UNIT 1

Introduction to Information Retrieval Systems

Syllabus

Introduction to Information Retrieval Systems: Definition of Information Retrieval System, Objectives of Information Retrieval Systems, Functional Overview, Relationship to Database Management Systems, Digital Libraries and Data Warehouses.

Information Retrieval System Capabilities: Search Capabilities, Browse Capabilities, Miscellaneous Capabilities.

Learning Objectives

- Introduction to Information Retrieval Systems (IRS)
- Objectives of IRS
- Major Functional Processes in IRS
- Relationship between IRS and DBMS
- Digital Libraries and Data Warehouses
- Various Capabilities of Information Retrieval Systems like Search, Browse and Miscellaneous Capabilities.

Introduction

An Information Retrieval System (IRS) refers to a system which is capable of representing, storing, organizing and accessing information. This information can exist in multiple forms such as text, audio, video, images and other multimedia objects. It is a proper way of representing and organizing information provides an easy access to its users. The main objective of an information retrieval system is to reduce user's overhead while searching for desired information. Overhead is the difference between the time required to obtain the information, the user is interested in and time taken to read the actual extracted information. IRS is a software that consists of all the features and functions needed to manipulate information items whereas, a DBMS handles structured data Information storage and retrieval system is another system needed for storing and extracting information in addition to digital libraries and data warehouse. It is basically developed to overcome some of the problems present in digital libraries.

PART-A SHORT QUESTIONS WITH SOLUTIONS

Q1. Define information retrieval system.

Model Paper-I, Q1(a)

Answer :

An Information Retrieval System (IRS) refers to a system which is capable of representing, storing, organizing and accessing information. This information can exist in multiple forms such as text, audio, video, images and other multimedia. Hence, a proper way of representing and organizing information provides easy access to its users.

Q2. What are the two measures associated with an information retrieval systems?

Model Paper-III, Q1(a)

Answer :

The two measures that are associated with information retrieval systems are,

1. Precision and
2. Recall.

1. Precision

It is defined as,

$$\text{Precision} = \frac{\text{Number of relevant items retrieved}}{\text{Total number of retrieved items}}$$

2. Recall

It is defined as,

$$\text{Recall} = \frac{\text{Number of relevant items retrieved}}{\text{Possible number of relevant items}}$$

Q3. List four major functional processes in information storage and retrieval systems.

Answer :

Information storage and retrieval system consists of the following four major functional processes,

1. Item normalization
2. Selective dissemination of information
3. Document database search
4. Index database search along with Automatic File Build process (AFB).

Q4. Write about selective dissemination of information.

Answer :

Model Paper-II, Q1(

Selective dissemination of information is also known as Mail process. The main purpose of this process is to perform dynamic comparison between the items recently received and the existing item statements which are desired by users. Once the comparison is completed, the process sends back the requested item (document) only to those users whose item statements are similar to the item contents.

The mail process comprises of the following elements,

1. Search process
2. User profiles
3. User mail files.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Q5. What are steps for performing search process.

Answer : Model Paper-III, Q1(b)

The following are the steps for performing the search process,

(i) User enters the query.

(ii) Query initiates the search process against all the items.

(iii) System checks if the user is offline or online. In case if he is online, then selective dissemination of information system is responsible for delivering processed items which are desired by the user.

Q6. Write short notes on types of index files.

Answer :

The two types of index files exist are,

(i) Public index files and

(ii) Private index files.

Each user has one or more private index files which lead to a large number of files. Each reference to a private index file results in a small subset of the entire items in the document database.

Public index files are managed by the library services personnel by indexing every item in the document database. A document database consists only of a small number of public index files. These files include access lists that permits its users to search and retrieve the desired data in contrast to the private index files which have very limited number of access lists.

Both private and public index files are created by implementing structured DBMS.

Q7. Write brief note on digital libraries.

Answer : (Model Paper-II, Q1(b) | Dec.19(R16), Q1(a))

A Digital library acts as a repository that stores massive volume of data. Apart from storing information, libraries are also responsible for extracting information based on the media, it has been developed on. Libraries are also termed as "electronic libraries' as they utilize electronic devices to perform their functions when the amount of information is increased. The concept of electronic libraries is growing rapidly because of emerging internet technology.

Later US government renamed electronic libraries as digital libraries. As Internet technology is growing exponentially, the digital libraries conception is also emerging.

Q8. What is the Impact on precision and recall in the use of stop lists?

Answer : Dec.19(R16), Q1(b)

Stop lists and be defined as a set of words that are non-relevant or meaningless and are ignored during the processing. These are also referred as stop words. The impact on precision and recall in the use of stop lists is as follows,

(i) Precision begins at 100% and this value is maintained throughout the retrieval process.

(ii) Recall starts from a value nearer to '0' and increases till all the relevant items have been retrieved from the database.

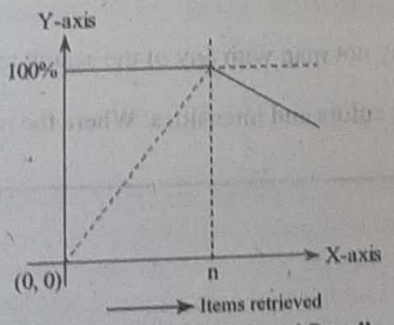


Figure: Ideal Precision and Recall

After retrieving all the relevant items the remaining are non-relevant. The retrieval of non-relevant items directly affects 'precision' and drops it to a value close to zero whereas, 'recall' remains unaffected by the retrieval of non-relevant items and remains at 100% once it is achieved.

Q9. Define proximity function.

Answer :

Proximity function is used to define the maximum units that can be allowed between the two search terms. The units can be in terms of characters, words, sentences or paragraphs.

The syntax for proximity function is,

Term1 with 'i' 'units' of Term2

Where 'i' is an integer number, Term1 and Term2 are the search terms.

Q10. Write about masking. **Model Paper-I, Q1(b)**

Answer :

Masking is the process of hiding some portion of the query term such that all the search terms that matches with the unhidden portion of the query term are treated as valid and accepted.

Query term masking is of two types. They are,

(i) Fixed length and

(ii) Variable length.

Fixed length masking provides a mask only for a single position in a word. The mask provided may be either to a symbol in a particular word or lack of a particular position in a word.

Variable length masking enables the user to mask any number of characters within a particular word. The mask can be applied either at the front or at the back or at both the front and the back (Imbedded mask).

Q11. Define highlighting.

Answer :

The highlighting is a mechanism by which the reason for selection of a particular item is indicated. Highlighting enables the user to concentrate only on the highlighted parts of the text for tracing the relevancy of an item.

There are different phases of highlighting that is, the strongly highlighted text indicates how strongly the text have taken part in the selection of an item. Similarly, the weakly highlighted text indicates how weakly the text have served the purpose of item selection.

Highlighted text within the item is displayed first and then the next highlighted text and so on. The RetrievalWare search system has DCARS system as a front-end which enables the user to search for the item by providing individual words (or) paragraphs as an input.

In some cases, the highlighted terms may not map with any of the search terms entered.

Highlighting can be done using different colors and intensities. Where the more darker shade (intensity) refers to the most relevant word in an item.

Q12. What is canned query?

Answer :

The canned query can also be termed as a stored query. It allows the user to store the query with a particular name so that it can be retrieved and executed in future sessions. Consider a situation where the user needs to repeatedly search a particular area at regular intervals, then instead of writing the same query again and again, the user can simply create a canned query and use it whenever needed. Additional search statements can also be added to the canned query.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

PART-B ESSAY QUESTIONS WITH SOLUTIONS

1.1 DEFINITION OF INFORMATION RETRIEVAL SYSTEM, OBJECTIVES OF INFORMATION RETRIEVAL SYSTEMS

Q13. Define an information retrieval system. Give a brief account on the origin of the Information Retrieval systems.

Answer :

Model Paper-II, Q3(a)

An Information Retrieval System (IRS) refers to a system which is capable of representing, storing, organizing and accessing information. This information can exist in multiple forms such as text, audio, video, images and other multimedia. Hence, a proper way of representing and organizing information provides easy access to its users.

IRS consists of six major subsystems which are as follows,

1. Document subsystem
2. Indexing subsystem
3. Vocabulary subsystem
4. Searching subsystem
5. User subsystem
6. Matching subsystem.

An information retrieval system is a software program that helps its users in obtaining desired information. The standard computer hardware (specialized hardware) is used by the system to support the search process and to transform non-textual information sources into a searchable media. The success of an information system can be determined by identifying how efficiently it minimizes the overhead of a user to retrieve the required information. From the user's perspective, overhead can be calculated as follows,

Overhead = Time required to obtain the desired information – Actual reading time required to read the data

Information retrieval systems were designed to cater to the need of organization information of in central repositories. Several catalogues were provided by these systems that aided identification and retrieval of data items.

With the commercial availability of computers, they have become a useful source of information storage and retrieval. The earlier Database Management System (DBMS) acts as an appropriate platform for electronically manipulating indexes for information. Catalogs and references are used by the libraries which migrate their hard copy information references to structured databases. Many independent developments occur in textual information retrieval systems because of the need to store and search large textual databases in military and government agencies. Also, the inexpensive advancements in powerful personal computer processing systems, such as high speed and large storage capacity, make use of large textual databases possible for the average users.

Q14. Discuss the objectives of information retrieval systems.

May/June-13(R09), Q1(a)

OR

Describe how the statement "language is the largest inhibitor to good communications" applies to information retrieval systems.

Answer :

(Model Paper-III, Q2(a) | Dec.-19(R16), Q2)

Objectives of Information Retrieval Systems

The main objective of an information retrieval system is to reduce user's overhead while searching for desired information. Overhead is the difference between the time required to obtain the information in which the user is interested and time taken to read the actual extracted information. The success of information retrieval system is dependent on the amount of information required and also on the user's wish to accept the search overhead. The required information specifies the information associated with the user's requirement or the amount of information required to accomplish a particular task. If a comparison is made between an ordinary retrieval system and detailed retrieval system, then the former requires less characteristic features than the latter. The disadvantage of detailed retrieval system is that the user is provided with abundant information due to which it becomes very difficult for a user to extract relevant and useful information. The word "relevant" has different meanings in different contexts. From a user's point of view, relevant and needed are treated as synonyms whereas system treats them in a different manner.

Other Objectives of Information System

The first and foremost objective of information system is to provide the search generation support to users.

The following are the hurdles that an IRS suffers from,

1. User requirement faces uncertainty that is present in natural languages.
2. Confines the capability of user, due to which he fails to express the amount of information required.

Thus, the above explanation justifies the statement "language is the largest inhibitor to good communications".

Ambiguity and usage of acronyms are the major problems inherent in all the natural languages. To remove this uncertainty, many unambiguous methods were developed. But the drawback of these methods are,

(i) Increased system overhead

(ii) Increased search time

(iii) Involves user's interaction.

It is very difficult for a user to generate an accurate search statement because of lack of experience in using Boolean logic statements. This problem is resolved by creating many information retrieval systems like Retrieval-Ware, Topic, INQUIRY, etc. These information systems have made it very easy for a user to understand the queries in natural language, which are becoming standard system characteristics.

In addition to these issues, multimedia data also creates problem in search specification. Text searching techniques can be employed while searching textual data but the same is not applicable while searching multimedia data due to lack of interfacing techniques. The multimedia search is possible by storing samples of known entities in advance and allowing the users to select those entities in order to initiate the search. However, this method becomes sophisticated when integrated with either boolean logic or specification of natural language. The complexities of query generation arises for an inexperienced user who begins the search process by taking into account how much information is needed but not what information is needed. Because of this, many search results are inaccurate and misleading. The same problem is faced even by the experienced users, since their search capability is limited because of lack of vocabulary. To overcome such problems many search tools are developed that supports the user in developing the search specification statements which specifies both the user requirements and also different writing styles of various authors. This specification also provides information about the specifications of multimedia data.

Another objective of information retrieval system is to represent the search result in a standard format that assists the user in identifying and extracting only relevant information. This is achieved by executing functions that presents the search results in accordance with the relevant ranking order.

If these functions are combined along with the search status and information formatting option, then the user can easily select the most relevant information as soon as the result is displayed. In order to provide necessary additional information about the item selection many complex methods employ item clustering and item summarization techniques. In contrast to item summarization technique that focuses on common area of the query, question/answer approach is much more difficult since, it requires the result to be examined by considering particular need of information.

Q15. What are the two measures associated with an information retrieval systems? Give the conditions under which 100% precision and recall are achieved. Also derive the relationship between these measures and user overhead.

OR

There are many measures/metrics associated with information systems. For the following scenario, suggest an appropriate measure or metric and justify your choice. A student searching a set of documents to find all of those that pertain to an assignment. All that the student wants is part of an answer or full answer for the assignment, so the student does not mind sifting through all of the documents that were returned.

Dec.-14(R09), Q1(a)

OR

Both precision and recall need to be taken into account when evaluating retrieval systems. Why is not sufficient to pick one and use only that?

Dec.-14(R09), Q1(c)

Answer :

The two measures that are associated with information retrieval systems are,

1. Precision and
2. Recall.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

1. **Precision**

It is defined as,

$$\text{Precision} = \frac{\text{Number of relevant items retrieved}}{\text{Total number of retrieved items}}$$

The number of retrieved items that are relevant (i.e., those items which helps the searcher in finding answers to his questions) is based on the user's search requirement whereas, the total number of retrieved items denote all those items that are retrieved as a result of query.

2. **Recall**

It is defined as,

$$\text{Recall} = \frac{\text{Number of relevant items retrieved}}{\text{Possible number of relevant items}}$$

The possible number of relevant items refers to all those relevant items that are available in the database.

Precision provides only one aspect of the overhead incurred in information retrieval to a user and measure it for a particular search operation. If the precision of search operation is 80% then, 20% of the user effort is the overhead involved in reviewing the non-relevant items (i.e., the items which do not provide any useful information to the user).

The use of one among precision and recall on information retrieval may generate many result, but none of them may be useful. For instance, if a system after the search returns all the documents that contains all the terms matching the query. Such search through gives 100% recall, but may be unhelpful. Similarly if another system returns only a single relevant document that makes the search there by yielding 100% precision search result may be some what helpful but it may not be considered better.

Figure (i) shows the variation in precision and recall values when the number of retrieved items increases. The following observations are made from figure (i),

(i) Precision begins at 100% and this value is maintained throughout the retrieval process.

(ii) Recall starts from a value nearer to '0' and increases till all the relevant items have been retrieved from the database.

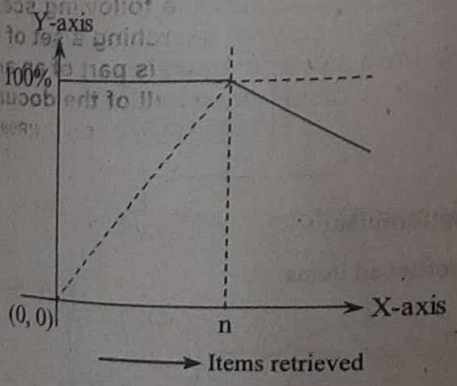


Figure (i): Ideal Precision and Recall

After retrieving all the relevant items the remaining are non-relevant. The retrieval of non-relevant items directly affects 'precision' and drops it to a value close to zero whereas, 'recall' remains unaffected by the retrieval of non-relevant items and remains at 100% once it is achieved.

