticism' and two occurrences of 'Kannada Literature' in the document, the vector ***(d)*** in the Figure 18.3 is obtained.

18.5.5 Probabilistic Retrieval Model

Probabilistic considerations may apply if one assumes that system characteristics such as the term assignment to the records, or the relevance properties of the records are probabilistic in nature. Other mathematical techniques that have been used include decisions theory, information theory, pattern classification, mathematical linguistics and feature selection methods.

In addition to a vector space model, the other dominant approach uses a probabilistic retrieval model. The model that has been most successful in this area is the Bayesian approach. This approach is natural to information systems and is based upon the theories of evidential reasoning (drawing conclusions from evidence). Bayesian approaches have long been applied to information systems. The Bayesian approach could be applied as part of index term weighting, but usually is applied as part of the retrieval process by calculating the relationship between an item and a specific query.

The probabilistic approach is based upon direct application of the theory of probability to information retrieval systems. This has the advantage of being able to use the developed formal theory of probability to direct the algorithmic development. The use of probability theory is a natural choice because it is the basis of evidential reasoning.

According to Van Rijsbergen [1979], probability theory is an intuitively pleasing model for describing and analysing information retrieval. In this approach one can estimate the probability of retrieval. In other words, the probability of relevance of retrieval is measured. Van Rijsbergen proposed that a measure of the probability of relevance of a given document to a particular theory be based on a Vector representing that document. He postulated that the pattern of index terms in relevant documents will differ from their pattern in non-relevant documents. Probability approach will help to analyse term-clustering, frequency weighting, relevance weighting and ranking.

Probability theory can also be used to rank, and order documents according to their probability of relevance. Robertson [1978] shows that the order of documents can be based on term values and on 'Optimal Retrieval Function'. However, if one attempts to rank order of the documents in a Boolean environment, some difficulties arise which are inherent to the Boolean logic. Bookstein [1978] suggested that the retrieved documents be ordered according to the number of Boolean expressions present in the document that are true.

Doszkoc [1978] suggests that probability associations are being used to find terms that are associated with other terms. The association procedures are based on term occurrences and the frequency of these terms in the database.

The probabilistic approach of a retrieval model is based on the assumption that the distribution of the indexing features tells something about the relevance of a document. This approach adopts a retrieval model which optimizes the retrieval effectiveness according to the Probability Ranking Principle (PRP).

William Cooper has formulated the Probability Ranking Principle as "If a reference retrieval system's response to each request is a ranking of the documents in the collections in order of decreasing probability of usefulness to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data has been made available to the system for this purpose, then the overall effectiveness of the system to its users will be the best that is obtainable on the basis of that data". In this method the system replies to a query by presenting the beginning of a list of documents that are ranked in descending order of scores that either represent probabilities themselves or could be mapped to probabilities by means of an order preserving transformation. These scores often called Retrieval Status Values (RSV) depend on document descriptions consisting of appropriate statistical information about the indexing features. Such score may also depend on domain dependent parameters that are estimated by means of additional data e.g. by a thesaurus.

Probabilities are usually based upon a binary condition – an item is relevant or not. But in information systems the relevance of an item is a continuous function from non-relevant to absolutely useful. The output ordering by rank of items based upon probabilities, even if accurately calculated, may not be as optimal as that defined by some domain specific heuristic. The source of the problems that arise in application of probability theory come from a lack of accurate data and simplifying assumptions that are applied to the mathematical model.

18.5.6 Linguistic Model

In linguistic model for information retrieval, study the information retrieval from the point of view of the properties of language. Information retrieval is provided by features of natural language as well as artificial language. The various ways of storage of information are essentially based on natural language. The human

communication itself is full of natural language. In short, the languages carry three types of functions:

i) They represent the contents of documents and other forms of information;

ii) The information problem of users are represented in terms of language; and

iii) Language is used in computer processing and in searching and retrieving of information.

The language works on three bases:

a) Semantic basis which conveys meaning from one human being to another;

b) The syntactic basis which helps formation of semantics in the use of grammar; and

c) The vocabulary, which supply different meaning to terms for formation of sentences, paragraphs and other structures.

The logical structure of a language and the taxonomy of the languages refers to relationship between vocabulary and concepts. The vocabulary generally refers to the logical structure. In modern times the vocabulary control also include thesaural control and technical glossary control. Use of transformational grammar as well as parsing techniques provide processing speed of the language for information retrieval. Besides this, indexing language with coordinative control provides a basic model for information retrieval. Use of associative mathematics in search logic and in search expression formulation, provide yet another type of language control in information retrieval. This linguistic model forms an essential base for information retrieval. In social science field, language plays an ambiguous role because the terminology of the field is not as rigorous as in the field of natural sciences.

18.5.7 Mathematical Model

Mathematical model generally pre-supposes a careful formal analysis of the problem and specification of the assumptions and explicit formulation of the way in which the model depends on the assumptions.