The graphs drawn for determining precision and recall (i.e., Precision/Recall graphs) depicts how the precision and recall values vary within a search results file (also called a Hit file) on viewing from the most relevant to the least relevant item. The ideal case shown in figure (ii) i.e., when all the retrieved items are relevant, it indicates 100% precision (1.0). Recall continuously increases towards the right on X-axis till it also becomes 100% (1.0). On continuing this process, the precision decreases down the Y-axis. This is shown in figure (iii).

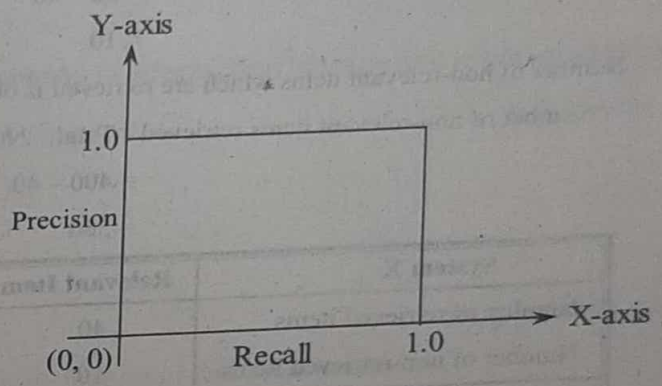


Figure (ii): Ideal Precision/Recall Graph

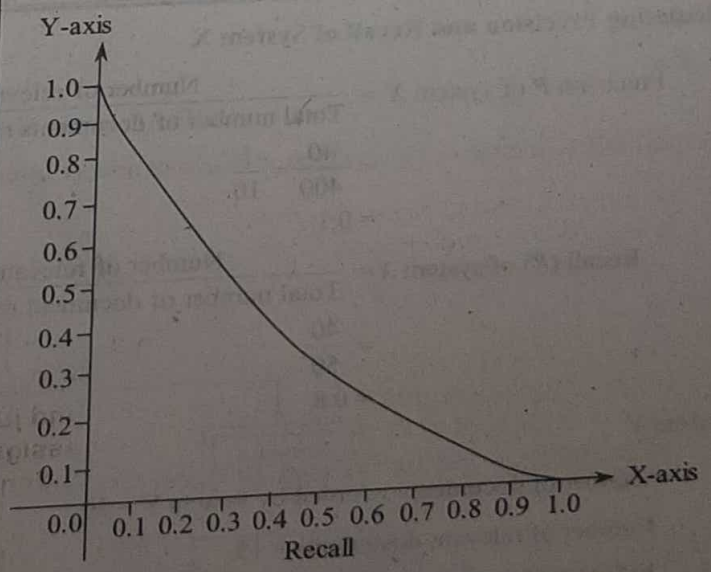


Figure (iii): Achievable Precision/Recall Graph

For, the given scenario, recall is an appropriate measures because students are interested in finding all the relevant documents pertaining to their assignment but not the exact answer. Hence, the suggested measures is "Recall".

Q16. Two retrieval systems, X and Y are being compared. Both are given the same query, applied to a collection of 1500 documents. System X returns 400 documents of which 40 are relevant to the query. System Y returns 30 documents, of which 15 are relevant to the query. Within the whole collection there are in fact 50 documents relevant to the query. Tabulate the results for each system and compute the precision and recall for both X and Y. Show your working.

Answer : Dec.-14(R09), Q1(b)

Given that,

Collection of 1500 documents.

System X

Number of documents returned by system $X = 400$.

Number of relevant documents = 40.

Number of non-retrieved items is obtained by using the following formula,

$$\text{Number of non-retrieved items} = \text{Total} - \text{Number of retrieved items}$$
$$= 50 - 40$$
$$= 10$$

Number of non-relevant items which are retrieved is obtained by using the following formula.

$$\text{Number of non-relevant items retrieved} = \text{Total} - \text{Number of relevant retrieved items}$$
$$= 400 - 40$$
$$= 360$$

System X	Relevant Items	Non-relevant Items	Total
Number of retrieved items	40	400 – 40 = 360	360 + 40 = 400
Number of non-retrieved items	10	1100 – 10 = 1090	1090 + 10 = 1100
Total	**50**	**1450**	**1500**

Calculating Precision and Recall of System X

$$\text{Precision } P \text{ of system } X = \frac{\text{Number of relevant items}}{\text{Total number of documents retrieved by system } X}$$
$$= \frac{40}{400} = \frac{1}{10}$$
$$= 0.1$$

$$\text{Recall } (R) \text{ of system } X = \frac{\text{Number of relevant items}}{\text{Total number of document retrieved by query}}$$
$$= \frac{40}{50}$$
$$= 0.8$$

System Y

Number of documents returned by system Y = 30

Number of relevant documents = 15

Number of non-retrieved items can be obtained by using the following formula.

$$\text{Number of non-retrieved relevant} = \text{Total} - \text{Number of relevant retrieved items}$$
$$= 30 - 15$$
$$= 15$$

Number of non-retrieved non-relevant items can be obtained in following manner.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Number of non-retrieved non-relevant item = Total – Number of non-retrieved relevant item

$$= 1470 - 35 = 1435$$

The table below shows the result of system Y.

System Y	Relevant Items	Non-relevant Items	Total
Number of retrieved items	15	15 (30 – 15 = 15)	30
Number of non-retrieved items	35	1435	1435 + 35 = 1470
Total	**50**	**1450**	**1500**

Calculating Precision (P) and Recall (R) of System (Y)

$$\text{Precision } (P) \text{ of system } Y = \frac{\text{Number of relevant items}}{\text{Total number of documents retrieved by system } Y} = \frac{15}{30} = 0.5$$

$$\text{Recall } (R) \text{ of system } Y = \frac{\text{Number of relevant items}}{\text{Total number of documents retrieved by query}} = \frac{15}{50} = 0.3$$

1.2 FUNCTIONAL OVERVIEW

Q17. Explain in detail about the four major functional processes in information storage and retrieval systems.

Model Paper-I, Q2

OR

Explain item normalization with example.

(Refer Only Topic: Item Normalization)

Oct./Nov.-20(R16), Q1(a)

Answer :

Information storage and retrieval system consists of the following four major functional processes,

1. Item normalization
2. Selective dissemination of information
3. Document database search
4. Index database search along with Automatic File Build process (AFB).

The following figure shows the complete information retrieval system in which boxes depicts functions and disks depicts the data storage,

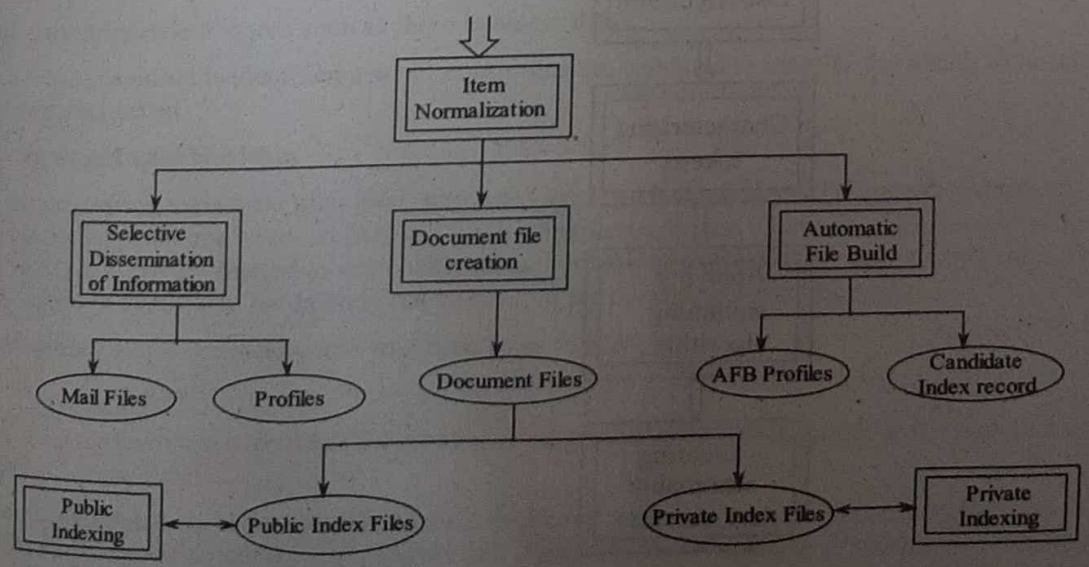


Figure: Complete Information Retrieval System

1. Item Normalization

Item normalization is a process of decomposing (i.e., normalizing) the incoming item into a standard form. This type of normalization is basically carried out in integrated systems. This process performs decomposition by logically restructuring the items once received, in addition to translating numerous external formats that can also be manipulated by functional processes like selective, dissemination of information, document database search, etc.

Extra functions which are to be performed in order to develop a searchable data structure are,

(a) Token identification

(b) Token characterization

(c) Token stemming.

Token identification and its characterization describe the searchable text once the entire text is received. The following are the steps to be considered while performing text normalization,

(i) Standardizing the input

(ii) Logical subsetting (zoning)

(iii) Identifying the processing tokens

(iv) Applying stop lists/algorithm

(v) Characterizing tokens

(vi) Applying stemming algorithm

(vii) Creating searchable data structure.

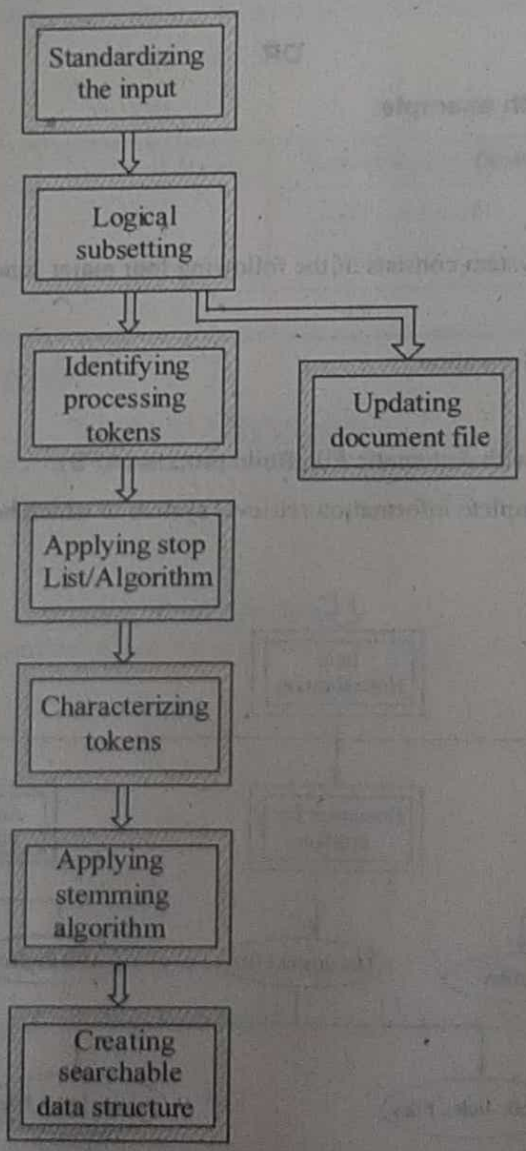


Figure: Steps for Text Normalization

WARNING. Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

(i) Standardizing the Input

In this step, different external formats of incoming data items are converted into a standard format. A system may support single or multiple format for input data items. Standardization could include translation of any international language into unicode format. Translation is done since, every language has its unique binary encoding. One of the standard encoding formats for English, French, etc., is ISO-Latin. Apart from this, there are many other internal encoding formats for different language group.

The advantage of translating the languages into standard format is the ease of displaying all the languages on to a single browser and the use of a single search system to perform the search process. The prerequisite for search engine (to search an item) is to comprehend the linguistic model of every language so as to perform accurate tokenization of individual language. In addition to text normalization, it is also essential to normalize multimedia data. Normalizing such data requires additional functions. There are different standards used for normalizing multimedia data.

- For video, the digital standard considered can include MPEG-1, MPEG-2.
- For audio, audio standards include Real Media, WAV.
- For images, image standards include JPEG, BMP.
- Encoding standards are more beneficial to multimedia rather than textual data that are being managed by interfaces.

(ii) Logical Subsetting (Zoning)

Once the input items are standardized, logical subsetting is performed on these items. In this subsetting, every standardized item is parsed into logical subparts which are relevant to the user. This parsing is user-transparent and basically done to enhance the search precision and utilize the display efficiently (by displaying multiple items per screen). Every individual item is fragmented into zones that may intersect with each other or can be hierarchical in nature. The term "zone" is considered rather than "field" because the former is of variable length and the latter has a connotation of independence. Once all the zoning information is gathered, it is transferred to the token identification process. This process archives the received information and performs the search only in some particular zone of interest. Zoning is very much different from multimedia when source structure is considered.

(iii) Identifying the Processing Tokens

In this step, the information required for performing the search process is to be identified. Identification of processing token is initially started by first identifying a word. This is done by partitioning every input symbol into the following classes,

- **Valid-word Symbol:** It includes alphabetic characters and numbers.
- **Inter-word Symbol:** It includes blanks, periods and semicolons.
- **Special Processing Symbol:** It includes symbols like hyphen that may require additional processing.

Word is defined as shared group of valid-word symbols that are constrained by inter-word symbols, which are generally non-searchable items. The definition of inter-word symbol rely on the concept of language area of the processing items. Selection of inter-word symbols is compromised depending on the search accuracy and also on the characteristic of language. Apart from these symbols they may be other special processing symbols like hyphens. There are many places where hyphen are used,

(i) At the end of line to mark the continuity of word.

(ii) To link two independent words such as large-business firms.

Once a hyphen symbol is identified a set of constraints are executed to identify the action to be taken in order to produce at least one processing token.

(iv) Applying Stop List/Algorithm

Once the processing tokens are identified, stop list/Algorithm is applied to all the identified tokens. The main goal of stop function is to consume system resources. It does this by deleting the irrelevant tokens from the set of processing tokens. Stop Lists comprises of words that are not treated as searchable token because of its frequency and semantic use. The algorithm enhances the processing speed of the system by deleting the words that are occurring more frequently.

Ziph developed a rank-frequency law which states as follows,

$$\boxed{F * R = C}$$

Where 'F' is frequency that specifies the number of times a word is occurring, 'R' is the rank that indicates the rank order of the word and 'C' is the constant.

The words that don't occur frequently are not considered during the search process as they are not treated as a part of vocabulary. The advantage of deleting such words is that it reduces the complexity of storing and accessing the structure. Stop algorithm is generally preferred for eliminating such words since it is very tedious to enlist those words individually.

Let us consider some examples of stop algorithms,

(a) Stop all characters that end with "*A*".

(b) Stop all processing tokens that consist of " – "

Stop algorithm do not have any search value and its objective is to delete distinct item numbers that occur more frequently.

(v) Characterizing Tokens

In this step, the characteristics of a particular word are identified. There characteristics help the system in defining the certainty of a specific word.

The characteristic of a word can be determined in the following ways,

(a) By performing morphological evaluation based on the parts of speech. For example for "budget", the system understands it could mean "the money that is available to a person or an organization" as a noun, "to plan to spend an amount of money for a particular purpose" as a verb.

(b) By classifying a token as a part of higher class token like "African Country" or "Financial Institution".

(c) By preserving the upper case characters. Upper cases are generally preserved for proper nouns, names of organizations or for acronyms. The usage of upper case in processing tokens is different from the way it is used in textual data.

Apart from textual data, characterization of numbers and dates are done in different manner.

(vi) Applying Stemming Algorithms

After completion of identification and characterization, many systems apply stemming algorithms in order to normalize the processing tokens to a standard format. Whenever stemming process is performed, there is a conflict between the search precision value and standardization. This is because, standardization decreases the system overhead in expanding the search term to enact a matching token representation with greater recall value. The more stemming is applied, the more irrelevant items are retrieved. Some of the stemming algorithms are RetrievalWare, Porter Stemming Algorithm, etc.

(vii) Creating Searchable Data Structures

In the final step of text normalization, the finalized processing tokens are used as updates for creating searchable data structures which are nothing but the internal representation of items. This representation is invisible to the user and generally represents items that are being searched by a user entered query. Data structure defines the concept of semantics and confines the search result when multimedia data needs to be searched, it is necessary to specify the relative time that defines the frequency of processing token. This frequency is required to determine the relationship between the text and multimedia source.

2. Selective Dissemination of Information

Selective dissemination of information is also known as Mail process. The main purpose of this process is to perform dynamic comparison between the items recently received and the existing item statements which are desired by users. Once the comparison is completed, the process sends back the requested item (document) only to those users whose item statements are similar to the item contents.

The mail process comprises of the following elements,

(a) Search process

(b) User profiles

(c) User mail files.

Processing of every individual item which is received is done against all the user's profile available in the information system. A user profile basically comprises of a search statements with a broader length and also a user list. The list specifies the mail files of the user along with the information about who should receive the item, only if the search statement specified matches the search profile. In contrast to ad hoc queries, search profiles generally includes more number of searching terms (i.e., more number of user profiles are processed against a single document) and considers larger areas of user interest. Another difference between them is that profiles focuses on every area of user's interest whereas, adhoc queries focus on providing solutions to a particular query.

When a system receives a search item statement, it performs the processing against every user's profile. If the item statement matches with any of the profiles, then the item related to that profile is stored on Mail files. Items in such Mail files are viewed on a regular basis. Once the time expires, the item is automatically deleted from the Mail file. It can also be deleted if the user enters a delete command during display. Mail files make it possible to dynamically update the content of the item in asynchronous manner. Due to this, it become sophisticated to extract the result of dissemination of information in relevant order as desired by the user.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

As detailed research has not performed on Mail dissemination process, only retrospective search algorithms are modified so that they can be applied to Mail profile. Dissemination process doesn't consider a static database while estimating ranking order for an item. This process doesn't rely on the other information system i.e., it is an independent process. But, the process relies on the knowledge factor to decide if desired received document is either user's profile or the received document. Because of this restriction, the dissemination process generates suboptimal system, (i.e., only partial information retrieval system is viewed) due to which the users are forced to receive massive amount of irrelevant information. On the other hand, if the entire information retrieval system is taken into account, then the existing mail as well as the index files are made available during the dissemination process. Due to which the profile of dissemination process is expanded so as to accommodate logic against the available files in the information system.

3. **Document Database Search**

For answer refer Unit-I, Q18, Topic: Document Database Search.

4. **Index Database Search**

For answer refer Unit-I, Q18, Topic: Index Database Search.

Q18. Distinguish between document database search, Index database search and Multimedia database search.

Model Paper-III, Q2(b)

Answer :

Document Database Search

Database search process comprises of the following components,

1. Search process
2. Query
3. Document database.

1. **Search Process**

It is responsible for performing the search against every item which is received, processed and stored by the system.

2. **Query**

The search process is initiated once the user enters the query.

3. **Document Database**

It acts as a repository that contains only those items that have been received, processed or stored by the system. This database is extremely large, comprising multiple number of items. Once the items are received, they are considered as static i.e., they do not alter their respective values.

The search process is based on looking backward (i.e., retrospective) strategy that deals with the past events. If search is made for information that had been processed before then, it is treated as retrospective search.

The following are the steps for performing the search process,

(i) User enters the query.

(ii) Query initiates the search process against all the items.

(iii) System checks if the user is offline or online. In case if he is online, then selective dissemination of information system is responsible for delivering processed items which are desired by the user.

Every document present in the database is considered while processing an individual query. When compared to profiles, queries are of shorter length and concentrate on particular domain of interest. Since, the items are static in nature, the value of information periodically reduces. Because of this, it is possible to fragment the entire database in accordance to the time and store the information based on time fragments. Document database also consist of document available in mail files as these documents act as logical input to both processes. Some user interfaces allow its users to apply database partitions when the items to be searched becomes older and are received after the specified time interval.

Index Database Search

Index database search process performs the search based on the index created by user for desired items which are stored in a file. Archiving these indexed items reduces the search time as well as avoids the need to go back to the actual item. The search process is performed in the following way.

Whenever a user finds the requested item, it stores the item in the file along with the index number which refer that item as well as some descriptive text. This descriptive text specifies the description about the item. If the user desires to retrieve the already stored information, then the user can apply the index number for extracting that information instead of initiating the search process from scratch.

In addition to the index records that provide references to the stored items, it is also possible to create index records that contains only important information but not the references of the items. There are many applications wherein index files systems are applicable. One such application is card catalog available in the library system. Index files can also be viewed as structural databases. The record in such databases can also refer items stored in document database. The advantage of index file search process is that it enables the system with the ability to perform combined file search. In this file search, the index is searched initially and then search is made for the items that satisfies the index part of the query and whose reference are stored in index record.

Types of Index Files

The two types/classes of index files exist are,

(i) Public index files and

(ii) Private index files.

Each user has one or more private index files which lead to a large number of files. Each reference to a private index file results in a small subset of the entire items in the document database.

On the other hand, public index files are managed by the library services personnel by indexing every item in the document database. A document database consists only of a small number of public index files. These files include access lists that permits its users to search and retrieve the desired data in contrast to the private index files which have very limited number of access lists.

Both private and public index files are created by implementing structured DBMS.

Multimedia Database Search

Multimedia data doesn't represent independent data structure but instead represents an augmentation to the actual available data structure present in information retrieval system. This type of data is basically present in every area specified by the document database. Search for multimedia data is performed with the help of augmented search structure. The actual sourced data is stored in normalized digitized form in their respective retrieval servers. The relationship between the areas of multimedia and textual data can be obtained by performing synchronization that basically enhances the accuracy of search process.

The two types of synchronization are,

1. **Time Synchronization**

It consist of text translated from audio or video sources.

2. **Positional Synchronization**

It consist of multimedia data that have been locally stored by employing hyperlink in a textual item.

Whenever both multimedia and textual search results in hits that are closely related, it is necessary to assign augmented significant weights to the data.

1.3 RELATIONSHIP TO DATABASE MANAGEMENT SYSTEMS

Q19. Bring out the relationship between information retrieval system and database management systems.

Answer : Model Paper-II, Q2(a)

The processing of items can be done by using two types of systems,

(i) Information Retrieval Systems (IRS) and

(ii) Database Management Systems (DBMS).

IRS is a software that consists of all the features and functions needed to manipulate information items whereas, a DBMS handles structured data.

A structured data refers to the collection of facts (appropriately defined data) represented by means of tables i.e., all the semantic descriptions among the attributes are expressed in database tables. This is done to avoid the ambiguity that results from using the data. For example, in an employee database the fields such as 'e_name' and 'e_sal' and their values doesn't cause any confusion to its users. On the other hand, if two persons are asked to create an abstract for some data item 'X' then these two abstractions can be different. Hence, fuzzy nature must be associated with information items in order to resolve the resulting ambiguity that arises in a language.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Using structured data, a user can send a request and receive the desired information. It is tabulated and displayed in the form of a report for easy-to-use purpose. On the contrary, a search process conducted to find all the desired information by a user, is very difficult. This can be made easy by refining the search process. This technique is called 'iterative search process'. An IRS provides this iterative search process to its users and guides them in finding the required items (known as Relevance feedback). The results are provided to the users in relevance ranked order. Though DBMS is easy to implement it doesn't provide features like ranking and relevance feedback. Use of structured data with information systems is also possible.

Hence, both DBMS and IRS must be integrated. Commercial database companies must implement such systems that can perform this integration. One of such commercial databases is INQUIRE DBMS which lasts for over fifteen years. Another example system is ORACLE DBMS. This can perform integration through an embedded capability called CONVECTIS. The RetrievalWare provided by the INFORMIX DBMS integrates the structured data and information associated with IRS.

Q20. Distinguish between database and IRS.

Answer : (Model Paper-III, Q3(a) | Oct./Nov.-20(R16), Q1(b))

	Database		IRS
1.	A database system is an organizational structure that consists of several components, which are used to specify and control the data stored in a particular database environment.	1.	An Information Retrieval System (IRS) refers to a system which is capable of representing, storing, organizing and accessing information.
2.	The Database Management System (DBMS) provides precise and appropriate semantics.	2.	The IRS mostly provides imprecise semantics.
3.	The data can be retrieved in the following three levels, (i) User level (ii) Physical level and (iii) Conceptual level.	3.	An IRS can be constituted in the following three areas, (i) User queries (ii) Items of information and (iii) Matching queries.
4.	The database system provides Data Modelling Facility (DMF) for manipulating and modeling the data.	4.	The IRS do not provide DMF as data modeling is restricted to classification of objects.
5.	The amount of capabilities and mechanisms for defining the data integrity constraints are developed in more quantity.	5.	In IRS, such mechanisms are developed in less quantity.
6.	In DBMS, query language is artificial.	6.	In IRS, query language is almost similar to natural language.
7.	In DBMS, query specification is complete.	7.	In IRS, query specification is incomplete.
8.	In DBMS, there exists structured data format.	8.	In IRS, there exists unstructured data format.

Q21. If a medical query system's collection of records is actually a database of journal and conference articles in medicine, is it more a database retrieval system than an information retrieval system. Justify.

Answer : Nov./Dec.-12(R09), Q1(a)

A medical query system that retrieves medical journals and articles is an Information Retrieval (IR) system.

IR system refers to a software system that is capable of representing, storing, organizing and accessing information. The information can exist in various formats such as text, audio, video and images.

Further, in IR system the interest is to find items that matches the query partially. The system uses natural language to query and it is less sensitive to errors i.e., small errors in retrieved document does not effect the system performance.

Users can query the medical system to retrieve articles or documents regarding various topics such as,

1. Childhood obesity and eating habits.
2. Support of family in treating schizophrenia.
3. Genetic disorders and its treatment.

When users search for "childhood obesity and eating habits" the system retrieves articles related to the search criteria from different journals.

A medical query system cannot be a database retrieval system as database retrieval system uses artificial query language restricted with syntax. A database retrieval system retrieves exact matching items and it is more sensitive to errors.

1.4 DIGITAL LIBRARIES AND DATA WAREHOUSES

Q22. What is the difference between the concept of a 'Digital Library' and an 'Information Retrieval System'? What new areas of information retrieval research may be important to support 'Digital Library'?

Answer : Model Paper-I, Q3(a)

Information storage and retrieval system is another system needed for storing and extracting information in addition to digital libraries and data warehouse. It is basically developed to overcome some of the problems present in digital libraries. This information system concentrates on searching and extracting textual data rather than developing standards on system content. It doesn't take into consideration the problems associated with unique identification and information tracking, which is needed by the copyright authoritative aspect. This copyright limits the boundary of the function to be within the library environments. On the other hand, there are intellectual property rights that are not handled by any country, due to which these rights create internet issues. Digital libraries generally convert hard copy and analog data into digital format which is incorporated as part of overall system by information retrieval system. Since more importance is given to actual copy of the data source, it becomes a major issue in digital libraries and to some extent in information retrieval system. An access to digital information and the availability of information in digital format can be achieved by means of digital libraries.

Q23. Write brief note on 'Digital libraries' and 'Data warehouses'. May/June-13(R09), Q1(b)

OR

Discuss digital libraries. Dec.-14(R09), Q8(a)

(*Refer Only Topic: Digital Libraries*)

OR

What is data warehouse? Explain.

(*Refer Only Topic: Data Warehouses*)

Answer : Nov./Dec.-12(R09), Q1(b)

The two systems that are used most often for extracting relevant information are,

1. Digital libraries
2. Data warehouses.

1. Digital Libraries

A Digital library acts as a repository that stores massive volume of data. Apart from storing information, libraries are also responsible for extracting information based on the media, it has been developed on. Libraries are also termed as "electronic libraries' as they utilize electronic devices to perform their functions when the amount of information is increased. The concept of electronic libraries is growing rapidly because of emerging internet technology.

Later US government renamed electronic libraries as digital libraries. As Internet technology is growing exponentially, the digital libraries conception is also emerging.

2. Data Warehouses

A data warehouse is a special type of database used by the organizations for storing large amount of business data in such a way that retrieving desirable information from the system is very easy and reliable. Data in the warehouse is stored in different location so that it doesn't collide with transactional database system which is responsible for storing day-to-day information. This storage system is also responsible for providing solutions to sophisticated queries, which involves many computation at finer-level of granularity.

A data warehouse mainly comprises of the following elements,

(i) Data

(ii) Information Directory: This directory elaborates (expands) the data content and its semantic stored in the database.

(iii) Input Function: This function is responsible for capturing data and moving it to the storage system.

(iv) Data Search and Manipulation Tools: These tools enable the user to access and evaluate the data available in the data warehouse.

(v) Delivery Mechanism: This mechanism transfers the data to different warehouses or data marts.

Data warehouses differs from information storage and retrieval system because the former concentrates on structural databases and decision support-mechanisms. Apart from actual search process, there are different sets of analytical tools used to extract (mine) the data. Data mining is a process of extracting knowledge from massive volume of database. It refers to a way of finding significant and useful information from an organization's database. Data mining is not confined to a particular organization, instead it needs techniques to explore the knowledge hidden in the data. The different techniques used for digging out data are artificial intelligence, statistical and mathematical techniques and pattern recognition techniques.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Q24. What is meant by information retrieval systems? Discuss its relation to DBMS, digital libraries and data warehouses.

Answer : Nov.-13(R09), Q1

Information Retrieval Systems (IRS)

For answer refer Unit-I, Q13 [First Two Paragraphs Only].

Relationship between IRS System and DBMS

For answer refer Unit-I, Q19.

Relationship between IRS and Digital Libraries

For answer refer Unit-I, Q22

Relationship between IRS and Data Warehouses

For answer refer Unit-I, Q23 [Last Paragraph Only].