Mathematical models are essentially based on representative mathematics as well as associative connections. In particular, cluster analysis and clustering techniques are used on experimental basis in automatic abstracting and indexing. Use of sets theory and Boolean logic is a very familiar method of mathematical modeling of information. Concept of similarity measures and choice of variable and the combinational aspects of clustering try to provide semantic structure for information represented. Cluster analysis today involves statistical packages or clustering software.

18.5.8 Psychological Model

The psycholinguistic approaches to information retrieval led to the study of formation of concepts in human mind the way in which the human thinking process arranges the ideas, its presentation at the time of enquiry, and the type of retrieval cues it demands while searching has led to a cognitive research linked with computer communication processes. The studies of Belkin, Brooks and Oddy [1979] on Anamolous State of Knowledge provides interesting insights in relation to information retrieval process. Further, the current day studies in the field of information retrieval and artificial intelligence have thrown sufficient light to bring in harmonious coupling of psychological theory into information retrieval.

18.5.9 Economic Model

The economic model of information retrieval centres round the measures of cost effectiveness and cost efficiency of information retrieval These two criteria are based on performance of information retrieval systems in relation to input cost as well as the number of successful outputs. The concept of provision of multiple access points being used gives a chance for measurement of information transfer. The field of information retrieval, which has developed several models of information measurement based on statistical and mathematical techniques used for studies in bibliometrics and scientometrics provides a scope for correlation of economic benefits. However, due to various intangible elements in information retrieval, which cannot be identified, the economic model does not yet provide a holistic approach to information retrieval.

Theoretically, the modelling of information retrieval can be looked at from output such as data, from the methodological approaches from different disciplines. But, in practice we may not find a single model operating in that manner. For that purpose, a collective modelling of various levels can be seen through. For example, the ERIC (Educational Resources Information Center) model for IR is one of the important areas of IR, which is a combination of all approaches to information retrieval [Henry and Diodate, 1991]. The idea behind these theoretical models is to help analytical studies for bringing in efficiency to different aspects of information retrieval. These multi-disciplinary approaches to information retrieval provide a better base.

18.5.10 Hypertext Linkage Model

Hypertext linkages are creating an additional information retrieval dimension. Traditional items can be viewed as two dimensional constructs. The text of the items is one dimension representing the information in the items. Imbedded references are a logical second dimension that has had minimal use in information search techniques. The major use of the citations has been in trying to determine the concepts within an item and clustering items. Hypertext, with its linkages to additional electronic items, can be viewed as networking between items that extend the contents. The imbedding of the linkage allows the user to go immediately to the linked item for additional information. The issue is how to use this additional dimension to locate relevant information.

Looking at the Internet at the current time there are three classes of mechanisms to help find information: manually generated indexes or directories, automatically generated indexes and web crawlers (intelligent agents). Yahoo (http://www.yahoo.com) is an example of the first case where information sources (home pages) are indexed manually into a hyperlinked hierarchy. The user can navigate through the hierarchy by expanding the hyperlink on a particular topic to see the more detailed sub-topics. At some point the user starts to see the end items. Lycos (http://www.lycos.com) and AltaVista (http://www.altavista.com) automatically go out to other Internet sites and return the text at the sites for automatic indexing. Lycos returns home pages from each site for automatic indexing while AltaVista indexes all of the text at a site.

Web crawlers (WebCrawler, Open Text, Pathfinder) and intelligent agents (Coriolis Groups' NetSeeker) are tools that allow a user to define items of interest and they automatically go to various sites on the Internet searching for the desired information. The Uniform Resource Locator (URL) hypertext links can map to another item or to a specific location within an item.

Self Check Exercise

4) Write down the different Information Retrieval Models based on theories and tools.

Note: i) Write your answer in the space given below.
ii) Check your answer with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

18.6 SUMMARY

An information retrieval system (IRS) is a system that is capable of storage, retrieval, and maintenance of information. Information in this context can be composed of text (including numeric and date data), images, audio, video and other multi-media objects. It consists of a software program that facilitates a user in finding the information the user needs. An IRS includes database, information base and Structured Query Language (SQL). Database consists of a set of records or files. An information base consists of a set of databases. SQL is a standard interactive and programming language for getting information from and updating a database. Many database products such as MS-Access, SQL Server and Oracle support SQL with proprietary extensions to the standard language. Queries take the form of a command language that lets user select, insert, update, find out the location of data, and so forth. Information retrieval systems process users queries as well as manipulate online databases using SQL. Thus, it is very important area of study. Retrieval models, like, Data Retrieval Model, Information Retrieval Model and Knowledge Retrieval Model are based on input and output. Data Retrieval Model handles data and can be taken as unprocessed information or preliminary phase of information. Census system is a data retrieval system. Information Retrieval Model is data oriented to a purpose. It combines several data into a relational structure. Knowledge Retrieval Model assimilates several types of information in order to facilitate decision-making and problem solving. It is a sophisticated model of information processing and organisation. The IR models based on theories and tools try to develop efficient IR systems.

18.7 ANSWERS TO SELF CHECK EXERCISES

1) Information Retrieval is a process of selecting information from a store.