1.5 INFORMATION RETRIEVAL SYSTEM CAPABILITIES

1.5.1 Search Capabilities

Q25. Discuss about search capabilities in IRS. Oct./Nov.-20(R16), Q2(a)

OR

Discuss in detail the search capabilities of information retrieval systems.

Answer : Dec.-19(R16), Q3

Search Capabilities

The main purpose of the search capability is to provide mapping between the search query specified by the user and the items available in the database. The search query can either be a natural language text or a statement with boolean operators.

In some cases, the search query will be a combination of both the textual statement as well as boolean operators. Moreover, weights can also be assigned to the search terms, that is, the search terms are ranked according to their order of relevancy.

Example

Search all documents related to hardware peripherals (0.7) or software components (0.2) of a computer.

The rank value will always be within the range 0 and 1, where 1 indicates the most preferable or important item.

The various search capabilities in the information retrieval systems are,

1. **Boolean Logic**

   For answer refer Unit-I, Q26, Topic: Boolean Logic.

2. **Proximity Function**

   For answer refer Unit-I, Q26, Topic: Proximity Function.

3. **Contiguous Word Phrase**

   For answer refer Unit-I, Q27, Topic: Contiguous Word Phrase (CWP).

4. **Fuzzy Searches**

   For answer refer Unit-I, Q28, Topic: Fuzzy Searches.

5. **Masking**

   For answer refer Unit-I, Q28, Topic: Masking.

6. **Numeric and Date Ranges**

Normalization process facilitates the system to represent characters or words as numbers or dates, which inturn enables the processing of numeric or date ranges against those words.

The query may contain the inclusive operator (–) as well as the range operators (<, >, <=, >=) for implementing Numeric and Date ranges.

7. **Concept/Thesaurus Expansion**

For answer refer Unit-I, Q30.

8. **Natural Language Query**

For answer refer Unit-I, Q31(a).

9. **Multimedia Query**

For answer refer Unit-I, Q31(b).

Q26. Why the use of proximity improve precision more than the boolean functions? Justify your answer.

Answer :

Boolean Logic

In the boolean logic, the required information is obtained by logically merging various concepts together. Boolean functions are applicable to the units (Characters, words, sentences or paragraphs) irrespective of their locations in an item.

There are four types of boolean operators available which are AND, OR, NOT and XOR but in practise only three of them i.e., AND, OR and NOT are frequently used.

In boolean search statements, the parentheses must be correctly placed in order to specify the order of boolean operations because by default they follow NOT, AND and OR sequence.

Examples of Boolean Search Statements

1. The search statement *pen* **OR** *pencil* **NOT** *eraser* will return all those items which describe pens or/and pencils but not eraser.
2. The statement *pen* **OR** (*pencil* **NOT** *eraser*) will return all the items that describe pens or/and items that describes pencils and do not describe erasers.
3. The statement *pen* **AND NOT** *pencil* **OR** *eraser* will return all those items that describe pens but not pencils or erasers.

The '**M** of **N**' logic is a special type of boolean search in which user lists a collection of potential search terms and then accepts all those terms that have the subset of that collection.

Example

Consider the following terms '*XXX*', '*YYY*' and '*ZZZ*'. In order to identify all the items that contain any two of the above terms, the boolean search query will be ((*XXX* **AND** *YYY*) **OR** (*XXX* **AND** *ZZZ*) **OR** (*YYY* **AND** *ZZZ*)).

Proximity Function

It is used to define the maximum number of units that can be allowed between the two search terms within a particular item.

The idea behind proximity function is that the more close the two terms in the text, the more chances for them to be related. The precision of the search can be augmented using the proximity function.

Example for Proximity Concept

In any given item, if the terms HARDWARE and COMPONENTS are present with very few words in between them then there is a high probability that the item is describing components of hardware.

The syntax for proximity function is,

Term1 with 'i' 'units' of Term2

Where 'i' is an integer number, units can be given in terms of characters, words, sentences or paragraphs.

The well structured item uses the distance between the characters instead of words. An individual image can be located from an item containing cluster of images with the help of text present between images.

In some cases, a direction operator can also be employed which defines the direction (either before or after) the second term.

Adjacent operator (ADJ) is one of the proximity operators which typically makes use of distance operator of one and always operate in a forward direction.

If the distance operator is set to zero, the second term is within the same unit as of the first term.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Examples of Proximity Function

1. The search statement 'Computer' ADJ 'Design' will return all the items that describes Computer Design but not Design Computer.
2. "United" within four words of "America" may return "United states of America", "United Bankers Association of America", "United Airlines of America" but it will not return items with "United Petroleum Corporation and Oil Reserves of America".
3. The statement 'Bush within zero paragraph of USA' will return all those items which has Bush and USA in the same paragraph.

Q27. Show that the proximity function cannot be used to provide an equivalent to a contiguous word phrase.

Answer :

Proximity Function

Proximity function is used to define the maximum units that can be allowed between the two search terms. The units can be in terms of characters, words, sentences or paragraphs.

The syntax for proximity function is,

Term1 with 'i' 'units' of Term2

Where 'i' is an integer number, Term1 and Term2 are the search terms.

Contiguous Word Phrase (CWP)

It is a combinational method for specification of query as well as special search operator. In simple terms, a contiguous word phrase refers to a unit comprising two or more words.

Example

'Kingdom of Saudi Arabia'.

This CWP can be used with both the boolean and proximity operators therefore, the query 'Oil' AND 'Kingdom of Saudi Arabia' return all those items containing both 'Oil' and 'Kingdom of Saudi Arabia' that is word and contiguous word phrase respectively.

CWP is almost similar to the adjacency operator (ADJ) provided by the proximity function but gives additional specifications. Making use of adjacency operator on two individual words is equivalent to a two word contiguous phrase. However, Nested adjacencies of proximity and boolean operators are needed to create a search statement equivalent to a contiguous word phrase comprising two words or more.

Contiguous word phrases are also called 'literal strings' and 'Exact phrases' in WAIS and RetrievalWare systems respectively. CWP are 'N' ary operators whereas, proximity and boolean operators are binary operators. Moreover, Nested adjacencies are not supported by most of the commercial systems so it is not possible to provide a search statement equivalent to the contiguous word phrase.

Q28. What are the similarities and differences between the use of fuzzy searches and term masking? How are these searches used for compensating errors?

Answer :

Fuzzy Searches

Fuzzy searches enable the user to locate all those words whose spelling matches the given search term. The primary objective of this search function is to counterbalance the errors that occur in the spelling of words. Since, fuzzy searches include all the words with similar spellings, the recall will be more. However the precision will be much lower because the other terms which may not even be related to the query will also be searched if they have similar spelling. Consider the following example, if the term 'computer' is searched via fuzzy search function, this will include the terms like 'computer'. 'compute', 'compitter', 'computter' etc.

All the alternate spellings that appear similar to the query term are included.

Fuzzy searches are mainly used in systems that use Optical Character Read (OCRed) method for fetching the items. In this OCRed method firstly, an item is scanned to obtain a binary image then, this image is partitioned into meaningful segments where each segment consists of various sub-segments each of which in turn represents a single character. The OCRed method will then identify and transform the characters into ASCII format. The identification of characters is truly based on the quality of input (i.e, item) provided for scanning. Even if good quality input is supplied, the accuracy will be within the range of 90% - 99%, this implies that there are some characters that can be wrongly recognized. Fuzzy searches compensate for these erroneous characters.

Masking

Masking is the process of hiding some portion of the query term such that all the search terms that matches with the unhidden portion of the query term are treated as valid and accepted.

Query term masking is of two types. They are,

(i) Fixed length and

(ii) Variable length.

Fixed length masking provides a mask only for a single position in a word. The mask provided may be either to a symbol in a particular word or lack of a particular position in a word.

Fixed length masking is rarely implemented because it doesn't provide any fundamental functionality to the system.

Example

Search statement for "Multi$millionaire" returns 'Multi-millionaire', 'Multiemillionaire', 'multimillionaire', but not 'multi millionaire' because it consists of two words.

Variable length masking enables the user to mask any number of characters within a particular word. The mask can be applied either at the front or at the back or at both the front and the back (Imbedded mask).

The search in which the mask is applied at the front is known as suffix search, the mask applied at the back is known as prefix search and the mask applied at both the ends of a word is known as an imbedded string search.

Example

1. **Suffix Search**

   '*star'→ matches 'superstar', 'semi-star' etc.

2. **Prefix Search**

   'star*'→ matches 'starter', 'stardom', start etc.

3. **Imbedded String Search**

   '*star*' → matches 'multi-starrer', 'super-starrer', 'kick-start' etc.

Variable length masking is oftenly used and is the default searching mechanism applied by most of the search systems.

Q29. What operations takes place in the system when the following search statements are entered,

(a) Multi$national

(b) *Computer*

(c) Comput*

(d) *Comput*.

Answer :

(a) Multi$national

Whenever 'multi$national' statement is entered, the search returns the matches like 'multinational', 'multi-national', 'multi$national' but not multi national because it consists of two words (tokens).

(b) *Computer*

When this statement is entered, the search returns the matches like minicomputer, microcomputer, computer or computerized i.e., all those words which include the word computer.

(c) Comput*

If 'comput *' is entered into the search it returns the matches like computer, computing, computes etc. That is, all those words starting with "comput".

(d) *Comput*

Searching this statement (i.e., * comput *) will result in the following matches, microcomputers, minicomputing, compute etc.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Q30. Are thesauri a subclass of concept classes? Justify your answer.

Answer :

Thesaurus is simply an extension of a particular term to one or more levels such that the other level terms must have a similar meaning as that of an initial term.

A concept class represents a hierarchical structure in which each meaning of a word is explored into different concepts related to that word. Sometimes, concept classes are used as network hierarchy that connect the words.

The representations of a concept class are useful for those users, who have little information about the concept domains. A concept based database also shows those relationships that are not shown by language based thesaurus.

Example

'Top scorers' may be linked to the 'newly announced Results' in the concept database even though they are not synonyms.

Thesauri can either be dependent on statistics or a semantic thesauri. The semantic thesauri is a collection of words along with their synonyms. RetrievalWare system and topic system provides commonly available language based electronic thesaurus. A user can manually eliminate all unnecessary synonyms by indicating the terms to be utilized in a query out of all the synonyms available.

The thesaurus or concept trees can be searched thereby inserting additional terms and their associated relationship. This feature enables users to search for the information relevant to their area of interest by providing specific terms.

The major drawback associated with thesauri is its generic nature. It is language generic and hence, it produces many different terms irrespective of those available in the document database. To overcome this drawback, another approach of thesauri is available, in which a database or its prototype is used to produce statistically related terms. This approach is completely based on the database being browsed.

However, it may not be compatible with other databases. Statistical thesaurus classes are very difficult to name and understand because they are automatically extended by eliminating the need of user interaction with the thesaurus.

In other words, concept trees and thesauri can be used to either broaden the search by adding additional terms to the search statement or replacing the terms with more specific terms to obtain specific results. From this point of view, we can say that broadening the search will decrease the precision because of the increase in recall and replacing the terms increase the precision because of the decrease in recall of the search. Mostly generalization process is used for extension of the search statement with additional terms.

Q31. Write short notes on,

(a) Natural language queries

(b) Multimedia queries.

Model Paper-II, Q2(b)

Answer :

(a) Natural Language Queries

Instead of using a specific boolean query, a user can simply enter a prose statement which give the description of the required information. This can be done with the help of natural language queries. The accuracy of the result depends upon the length of the prose statement. That is, the larger the prose statement the more accurate the result.

It is difficult to specify negation in the search statement using natural language query. Moreover, making the system to recognize it is even more difficult.

The prose statement is processed by the system in a manner similar to that of the query. All the mechanisms applied for the identification of the items similar to the search statement are also applicable for the identification of other items, but these mechanisms do not include conditions for excluding some part of the search statement.

Relevance feedback is a function associated with natural language queries. The user must not input the natural language query, instead it must identify all the items that appears similar to that query. This capability enables the user to search for items by simply providing some portion of the item as the search statement.

In order to include negation function most of the system make use of a user interface that can accommodate the capabilities of both the boolean logic as well as the natural language. Such that the boolean logic handles the negation function. Information retrieval system does not provide this user interface.

Whenever the negation function is required, natural language interface will increase the recall of the systems thereby decreasing precision.

(b) Multimedia Queries

The introduction of the multimedia items further enhance the complexity of user interfaces.

Now-a-days, still images can be used for multiple purpose like searching images that are portions of an item locating a specific scene in a video etc.

The scene that changes in a video are considered as a sequence of images. If the video consists of static text then OCRing is done for that text. In order to search for the audio, we need to simulate the audio segment before looking for a match. An alternative approach is audio transcription, which can be performed on the audio to convert it to textual format such that queries can be applied to the resultant text.

The resultant text data (Transcribed Audio) contains many different errors which must be allowed by the search algorithms. Unlike Optical Character Reading (OCR), all the errors of Automatic Speech Recognition (ASR) are treated as valid words because these words are selected from a dictionary. Whereas, in the OCR the error some words may not be valid because the words are created by reading the characters.

1.5.2 Browse Capabilities

Q32. What are browse capabilities in information retrieval systems? Explain the concepts of zoning and highlighting in information system.

Answer :

Model Paper-II, Q3(b)

Browse Capabilities

After the completion of the search, browse capabilities enable the user to identify and display the required items. Items summary can be displayed in two ways,

1. Line the status of items
2. Data visualization.

The user can select the items and the zones from either of these summary displays. Moreover, the user can easily change from one display to another and also between the items.

Browse capabilities facilitates the user in selecting the most relevant items from the search resultant items because the search returns many irrelevant items in addition to the relevant ones.

Zoning

The user will always prefer to view minimal information in order to determine the relevancy of an item. After the item is identified to be potentially relevant, the complete item must be displayed for a detailed review.

Since, the display screen is of limited size only selected portions of the items can be displayed at a time such that more number of items can be covered within a single display screen. Thus, relevancy of multiple items can be determined simply by scanning a single image.

For example, relevancy of an item can be predicted by viewing title and abstract information about an item. Focusing only on these two zones will display multiple items using a single display screen which in turn augments the speed of identification of relevance of multiple items simultaneously.

Locality and passage based search and retrieval approach is similar to zoning and is used for minimizing the review information needed by the user.

In this approach, the item is partitioned into various passages that are indexed and where the passage boundaries are dynamic. Hence, instead of displaying the complete item, a particular passage or locality is displayed, which can be expanded to retrieve the complete item.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Highlighting

The highlighting is a mechanism by which the reason for selection of a particular item is indicated. Highlighting enables the user to concentrate only on the highlighted parts of the text for tracing the relevancy of an item.

There are different phases of highlighting that is, the strongly highlighted text indicates how strongly the text have taken part in the selection of an item. Similarly, the weakly highlighted text indicates how weakly the text have served the purpose of item selection.

In many systems, highlighted text within the item is displayed first and then the next highlighted text and so on. The RetrivalWare search system has DCARS system as a frontend which enables the user to search for the item by providing individual words (or) paragraphs as an input.

Boolean systems make use of highlighting to indicate the reason for retrieval because of the direct mapping between the search terms and the terms in the items.

In some cases, the highlighted terms may not map with any of the search terms entered. This can happen while using natural language processing and may lead to serious problems.

The ranking system involves ranking of different terms based on their level of contribution in making a decision for retrieving an item. Highlighting can be done using different colors and intensities. Where the more darker shade (intensity) refers to the most relevant word in an item.

Q33. 'Ranking is one of the most important concepts in Information Retrieval Systems'. What difficulties are encountered in applying ranking to boolean queries?

Answer :

In boolean systems, the total number of items retrieved by a query represents the status. Each and every item retrieved satisfy all the condition of a boolean query. Highlighting can be done on retrieved item to specify the reasons for the selection of those items.

Ranking based upon the prediction of relevance values augment the status report to include relevance value of each item along with its brief description.

The search system estimates a relevance value for an item based upon how well the item satisfies the conditions of a search statement. The relevance value will be normalized within the range of 0.0 to 1.0 where 0.0 refers to non-relevant and 1.0 refers to the most relevant item.

Typically, the search system contains a minimum default value that blocks the returning of all those items whose relevance value is below this default value. This default value can be altered by the user whenever required.

The selection and arrangement of the output can also be done by using collaborative filtering option in which the system collects the user's feedback on items relevancy and then uses this information to arrange output for similar queries in future.

A portion of a zone is assigned for displaying relevance weight of the item which in turn shortens the zone, because each item is defined by atleast one line in the summary display. This portion is usually the title and provides relevance weight along with the additional information to the user such that non-relevant items are not selected. Providing the range for relevance number is much easier for a user rather than providing the exact relevance number.

The graphical visualization technique that displays the relevance relationship of hit items can also be implemented, in which items are grouped according to the topics and user can easily navigate through these groups. Hence, by means of this visualization techniques the user can get all the required items within a single image.

1.5.3 Miscellaneous Capabilities

Q34. How vocabulary browse and thesauri/concept classes are related to each other?

Answer : **Model Paper-I, Q3(b)**

Vocabulary Browse

The vocabulary browse enables the system to display the words stored in a document database in an alphabetical order.

In vocabulary browse, whenever the user enters a word or a part of a word, the system returns the dictionary around that word.

Example

If the user enters a word 'Product', then the system alphabetically returns the word entered along with some other words contained in the document. The user can scroll upwards and downwards to view the additional words which are displayed along with the entered word.

Term	No. of Occurrences
Prodigy	2
Produce	20
Producer	12
Product	1258
Production	165
Productivw	1
Productive	1475
Productivity	247
Profane	21

Figure: Vocabulary Browse List with 'Product'

Vocabulary browse serves the purpose of both the fixed length as well as variable length masks. It also identifies mis-spellings.

In the above example, even if the user enters the word 'Productivw' instead of 'Productive' the search result will not be effected.

Vocabulary browse shows the number of occurrences of a given term in various documents. Therefore, ORed term cannot be used to make the query more focused because it requires additional ANDed terms. Moreover, if 'OR' term is used, the number of hits displayed will be excessive.

Both the vocabulary browse and the thesauri introduces various search terms that may not even be found in the document database.

Concept/Thesaurus Examples

For answer refer Unit-I, Q30.

Q35. Write short notes on the following,

(a) Iterative search and search history log

(b) Role of multimedia in information display

(c) Canned query.

Model Paper-III, Q3(b)

OR

Demonstrate canned query with example.

(Refer Only Topic: Canned Query)

Answer :

(a) Iterative Search and Search History Log

Oct./Nov.-20(R16), Q2(b)

Iterative search refers to the process of refining the previous search result in order to obtain the relevant items, that is, whenever a user enters the query onto the search, many items are returned in a Hit file. This Hit file contains a collection of items most of which are non relevant to the user's query. Instead of writing a new query, this resultant item collection can be refined by inserting additional statements to the previous query. The additional statements can be added to the query in an AND condition. Iterative search can be applied even when relevance feedback method is being implemented.

Search history log keep track of all the search queries that were executed along with the results obtained. So that, previous search results serve as a base for the new search.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

(b) Role of Multimedia in Information Display

The display of multimedia content is difficult and often requires complex procedures. Grouping of hit files and the use of graphical display will not prove to be efficient in case of multimodal information. Instead of using the above techniques, textual aspect of multimedia can be used.

In order to display the requested image, the thumb nail view of the image is displayed along with the hit item, as a result the length of the hit goes beyond one line so the number of hits that can be displayed on a single screen will be reduced.

In audio searches, the human related processing problems must be considered. If the search result is also audio then the processing speed of listening decreases. The implementation of transcribed audio will be a major enhancement for users, the user can listen to audio and can review the transcribed text. Both these operations can be done at the same time.

The study on the items revealed that, the combination of the transcribed audio reviewing and original audio listening have reduced the processing time to half the actual time needed with only audio listening.

In textual source, the processing is done by using equipments. Moreover the text data can be easily and quickly scanned by the users.

The transcribed text for the following audio that is "still alone looking for someone" will be 'SALFS'. This transcribed text can also be used as an index for future retrieval of the above audio. The problem with the transcribed audio is that, most of the users must be capable of exploring the transcriptions so that, they can include additional information like tone of the original audio.

Ranking of multimedia output is often a complex task that is, a hit file containing different modalities like text, image and audio needs an aggregate weight to be assigned for each hit region. Aggregate weight is the combination of all the weights assigned to individual modalities and depends upon their level of satisfying the query.

(c) Canned Query

The canned query can also be termed as a stored query. It allows the user to store the query with a particular name so that it can be retrieved and executed in future sessions. Consider a situation where the user needs to repeatedly search a particular area at regular intervals, then instead of writing the same query again and again, the user can simply create a canned query and use it whenever needed. Additional search statements can also be added to the canned query.

Example

Consider an organization with different branches, for which the administrator wants to retain the salary information of all the employees. Instead of writing a separate query for each and every branch, the administrator will create a canned query and then modify some of the statements like branch_name, branch_id etc., to retain the information.

The reusability feature of the canned query make the search more accurate and comprehensive. Canned query also supports the insertion of variables into the query so that the binding of variables takes place during the query execution.

UNIT 2

Cataloging and Indexing, Data Structure

Syllabus

Cataloging and Indexing: History and Objectives of Indexing, Indexing Process, Automatic Indexing, Information Extraction.

Data Structures: Introduction to Data Structure, Stemming Algorithms, Inverted File Structure, N-Gram Data Structures, PAT Data Structure, Signature File Structure, Hypertext and XML Data Structures, Hidden Markov Models.

Learning Objectives

- Changed Objectives of Indexing based on its Evolution
- The Process of Indexing and Automatic Indexing
- The Process of Extracting Information
- Concept of Data Structures
- Use of Stemming Algorithms in IRS
- Various Data Structures Like Inverted File Structure, N Gram, PAT, Signature, Hypertext and XML Data Structures
- Concept of Hidden Markov Models.

Introduction

The process of finding the contents of items, so as to make their retrieval easy is known as "Indexing". It is the oldest technique for finding the items and their contents. Originally, indexing is called as "Cataloging". The evolution of Information Retrieval Systems have changed (effected) the objectives of indexing. Earlier the full text of an item, available in a searchable form was used in finding the rules for manual indexing but the document file acts as a full text searchable data structure and provides a new type of indexing called "total document indexing". The two major data structures that are used in any information system are, the first data structure contains the received items in the normalized form. The second data structure contains the processing tokens and the related data structure to support search functions.

PART-A SHORT QUESTIONS WITH SOLUTIONS

Q1. Define indexing and List types of indexing.

Answer :

(Model Paper-II, Q1(c) | Dec.-19(R16), Q1(c))

Indexing

The process of finding the contents of items, so as to make their retrieval easy is known as "Indexing". It is the oldest technique for finding the items and their contents. Originally, indexing is called as "Cataloging".

Types of Indexing

Indexes can be stored in three different forms or types.

1. Direct index
2. Document index
3. Lexicon index
4. Inverted index.

Q2. Define automatic Indexing.

Answer :

Model Paper-I, Q1(c)

The ability of a system to automatically identify the index terms that must be assigned to an item is called automatic indexing. In the simplest scenario of implementing automatic indexing all the words in the document are used as index terms. The complex scenario results when the goal is to perform like a human indexer and find a limited number of index terms for the vital concepts in the item.

Q3. How does the process of Information Extraction differ from the process of Document Indexing?

Answer :

(Model Paper-III, Q1(c) | Dec.-19(R16), Q1(d))

The processes of information extraction are,

- Facts determination
- Text extraction.

Facts determination deals with the identification and extraction of important facts in an item. These facts are then sent to the database and reside in the structured fields. In text extraction, the extracted text is used to summarize the data. The summarization includes representation of major concepts in an item in the form of a summary. However, the process of document indexing provides representation of documents for their faster and effective retrieval. Moreover, it associates the information with a file which is utilized by document management system.

Q4. List three disadvantages of human indexing over automatic indexing.

Answer :

The three disadvantages of human indexing over automatic indexing are,

1. Cost
2. Consistency and
3. Processing time.

Q5. What is unweighted indexing?

Answer :

Unweighted indexing system keeps the existence of an index term and its location/locations as a part of the searchable data structure. In the representation of concepts in the item, the values of index terms are not discriminated. Similar type of commercial systems were used in 1980's. In unweighted systems, queries are based on boolean logic and resultant Hit file's items have identical values.

Q6. Define multimedia indexing.

Answer :

Indexing used for multimedia differs from all other types of indexing. Automatic indexing is done by using a number of passes of information.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

- First pass converts the analog input to digital form.
- Algorithms are applied to the digital information, in order to retrieve the unit of processing of different modalities. The units are then used to represent the item.
- A final processing of the units results in extraction of searchable features of the unit.

Multimedia indexing can be done at three different levels,

(i) Raw data level

(ii) Feature level

(iii) Semantic level.

Q7. Define,

(i) Over generation

(ii) Fallout.

Answer :

(i) **Over Generation**

The amount of unnecessary information that is extracted is given by "overgeneration". It can be a result of filling the slots and topics with irrelevant data and templates respectively.

(ii) **Fallout**

The amount of incorrect assignment of slot fillers made by a system with an increase in incorrect slot fillers is called "fallout".

Q8. What are dictionary look-up stemmers?

Answer : **Model Paper-I, Q1(d)**

Dictionary look-up stemmers is another approach for determining a stem apart from those approaches that are fully dependent on certain algorithms such as Porter stemming algorithm. However, in this mechanism some stemming rules need to be applied, which are generated from those that have minimum number of exceptions. When the actual terms or stemmed version of it is entered, then a search is made on all the entries available in the dictionary. This is done so as to look-up for the exact match and substitute that term with the best suited stem. This approach is basically applicable in the following systems,

1. Inquery system
2. Retrievalware system.

Q9. What are successor stemmers?

Answer :

Successor stemmers depends upon the length of prefixes and upon an analogy in the structural language, which examine words and morpheme boundaries depending upon the distribution of phonemes. The phonemes are a small part of speech that differentiates one word from another. The successor stemmer identifies the successor varieties for a word so that it can be divided into segments and one of the segments of a word is selected as stem.

Q10. Define inverted file structure.

Answer : **Model Paper-II, Q1(d)**

Inverted file structure is a common data structure that is used in database management as well as in information retrieval systems. The inverted file structures are divided into the basic files,

- Document file
- Inversion lists or Posting files
- Dictionary.

The inverted file is called so because it has an ability to store an inversion of documents. A unique numerical identifier is given to every document in the system. Hence, the documents are represented as : document #1, document #2, document #3 and so on.

Q11. What are N-grams?

Answer : **Model Paper-III, Q1(d)**

N-grams are the data structures that are used in information systems. They are considered as a special technique for stemming. They are not concerned about the semantics of the word. Instead, they depend upon the fixed consecutive series of '*n*' characters. The data structure is converted into overlapping *n*-grams so that a searchable database can be created. Some systems allow the usage of interword symbols like #, for *n*-gram when *n* is greater than two. This symbol replaces a blank, period, semicolon, colon in the *n*-gram set. The n-grams that are created are a sequence of searchable processing. Also the similar *n*-grams can be created a number of times from a particular word.

PART-B ESSAY QUESTIONS WITH SOLUTIONS

2.1 CATALOGING AND INDEXING

2.1.1 History and Objectives of Indexing

Q12. Define Indexing and explain about it.

Model Paper-I, Q4(a)

Answer :

Indexing

The process of finding the contents of items, so as to make their retrieval easy is known as "Indexing". It is the oldest technique for finding the items and their contents. Originally, indexing is called as "Cataloging".

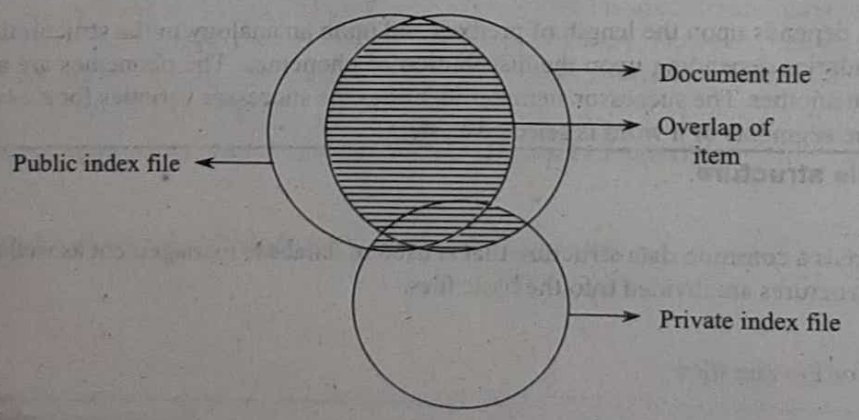
Overlap of Item

The value of a concept is not always given by the words used in an item. However, it is given by the combination of words and their semantic implications. "The User Needs" also helps in finding the utility of a concept. These needs will be considered by the public file indexer. The public file indexer takes into consideration the needs of all the users of the library system. Each user has his/her interest in a specific domain. Thus, bounding the concepts to a specific domain. To know whether a concept should be indexed or not, manpower is required to evaluate the quality of that concept. "User Needs" defer from individual users to library class of indexers, hence making "private index files" is a vital part of a good information system.

The private index file allows a user to make the total document file a subset of the folders the user is currently interested in and in the ones the user may be interested in future. The user also gets an opportunity to decide the utility of a concept. The decision is based on the user need rather than the system need. The precision of a search is increased by performing selective indexing using the value of concepts.

If the full document indexing is available then the indexer does not enter the index terms which are same as the word in the document. Public index files may be used by the users as a constituent of their searching criteria so that the recall can be increased. Similarly the precision of a search can be increased if users limit.

Search to their private index file, the below figure depicts the relationship between the use of words to define the concepts.



Figure

In order to achieve abstraction, the public indexing of the concept inserts more index terms over the words in an item. However, the index file will use less number of terms as only the important concepts are indexed. Private index files limit the number of items to be indexed to,

(i) The items that have value to the user.