2) Information retrieval techniques are broadly categorised into three types:
   i) Exact match retrieval:
   ii) Best match retrieval
   iii) Partial match retrieval.

3) There are three basic information retrieval models.
   i) **Data Retrieval Model** – It essentially handles data and can be taken as unprocessed information or preliminary phase of information. Census system is a data retrieval system.
   ii) **Information Retrieval Model** – It is data oriented to a purpose. It combines several data into a relational structure.

iii) **Knowledge Retrieval Model** – It assimilates several types of information in order to facilitate decision-making and problem solving. It is a sophisticated model of information processing and organisation.

4) The different information retrieval models based on theories and tools are:

- Boolean Retrieval Model (Standard and Weighted)
- Fuzzy Logic Model
- Set Theoretic Model
- Vector Space Model
- Probabilistic Retrieval Model
- Linguistic Model
- Mathematical Model
- Psychological Model
- Economic Model
- Hypertext Linkage Model

18.8 KEYWORDS

Anomalous State of Knowledge (ASK) : Lack of a particular fact in the information seeker's knowledge base, it may represent a much larger area of missing information, or it may represent a lack of knowledge structure.

Fuzzy Retrieval : The capability to locate spellings of words that are similar to the entered search term. Fuzzy retrieval increases recall and decreases precision. In the process of expanding a query term, fuzzy retrieval includes other terms that have similar spellings. This function is primarily used to compensate for errors in spelling of words.

Retrieval Status Values (RSVs) : In retrieval process, the query description and document descriptions are matched and a score is generated for every document pair. These scores are called Retrieval Status Values (RSVs).

Weighting : Weighting is the process of assigning an importance to an index term's use in an item. The weight represents the degree to which the concept associated with the index term is represented in the item. In a weighted indexing system, an attempt is made to place a value on the index term's representation of its associated concept in the document.

18.9 REFERENCES AND FURTHER READING

Belkin, N.J., Brooks, H.M. and Oddy, R.N. (1979). *Representing and classifying Analogous States of Knowledge*. In: *The analysis of meaning: informatics*, edited by M. Maccafferty and K. Gray. London: Aslib.

Bookstein, Abraham. (1978). On the perils of merging Boolean and weighted retrieval system. *Journal of American Society for Information Science*, 29(3), 156-8.

Brookes, Bertram C. and Grifiths, Joseph. (1978). *Frequency rank distributions*. *Journal of the American Society for Information Science,* 27(1), 13-17.

Carter, M.B. (1986). A methodology for the economic appraisal of management information. *International Journal of Information Management,* 193-203.

Doszkoc, Tamas E. (1978). AID: an Associative Interactive Dictionary of online searching. *Online Review*, 2(2), 163-73.

Gopinath, M.A. (1999). *Information retrieval*. In: MLIS-03 course materials. Unit 13. New Delhi: Indira Gandhi National Open University.

Gopinath, M.A. (1999). *ISAR systems: operations and design*. In: MLIS-03 course materials. New Delhi: Indira Gandhi National Open University.

Gopinath, M.A. (1999). *Objectives of information storage and retrieval systems*. In: MLIS-03 course materials. New Delhi: Indira Gandhi National Open University.

Henry, G. and Diodato, V. (1991). The rates of assignment of thesaurus terms in the ERIC information retrieval system: an analysis of hierarchies and levels. *Journal of Documentation*, 47(3), 276-283.

Karen, Spark Jones (1973). *Linguistics and information science*. New York: Academic.

Kemp, Alister (1988). *Knowledge base retrieval system*. London: Aslib.

Levitan, K.B. (1982). Information resources as 'Goods' in the life cycle of information production. *Journal of American Society of Information Science*, 33, 44-54.

McGill, Michael J. (1978). Knowledge and information spaces: implications for retrieval systems. *Journal for the American Society for Information Science,* 27(4), 205-10.

Mock, T.I. and Vasarhelyi, M.A. (1980). A synthesis of the information economics and lens model. *Journal of Accounting Research,* 477-505.

Murthy, S.G.K. and Biswas, R.N. (2004). A fuzzy logic based search technique for digital libraries. *DESIDOC Bulletin of Information Technology*, 24(6), 3-10.

Rijsbergen, Van C.J. (.1979). *Probabilistic retrieval*. In: *Information retrieval*, edited by C.J. Van Rijsbergen. 2nd Ed. London: Butterworths.

Robertson, Stephen E. (1978). On the nature of fuzz: a diatribe. *Journal of the American Society for Information Science,* 29(6), 304-7.

Salton, Gerard. (1979). Mathematics and information retrieval. *Journal of Documentation*, 35 (1), 1-29.

Salton, G. and McGill, M.J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Smith, L.C. (1976). Artificial Intelligence in Information Retrieval System. *Information Processing and Management*, 12(3) 189-222.

Wikipedia.org. (2005). *Information retrieval*. <http://en.wikipedia.org/wiki/information_retrieval/>.