(ii) The items that are in user interested domain.

Thus, there is an overlap between the items of a public index file and a private index file.

Electronic indexes to items is a good method of finding relevant information and assists its users in most of the applications.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Q13. Discuss the objectives of Cataloging and Indexing.

Answer : May/June-13(R09), Q2(b)

Objectives

The evolution of Information Retrieval Systems have changed (effected) the objectives of indexing. Earlier the full text of an item, available in a searchable form was used in finding the rules for manual indexing but the document file acts as a full text searchable data structure and provides a new type of indexing called "total document indexing". In this, the index descriptors of the subject(*s*) of an item are the words in that item.

The need for manual indexing looks absurd when total document indexing is available. However there are some valid reasons for its requirement. They are as follows,

1. Controlled Vocabulary

It refers to the finite set of index terms from which all index terms should be selected. This vocabulary acts as a domain for the term indexes. Manual indexing uses the controlled vocabulary so that the search process is simplified. But ,there is an adverse effect of using a controlled vocabulary due to which the indexing process becomes slower. The reason for such an adverse effect is, the indexer who tries to find some index terms for concepts which are not specifically present in the controlled vocabulary set. The effect of using a controlled vocabulary set can be reversed by using an uncontrolled vocabulary set because uncontrolled vocabulary set makes the process of indexing faster and search process difficult.

2. Items in Electronic Form

The objectives of manual indexing changes due to the availability of items in an electronic form. Extracting the source information or citation data can be done automatically. However using indexes for index term standardization is still needed. The need for controlled vocabularies can be reduced by using modern systems along with the automatic use of some reference databases and thesauri.

Despite the above two reasons, total document indexing rules manual indexing. The reason is that, most of the concepts in a document are locatable by making a search of the total document index. Thus, shifting the use of manual indexing to the abstraction of concepts and judgements on the value of the information. Consistent abstraction on all the concepts in an item is not possible even by the automatic text analysis algorithms. Thus, enhancements in information retrieval systems along with the artificial intelligence techniques are applied.

2.1.2 Indexing Process

Q14. Explain in detail the indexing process for IRS with a neat diagram.

Answer : (Model Paper-II, Q4(a) | Dec.-19(R16), Q4)

In an organization consisting of multiple indexers, the decision about the creation of a public or a private index is based on some procedural decisions of creating index terms that aids the indexers and end users in determining what is expected from the index file. The two procedural decisions are as follows,

(i) Indexing scope

(ii) Linking of index terms.

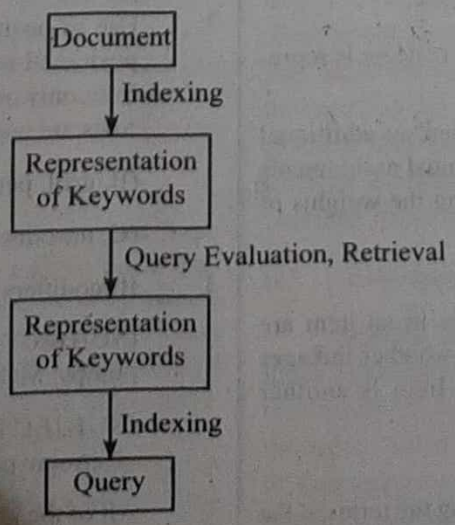


Figure: Indexing Process

(i) Indexing Scope

Scope of indexing defines the level of detail the subject index may contain. Manual identification of bibliographic terms representing the concepts of an item, is very difficult. The reason is the difference in interaction between the author and the indexer. Vocabulary domain of an author may differ from that of the vocabulary domain of an indexer. The indexer should be aware of the level at which indexing should continue or stop.

Exhaustivity and specificity are the two factors that help in deciding about the level at which the concepts of an item can be indexed.

(a) Exhaustivity

The degree to which different concepts of an item are indexed is called exhaustivity.

(b) Specificity

Specificity deals with the preciseness of an index term that is used in indexing. For instance, the term processor or intel should be used depending on the specificity decision.

If indexing is done only on important concepts in an item using general index terms then specificity and exhaustivity becomes low. Thus, minimum number of index terms for each item is desirable. For instance, in order to index this page it would need only the index term "indexing". If exhaustivity and specificity are high then this page would need the index terms "exhaustivity" "specificity", "indexing" etc. Precision and recall are badly effected by low exhaustivity. However, low specificity has a bad affect only on precision but not on recall.

Finding the portion of an item to be indexed is another decision in indexing. The best solution to such a decision is limiting the process of indexing to either title region or both title and abstract regions. Thus, indexing only that material which is considered to be important by the author reduces the cost of indexing an item. But, both recall and precision are lost.

The process of giving significance to the usage of an index term in an item is called weighting. However, it is not commonly used in manual indexing system. The weight should,

(i) Try to minimize the number of times the concept is used in the items stored in the database.

(ii) Indicates how clearly an index term's concept is represented in an item.

If weights are assigned manually, then an additional responsibility is imposed on the indexer. Manual assignments also need a complex data structure for storing the weights of the index terms.

(ii) Linking of Index Terms

The related attributes of the concepts in an item are correlated with the help of linkages. Finding whether linkages are present among the index terms of an item is another important decision in indexing.

(a) Pre-coordination

The process of creating linkages among the terms at the time of index creation is called pre-coordination.

(b) Post-coordination

The process of creating linkages among the terms after the indexing process is called post-coordination. "AND ing" of index terms is used to implement post-coordination. It helps in finding only those indexes that contain all the search terms.

In the linkage process, the factors that should be identified are,

(a) Ordering constraints/limits on the linked terms

(b) Additional descriptors for the index terms

(c) The number of terms that can be related.

The number of terms that can be related is not an important implementation issue. The design of the indexer's user interface is only effected by the range of terms to be linked. If a large number of terms are used then the chances for the relationships between them is more. The additional role descriptors for the index terms can be given either by the order of terms or by using a modifier. If order of terms is used then there is a limit on the number of index terms that can have role descriptors. However, if modifiers are used then there is no such limit.

Types of Linkages

To know about the types of linkages, consider an example scenario. An item discusses the drilling of petrol pumps in country 'A' by country 'B' and the introduction of electronic petrol dealers in country 'D' by country 'C' when the terms are not linked then the index terms are petrol pumps, A, B, electronic dealers C, D, drilling etc.

1. If terms are linked at index creation time (precoordination). Then the index term sets are,

   (petrol pumps, A, drilling, B)

   (C, electronic petrol dealers, D, introduction)

   However, this type of linkage does not specify which country is introducing electronic petrol dealers in which country.

2. Use of positional roles can eliminate the ambiguity. A positional role treats the data as a single vector and allows only one value for one position. Thus if positional roles are used, then the index terms sets are,

   (B, drill, petrol pumps, A)

   (C, introduction, electronic petrol dealers, D)

3. If modifiers are used, then the index terms sets are,

   (SUBJECT: B; ACTION: drilling; OBJECT: petrol pumps; MODIFIER: in A)

   (SUBJECT: C; ACTION: introduces; OBJECT: electronic petrol dealers; MODIFIER: in D)

All of the above types of linkages use pre-coordination indexing.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

2.1.3 Automatic Indexing

Q15. Define automatic indexing. What are the benefits of using automatic indexing? Also discuss the two classes of indexing resulted from automatic indexing.

Model Paper-III, Q4(a)

OR

Discuss the different classes of automatic indexing.

Nov./Dec.-12(R09), Q3

(Refer Only Topic: Types of Automatic Indexing)

Answer :

Automatic Indexing

The ability of a system to automatically identify the index terms that must be assigned to an item is called automatic indexing. In the simplest scenario of implementing automatic indexing all the words in the document are used as index terms. The complex scenario results when the goal is to perform like a human indexer and find a limited number of index terms for the vital concepts in the item.

Advantages of Human Indexing Over Automatic Indexing

1. The ability to find concept abstraction.
2. The ability to judge the concept value.

Disadvantages of Human Indexing Over Automatic Indexing

1. Cost
2. Consistency and
3. Processing time.

Advantages of Automatic Indexing Over Human Indexing

The disadvantages of human indexing can be overcome by using automatic indexing. The advantages of automatic indexing are as follows,

1. Cost

The automatic indexing do not incur any additional indexing costs if, the hardware cost is amortized. The cost is included as a part of maintenance and normal functioning of a computer. But, in human indexing, an additional indexing cost (salaries and benefits) are to be regularly paid to human indexers.

2. Consistency

Consistency lacks when human indexing is used. They produce different or multiple indexing for a given document. For example, in an experiment in TREC-2, 20% difference was observed in judging a single item generated in different documents by human indexers. However, the advantage of using automatic indexing is predictability of algorithms. This advantages in turn maintain consistency in the index term selection process. The reason for such an advantage in automatic indexing is a skilled researcher who makes a deep study and understands the complete automatic process and finds its advantages and dis-advantages.

3. Processing Time

The time taken by human indexers to process an item varies significantly. The variation is due to,

- Exhaustivity and specificity rules.
- Indexer's knowledge about the concept being indexed.
- The accuracy and amount of preprocessing through automatic file build.

Even for the short items containing 400-600 words the processing time is 5 minutes for every item. A major part of this processing time is the time taken by human indexer to communicate or interact with the computer.

However, if automatic indexing is used then the processing time is just a few seconds. It can be even lesser if the complexity of the algorithm used to produce the index and the size of the processor are properly given.

Types of Automatic Indexing

The two classes or types of automatic indexing are,

(i) Weighted indexing

(ii) Unweighted indexing.

(i) Weighted Indexing

In weighted indexing, the concept value in a document is placed on the representation of the index term. The weight of an index term depends on a function. Thus, a function represents frequency of appearance of a term in an item. Among the pioneers of automatic indexing, Luhn says that "the importance of a concept in an item is proportional (directly) to the frequency of the use of the word that is related to the concept in a document". He introduced the concept of "resolving power" of a term. Similarly, studies made by other pioneer professionals illustrates the poisson distribution of "concept bearing" words (i.e., the words are not randomly distributed) and the occurrence of words "dump" within the items. If automatic indexing tries to retain the original text of an item or map the item to a totally different representation, then another form of indexing called "concept indexing" is introduced.

(ii) Unweighted Indexing

Unweighted indexing system keeps the existence of an index term and its location/locations as a part of the searchable data structure. In the representation of concepts in the item, the values of index terms are not discriminated. Similar type of commercial systems were used in 1980's. In unweighted systems, queries are based on boolean logic and resultant Hit file's items have identical values.

Q16. Write short notes on the following,

(a) Indexing by term

(b) Indexing by concept and

(c) Multimedia indexing.

OR

Illustrate indexing by term with example. Oct./Nov.-20(R16), Q3(a)

(Refer Only Topic: Indexing by Term)

Answer :

(a) Indexing by Term

Indexes can be created by using either,

(i) Statistical technique

(ii) Natural language technique.

These two techniques can be used only if the terms of original item are used as the basis for the indexing process.

(i) Statistical Technique

These techniques are based on vector probabilistic models and a special case of probabilistic models called Bayesian Models. The models are categorized as statistical because the weight calculations in these models use statistical information.

❖ **Vector Model**

Dr. Gerald Salton created SMART system in which an association is made between weighted systems and vectorized information system. SMART system focuses on weights and store them in the form of a vector. The vector in SMART system represents a document and each vector position represents a unique word in the data base. The weight of a term is given by the value assigned to a position. If the value is zero (0), then it means that the word does not exist in the document. Queries to SMART system are given in the form of vectors. This translation helps the search process by evaluating the distance between the document vectors and a query vector.

❖ **Bayesian Model**

Most dominant approaches use the special case of probabilistic models called the Bayesian model. This model is based on the "theories of evidential reasoning". This approach is usually applied to the retrieval process, however can be applied to the weighting process. A Directed Acyclic Graph (DAG) is used to represent a Bayesian network in this DAG,

(i) A random variable is represented as a node and

(ii) The probabilistic dependence between a node and its parents is given by an arc between these nodes.

The below figure depicts a two-level Bayesian Network,

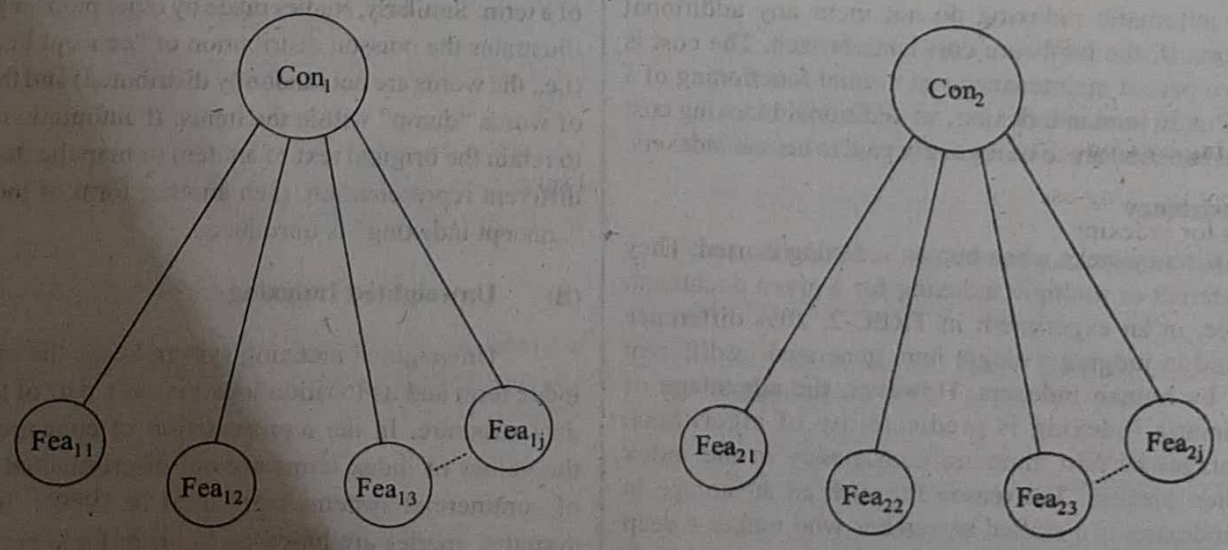


Figure: DAG Representation of a Two-level Bayesian Network

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

In the above figure,

1. Nodes Con_1 and Con_2 means the item has concept Con_m.
2. Nodes Fea_{11}, Fea_{12},...Fea_{1l}, Fea_{21}, Fea_{22},...Fea_{2l} means " the item has feature Fea_k".

The goal of the Bayesian network is to evaluate the probability of Con_m for a given Fea_{kl}. For this, the two probabilities that are required are,

(i) The conditional probability that the features Fea_{kl} are available in an item, provided that the item has topic 'Con_m'.

(ii) The prior probability that an item is relevant to the concept 'Con'.

In mathematical terms,

$P(Fea_{kl}/Con_m)$ is the conditional probability for $l = 1$.

$P(Con_m)$ is the prior probability

The task of automatic indexing is to evaluate posterior probability which is the probability of containing a concept 'con_m' by a item for the features 'Fea_{kl}'. Mathematically it is expressed as follows,

$P(Con_m/Fea_{kl},.....Fea_{ki})$ is the posterior probability.

To evaluate posterior probability the Bayes inference formula is,

$$P(Con_m/Fea_{kl},.....,Fea_{ki})$$

$$= \frac{P(Con_m)P(Fea_{kl},...., Fea_{kl}/Con_m)}{P(Fea_{kl},...., Fea_{ki})} \quad ...(1)$$

However, equation (1) can be simplified to linear decision rule as,

$$S(Con_m/Fea_{kl},..., Fea_{ki}) = \sum_j I_n(Fea_{kj})\, W(Fea_{kj}, Con_m)$$

Where,

- S is a function given by the sum of feature weights.
- W is a coefficient associated with a specific pair of feature/concept.

$I_n(Fea_{kj})$ is an indicator variable.

$$I_n(Fea_{kj}) = \begin{cases} 1 & \text{if } Fea_{kj} \text{ is in item} \\ 0 & \text{otherwise} \end{cases}$$

(ii) Natural Language Technique

Natural language processing is another approach of defining indexes to items. The Document Retrieval through Linguistic Knowledge (DR-LINK) system processes items at different levels. These levels are lexical, semantic, morphological, syntactic and discourse. Each level performs additional analysis by using information from the previous level. For example, the discourse level identifies the abstract concepts by applying the predefined models of event relationships. This is how the discourse level abstracts the information beyond the sentence level. However, normal indexing is poor at finding and extracting verbs. It also fails to identify and extract the object relationships using the verbs in the best way.

(b) Indexing by Concept

Concept indexing uses a test set of terms to find a canonical set of concepts. After finding the set, concept indexing use it as the basis for indexing all the items. It is also known as "latent semantic indexing". An example system is match plus which uses concept indexing. This system was developed by HNC Inc. Machine learning of word concept relationships and sensitivity to similarity of use are facilitated by match plus system using neural networks. The system uses two neural networks.

- One neural network to modify a query in accordance with the user's feedback.
- Another neural network learning algorithms to output the stem context vectors which are sensitive to "similarity of use".

A context vector is a high dimensional vector that represent queries, items and word stems. The vector may be at least 300 dimensions or more. In this vector, a dimension may be viewed as an abstract concept class. In a study made by Pollack and Waltz, a context vector is defined by selecting a set of 'm' features. A context vector C^i is an m-dimensional vector for a word stem 'i' such that,

$C^i > 0$, if 'i' is strongly connected with feature k

$C^i < 0$, if 'i' contradicts feature k

$C^i \approx 0$, if 'i' is not connected with feature k

The above conditions are defined for each kth component,

The components of concept vectors are interpreted to be same as the weights in neural networks. Each feature is viewed as an abstract concept class.

(c) Multimedia Indexing

Indexing used for multimedia differs from all other types of indexing. Automatic indexing is done by using a number of passes of information.

- First pass converts the analog input to digital form.
- Algorithms are applied to the digital information, in order to retrieve the unit of processing of different modalities. The units are then used to represent the item.
- A final processing of the units results in extraction of searchable features of the unit.

Multimedia indexing can be done at three different levels,

(i) Raw data level

(ii) Feature level

(iii) Semantic level.

Example

Consider an example system 'Virage' to process video this system performs the following steps for processing a video,

1. Periodically collects a video input frame.
2. Compare the frame with the last captured frame, in order to find the difference between the frames. After comparison the frame will be discarded if the difference is less than a threshold, else, it goes to step-3.
3. A vector is defined for a frame to be processed. This vector represents the different features of that frame, where each dimension gives a unique feature level of the frame.

Finally the vector becomes a unit of processing that can be used in the search system. Image processing is same as video processing. Indexing at semantic level needs pattern recognition of objects in the images. Examples are Virage's multimedia datablade, IBM's QBIC and MIT's photobook. In these examples, for an analog audio input, these systems perform a conversion of audio to digital form and finds the phonemes of the utterances. The conversion is useful as the phonemes will be given as input to HMS (Hidden Markov Search) model. The model will then determine the words that were spoken at a confidence level. In HMS model, a phonemes can be broken into four states.

A multimedia item not only stores the extracted index searchable data but also a mechanism to link the modalities in the search process. The two mechanisms that can be used are,

(i) Positional

When the modalities are interspersed in a linear sequential form, the modality mechanism called "positional" is used. Consider an example document with audio or images inserted in it. Such a document can be treated as a linear structure with only Juxtaposition relationship between the modalities. Thus, allowing a query to specify the position of a boat (image) in a paragraph of a book.

(ii) Temporal

If modalities execute concurrently then the mechanism called 'temporal' is used. It is completely based on time. The source for a television is a multimedia source having audio, video and closed captioning. By using SMIL (Synchronized Multimedia Integration Languages) multimedia presentations can be created easily SMIL is a mark-up language that supports audio, video, images and text.

2.1.4 Information Extraction

Q17. How does the process of information extraction differs from the document indexing process?

Answer : Model Paper-I, Q4(b)

The processes of information extraction are,

- Facts determination
- Text extraction.

Facts determination deals with the identification and extraction of important facts in an item. These facts are then sent to the database and reside in the structured fields. In text extraction, the extracted text is used to summarize the data. The summarization includes representation of major concepts in an item in the form of a summary. However, the process of document indexing provides representation of documents for their faster and effective retrieval. Moreover, it associates the information with a file which is utilized by document management system.

Automatic file build is the process of extracting facts. Its objective is to process incoming items and extraction of index terms which will reside in a structured database.

The goal of an information retrieval system is to give a deep representation of all the contents in an item. An information extraction system analyzes only those parts of a document that has information relevant to the extraction criteria. Mostly, the goal of data extraction is to update a structured database so as to contain additional facts. Controlled vocabulary or substrings from an item may request for such updates. Information of a specific type which has to be extracted is called "slot". Organization of slots is in the form of either semantic frames or templates.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Metrics

Information extraction can be compared by using metrics. These metrics are defined by slightly modifying precision and recall.

Recall

It deals with the amount of actual information that is extracted from an item versus the amount of information extraction that is required.

Precision

Precision deals with the amount of information that is accurately extracted versus the total information extracted.

Overgeneration

The amount of unnecessary information that is extracted is given by "overgeneration". It can be a result of filling the slots and topics with irrelevant data and templates respectively.

Fallout

The amount of incorrect assignment of slot fillers made by a system with an increase in incorrect slot fillers is called "fallout".

The above listed measures are applicable to both automated and human extraction processor.

Data Extraction

Perfection is always scarce if extraction is done either by human beings or by automated systems. However, there are sources based on which data extraction can be analyzed. The best of these sources is message understanding conference proceedings, there conferences are sponsored by Intelligent Systems Technology Office of the department of defense and Advance Research Project Software/Agency. If any organization wishes to evaluate its algorithms then they are facilitated with large databases. One such example is data extraction in MUC-5. A detailed extraction of 120 documents was done by four human analysts. Then, the extraction was compared with top three automated information extraction systems, and the following observations were made.

1. 79% of recall and 82% of precision was achieved by human analysts with a processing time of 15 to 60 minutes for an item.
2. 53% of recall and 59% of precision was achieved by the automated systems with a processing time of 30 seconds to three minutes for an item.

Thus human analysts defeated the automated systems in terms of recall and precision, but not in terms of cost (the automated system are cheaper). If both human analysts and information extraction systems are combined to perform data extraction then the benefits of both the portion can be achieved.

Document Summarization

The goal of document summarization is "extraction of item's summary that maintains the major ideas with a significant reduction in size". It does not attempt to find specific facts. Table of contents, titles, abstracts, etc., are the examples of summaries that are most likely to be the part of any item. There are two approaches to document summarization they are (i) human and (ii) automated systems. Different humans may create different summaries for the same document. Similarly, different automated algorithms may generate different abstracts/summaries, but the difference is not big (as those generated by human beings). The reason being, automated algorithms generates summaries by evaluating a score for every sentence and the sentences with the highest scores are then extracted. Such example algorithms are usage of term location, rhetorical relations, syntactic coherence and contextual inference using cue words and statistical weighting properties. Edmundson, Kupiec et al, devised an algorithms, for document summarization. The following five feature sets were used as the basis for their algorithm.

(a) **Sentence Length Feature**

This feature requires that the length of words in a sentence should be more than five.

(b) **Fixed Phrase Feature**

It deals with the existence of the phrase "cues".

(c) **Paragraph Feature**

It focuses on location of sentences and also on first 10 and last 5 paragraphs in an item.

(d) **Thematic Word Feature**

It uses the frequency of a word.

(e) **Upper Case Word Feature**

It focuses on acronyms and proper names.

2.2 DATA STRUCTURE

2.2.1 Introduction to Data Structure

Q18. Give a brief account on the two major data structures used in any information system.

Answer : **Model Paper-II, Q4(b)**

The two major data structures that are used in any information system are,

- The first data structure contains the received items in the normalized form.
- The second data structure contains the processing tokens and the related data structure to support search functions.

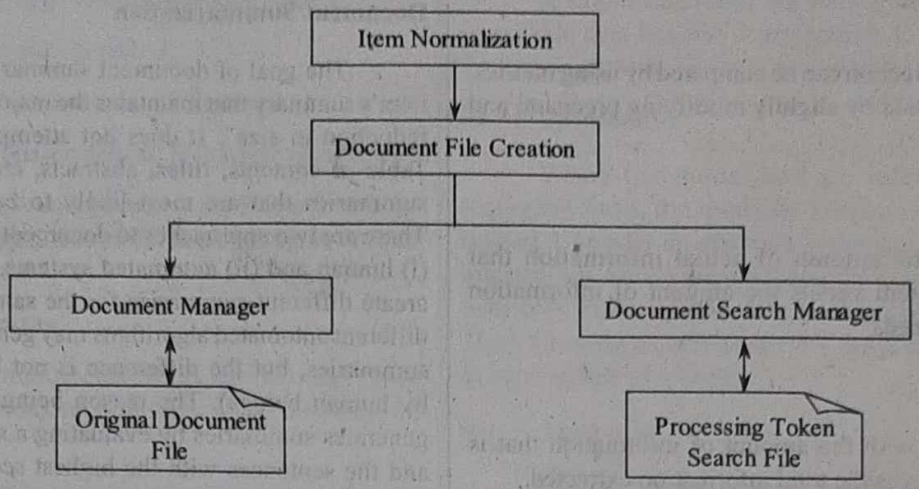


Figure: The Two Major Data Structure

From the figure it is clear that, document file creation is divided into document manager and document search manager to generate the original document file and processing token search files, respectively. The references to items are obtained from the results of search function. These references satisfy the search statement and are passed to the document manager for retrieval purpose.

The first transformation applied to data structure before placing it in the searchable function is referred to as 'stemming'. This transformation converts the representation of a concept or word to a canonical morphological representation. The problem with stemming transformation is that the discrimination concept related information may be lost thereby reducing the precision and performance capability. On the other hand, stemming has the ability to improve recall.

The data structure that is used in both database and information system is called the inverted file system. When multiple search items are placed across the total database, then the inverted file data structure reduces the secondary storage access capability. The inverted data structure is used as the searchable data structure in many commercial and educational systems. The variation of searchable data structure is referred as N-gram data structure, which fragments the processing tokens into small string units. These fragments are used for searching purpose. The efficiency of N-gram is more when compared to full-word inversion data structure. Also, the conceptual manipulations are easily done in the N-grams. With the use of PAT trees and arrays, the text of an item is seen as a single long stream rather than a juxtaposition (combination) of words. Around this paradigm, some search algorithms such as signature files are based on the text string. This algorithm is based on the concept of eliminating the non-relevant items, thereby reducing the number of searchable items to a manageable subset. The subset can be reviewed by the user or may be applied across other search algorithms so that any false hits that occur at the signature filter can be removed.

Typically, a special data structure that is used commonly over the internet is referred to as hypertext. This data structure has an imbedded link within an item so that it can be linked to its related item.

2.2.2 Stemming Algorithms

Q19. What are stemming algorithms? Discuss their uses in information retrieval systems. Also explain the affix removal technique.

Answer :

Model Paper-III, Q4(b)

Stemming Algorithms

The words can be inflected to their respective base, root or stem form or as a written word form, this process is called stemming. It is not necessary to have a stem that is similar to the word's morphological root. But, the related stem should be mapped onto a single stem, and that stem can be an invalid root. Stemming process is generally called as 'conflation' that is mainly used in search engines. They deal with indexing, query expression and other problems arising due to natural language processing. The programs that employ stemming process are stemming programs and can also be termed as stemmers or stemming algorithms.

The following are the two uses of stemming algorithms in information retrieval systems,

1. Stemming algorithms improve the efficiency of an information system.
2. They have the potential to improve recall.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

1. Improving Efficiency of Information System

In this, the premise is that the stem contains the meaning of the concept related to the word and the suffixes (endings) create certain modifications to the concept or used for creating sentences by arranging words and phrases. The use of languages is not only defined by their grammars, but are also developed based upon human usage. Hence, languages consist of exceptions and non-consistent variants, which require exception look-up tables and normal reduction rules.

Multiple representations of a word is equated as a single representation stem. This idea of equating provides significant compression along with the savings that occur in storage and processing. Now consider an example, where the stem "distribut" is associated to "distributable", "distributability", "distribution", "distributerized" one significant compressed word. Since inversion lists use weighted information, the savings is required only in the dictionary. Based upon the stemming data structure implementation, the information is needed in a single inversion list for the stem rather than multiple inversion lists. As the size of inversion lists is the basic factor for storage, the compression of stemming does not minimize the storage requirements.

2. Improving Recall

As there exists a semantically consistent stem for a set of words, the generalization process of stemming will not lose its relevant items. The stemming of words for example, "determine, determines, determination, determinations, determining" to a single stem ("determin") will protect those terms that are entered by the user. Also, stemming finds all the variants that exist in their items. The risk with stemming is that it decreases precision i.e., the value of precision does not depend on identifying all the relevant items but reduces the retrieval of information associated with non-relevant items.

Certain systems like Natural Language Processing (NLP) systems can also face problems through stemming because the retrieval information is lost at the aggregate levels of NLP. The tenses of verbs are required in identifying whether a particular concept contains an indexing in the past or future. But, while creating a stem, the tenses of verbs may get lost. For example, time is a kind of relationship that is defined in NLP systems.

Affix Removal Technique

This technique removes prefixes and suffixes for deriving the final stem. Stemmers are usually iterative and remove the longer prefixes and suffixes from the terms. The Porter algorithm is the basic affix removal technique. But, it minimizes precision, thereby introducing certain anomalies to user query based on the integrity of the system. Stemming process is applied both to the user's query and incoming text. During this process, if the query term moves to a distinct semantic meaning, then the user will not be able to understand the reason of returning a particular item and may start questioning the integrity of the system.

Techniques such as successor stemming and table lookups are the other alternatives of affix removal technique and requires additional overheads. Successor stemmers identifies prefix overlap as an increased length of a stem, which inturn identifies the optimal length for every individual stem from a statistical perspective rather than linguistic perspective. Table lookup needs a very large data structure such as RetrievalWare, which is based on a large thesaurus/concept network. The Kstem algorithm combines stemming rules with dictionary and is used in the INQUERY system to determine processing tokens.

Q20. Explain the Porter stemming algorithm.

Answer : (Model Paper-I, Q5(a) | Nov./Dec.-12(R09), Q2(a))

Porter algorithm depends upon the conditions of stem, prefix and suffix and also upon its related actions.

Following are the examples of stem conditions,

(i) Let 'm' be a function used to measure a set of vowels followed by a consonant (a, e, i, o, u, y). If V is a set of vowels and C is a set of consonants, then the function m is defined as,

$C(VC)^m V$

Where, C(initial) and V(final) are optional and 'm' denotes the number of repetitions of VC.

Measure(m)	Examples
0	shy, tree
1	cold, felt, west
2	enrol, island
3	graceful, isotherm

(ii) *<L>

This symbol represents that the stem should end with letter L (example : trial).

(iii) *V*

This symbol represents that the stem should contain a vowel (example : flag).

(iv) *d

This symbol represents that the word should end with double consonant (example : least).

(v) *O

This symbol represents that the stem should end with consonant - vowel - consonant sequence, but the final consonant should not be *w*, *x* or *y*. (Example : deformed).

The conditions of the suffix are in the following form : current_suffix = = pattern.

Actions take the following form : old_suffix → new_suffix

Rules are divided into a sequence of steps. Some examples of rules are,

Step	Condition	Suffix	Replacement	Examples
1a	NULL	sses	ss	classes → class
1b	*V*	ing	NULL	taking → tak
1b1	NULL	at	ate	animat(ed) → animate
1c	*V*	y	i	courtesy → courtesi
2	m > o	aliti	al	casualiti → casual
3	m > 0	icate	ic	triplicate → triplic
4	m > 1	able	NULL	addressable → address
5a	m > 1	e	NULL	animate → animat
5b	m > 1, *d, *<L>	NULL	single letter	animall → animal

For the word "triplicate", the following sequence of steps are involved in stemming process,

Measure(m)	Examples
4	triplicat
1b1	triplicate
3	triplic

From the above example, it is observed that only one rule from each step must be applied at any time i.e., applying another rule in step 4 is not possible.

Q21. Apply the Porter Stemming Algorithm steps to the following words: Irreplaceable, Informative, Activation and Triplicate.

Answer : Dec.-19(R16), Q5

The steps of porter stemming algorithm are as follows,

Step-1a

If any of the following suffix is seen in the word then it is replaced with the respective new suffix.

1. SSES → SS
2. IES → I
3. SS → SS
4. S → NULL

Step-1b

1. (m > 0) EED → EE
2. (*V*) ING → NULL
3. (*V*)ED → NULL

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Step-1b(i)

In the above step-1b, if the last two rules are successful then the following rules are performed.

1. AT → ATE
2. 1Z → 1ZE
3. BL → BLE

Step-1c

1. (*V*) Y → I

Step-2

If the following condition is satisfied then the old suffix is replaced with new suffix.

1. (m > 0) ATIONAL → ATE
2. (m > 0) TIONAL → TION
3. (m > 0) ENCI → ENCE
4. (m > 0)ANCI → ANCE
5. (m > 0) IZER → IZE
6. (m > 0) ABLI → ABLE
7. (m > 0) ALLI → AL
8. (m > 0) ELI → E
9. (m > 0) ENTLI → ENT
10. (m > 0) OUSLI → OUS
11. (m > 0) ATION → ATE
12. (m > 0) IZATION → IZE
13. (m > 0) ALISM → AL
14. (m > 0) ATOR → ATE
15. (m > 0) FULNESS → FUL
16. (m > 0) OUSNESS → OUS
17. (m > 0) IVENESS → IVE
18. (m > 0) IVITI → IVE
19. (m > 0) BILITI → BLE
20. (m > 0) ALITI → AL

Step-3

If the following condition is satisfied, then the old suffix is replaced with new suffix.

1. (m > 0) ICATE → IC
2. (m > 0) ALIZE → AL
3. (m > 0) ICITI → IC
4. (m > 0) ICAL → IC
5. (m > 0) ATIVE → NULL
6. (m > 0) NESS → NULL
7. (m > 0) FUL → NULL

Step-4

If the following condition is satisfied, then the old suffix is replaced with new suffix.

1. (m > 1) AL → NULL
2. (m > 1) ANCE → NULL
3. (m > 1) ENCE → NULL
4. (m > 1) IC → NULL
5. (m > 1) ER → NULL
6. (m > 1) ANT → NULL
7. (m > 1) ABLE → NULL
8. (m > 1) ANT → NULL
9. (m > 1) ENT → NULL
10. (m > 1) MENT → NULL
11. (m > 1) EMENT → NULL
12. (m > 1) OU → NULL
13. (m > 1) ATE → NULL
14. (m > 1) OUS → NULL
15. (m > 1) ISM → NULL
16. (m > 1) ITI → NULL
17. (m > 1) IZE → NULL
18. (m > 1) IVE → NULL
19. (m > 1 and (*S or *T))ION → NULL

Step-5a

If

1. (m > 1) E → NULL
2. (m = 1 and not * 0) E → NULL

Step-5b

If (m > 1 and *d and *L) → Single letter

The given words are as follows,

(i) Irreplaceable

(ii) Informative

(iii) Activation

(iv) Triplicate

(i) Irreplaceable

For the word "Irreplaceable", the following sequence of steps are involved in stemming process.

(a) Initially, identify the suffix in the given word. In the above word, the suffix is 'able'.

```
I  r  r  e  p  l  a  c  e  a  b  l  e
↓  ↓  ↓  ↓  ↓  ↓  ↓  ↓  ↓
v  c  c  v  c  c  v  c  v
↓  ↘  ↙  ↓  ↘  ↙  ↓  ↓  ↓
v     c     v     c     v  c  v
```

v(cv) (cv) (cv) ⇒ v(cv) (cv) (cv)

[v] (cv) (cv) (cv) [c]

It can be represented in the following form,

$[v]\ (cv)^3\ [c] \Rightarrow m = 3$

(b) The suffix does not match with the suffixes found in step 1, 2 and 3.

(c) Now, go to step 4 for the above word, the value of m is 3. So it satisfies the condition i.e., m > 1 (since m = 3).

(d) As it satisfies the condition, the old suffix is replaced with new suffix (i.e., able → NULL). It is shown as follows,

Irreplaceable → Irreplace

The stem word of Irreplaceable is Irreplace.

(ii) Informative

For the word "Informative", the following sequence of steps are involved in stemming process.

(a) Initially, identify the suffix in the given word. In the above word, the suffix is 'ative'

```
I  n  f  o  r  m  a  t  i  v  e
↓  ↓  ↓  ↓  ↓  ↓
v  c  c  v  c  c
↓  ↘  ↙  ↓  ↘  ↙
v     c     v     c
```

(vc) (vc)

It can be represented in the following form,

$[c]\ (vc)^2\ [v] \Rightarrow m = 2$

(b) The suffix does not match with the suffixes found in step 1, 2.

(c) Now, go to step 3. For the above word, the value of m is 2. So it satisfies the given condition (i.e.,) m > 0 (since m = 2).

(d) As it satisfies the condition, the old suffix is replaced with new suffix (i.e., ative → NULL). It is shown as follows,

Informative → Inform

The stem word of Informative is Inform.

(iii) Activation

For the word "Activation", the following sequence of steps are involved in stemming process.

(a) Firstly, identify the suffix in the given word. In the above word, the suffix is 'ation'.

```
A  c  t  i  v  a  t  i  o  n
↓  ↓  ↓  ↓  ↓
v  c  c  v  c
↓  ↘  ↙  ↓  ↓
v     c     v  c
```

v(cv)c

It can be represented in the following form,

$[c]\ (vc)^1\ [v] \Rightarrow m = 1$

(b) The suffix does not match with the suffixes found in step 1.

(c) Now, go to step 2. For the above word the value of m is 1. So it satisfies the condition i.e., m > 0 (since m = 1).

(d) As it satisfies the condition, the old suffix is replaced with new suffix (i.e., ation → ate). It is shown as follows,

Activation → Activate

The stem word of Activation is Activate.

(iv) Triplicate

For the word "Triplicate", the following sequence of steps are involved in stemming process.

(a) Firstly, identify the suffix in the given word. In the above word, the suffix is 'icate'.

```
T  r  i  p  l  i  c  a  t  e
↓  ↓  ↓  ↓  ↓
c  c  v  c  c
↘  ↙  ↓  ↘  ↙
   c     v     c
```

c(vc)c

It can be represented in the following form,

$[c]\ (vc)^1\ [v] \Rightarrow m = 1$

(b) The suffix does not match with the suffixes found in step 1, 2.

(c) Now, go to step 3. For the above word the value of m is 1. So it satisfies the condition i.e., m > 0 (since m = 1).

(d) As it satisfies the condition, the old suffix is replaced with new suffix (i.e., icate → ic). It is shown as follows,

Triplicate → Triplic

The stem word of Triplicate is Triplic.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Q22. Write short notes on,

(a) Dictionary look-up stemmers

(b) Successor stemmers. **Model Paper-II, Q5(a)**

OR

State and explain dictionary lookup stemmer algorithm. **Oct./Nov.-20(R16), Q4(a)**

(Refer Only Topic: Dictionary Look-up Stemmers)

Answer :

(a) Dictionary Look-up Stemmers

Dictionary look-up stemmers is another approach for determining a stem apart from those approaches that are fully dependent on certain algorithms such as Porter stemming algorithm. However, in this mechanism some stemming rules need to be applied, which are generated from those that have minimum number of exceptions. When the actual terms or stemmed version of it is entered, then a search is made on all the entries available in the dictionary. This is done so as to look-up for the exact match and substitute that term with the best suited stem. This approach is basically applicable in the following systems,

1. Inquery system
2. Retrievalware system.

1. Inquery System

The stemming mechanism employed in this system is "Kstem" which is generally a morphological evaluator. Its purpose is to integrate different terms with same meaning so as to form a root. Consider an example where "moonscape" and "moonshine" reduces to moon. But both these terms have different meaning. Like other stemmers, Kstem is also associated with the Natural Language Processors (NLP) and dictionaries. Kstem does not return truncate words rather it returns original words. The prerequisite for Kstem working is that a word needs to be present in the dictionary before performing substitution. If the word is available in the dictionary, then it is considered that the word is not related to the root word after performing the stemming process. The conflation process is not performed while a word is being processed. It is essential to directly relate a word to the desired root in case of irregular morphologies. The following are the major data files used by Kstem system for controlling and configuring the stemming process.

(i) Dictionary of words called "lexicon".

(ii) Additional list of words for the dictionary.

(iii) Exception list for words that have "e" at the end. For example "rages" to "rage" but "raged" to "rag".

(iv) Direct_conflation : It permits the definition of direct conflation by making pairs of words that can override the stemming algorithm.

(v) Counter_nationality : It deals with the conflations that exist between nationalities and countries. For example, Irish maps to Ireland.

(vi) Proper Nouns : It refers to a list of non-stemming proper nouns.

2. Retrievalware System

This system depends upon the meaningful network support data structure. This structure consists of approximately 400,000 words, which are available in the dictionaries that comprises of morphological word variants. A word which is not in a special form is lookup in the dictionary inorder to find an easier form of that word. This is done by removing the suffixes and spelling it again in plural form as defined in the dictionary.

(b) Successor Stemmers

Successor stemmers depends upon the length of prefixes and upon an analogy in the structural language, which examine words and morpheme boundaries depending upon the distribution of phonemes. The phonemes are a small part of speech that differentiates one word from another. The successor stemmer identifies the successor varieties for a word so that it can be divided into segments and one of the segments of a word is selected as stem.

The successor variety for a word is defined as the number of distinct letter that utilize the segment length plus one character. Consider an example, where the successor variety for a five-letter word has first three letters similar and a distinct fourth letter plus one character for the existing word. Figure below shows successor variety representation as a symbol tree for the terms like cap, card, cheat, cheap, close and clot.

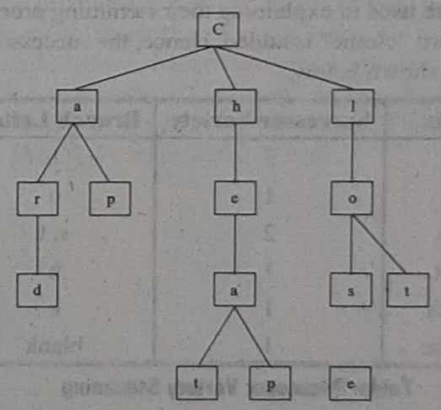


Figure: Symbol Tree Representation for Cap, Card, Cheap, Cheat and Clot

The successor variety for a word is described as the number of children that are related with root i.e., "*c*" in figure. Hence, the first letter "*c*" has three children i.e., successor varieties and the prefix "*ca*", "*ch*", "*cl*" has 2, 1, 1 successor varieties, respectively.

The following methods are used to determine the successor varieties for segmenting a word,

(i) Cut-off Method

The length of the stem is defined by selecting a cut-off value. This value is different for different set of words.

(ii) Peak and Plateau

This method is used to break a segment when the size of successor character is more than the size of the preceding character and the one following it.

(iii) Complete Word Method

This method is used to break a segment on the boundaries of complete words.

(iv) Entropy Method

This segment is used for the distribution of successor varieties of letters. Suppose $|A_{sl}|$ be the number of words starting with the sequence of letters 's' and length 'l' $|A_{sli}|$ be the number of words in A_{sl} that has the successor 'i'. Thus, the probability of A_{sl} is given by,

$$\frac{|A_{sli}|}{|A_{sl}|}$$

The entropy of $|A_{sl}|$ is,

$$H_{sl} = \sum_{x=1}^{26} -\left[\frac{|A_{sli}|}{|A_{sl}|}\right]\left[\log_2\left(\frac{|A_{sli}|}{|A_{sl}|}\right)\right]$$

Using the above formula, a set of entropy units and its associated predecessors are evaluated for a word. When a cut-off value is reached, it is selected and a boundary is identified. After experimenting with the four methods, Hafer and Weiss discovered that the combination of these methods produce good results and are used in explaining their stemming processes. In table, the word "clothe" is added. Hence, the successor variety stemming is shown below,

Prefix	Successor Variety	Branch Letters
c	3	a, h, *l*
cl	1	o
clo	2	s, t
clot	1	h
cloth	1	e
clothe	1	blank

Table: Successor Variety Stemming

Now, apply all the four methods for the term "clothe",

(i) If the cut-off method is applied and a value five was selected, then the stem would be "cloth".

(ii) The peak and plateau method cannot be applied to the term since the successor variety gradually decreases.

(iii) If the complete word method is applied, then the stem is "clot".

(iv) Also, the entropy method cannot be applied since the term does not have enough values.

The basic advantage of peak and plateau and the complete word methods is that they don't require a cutoff value to be selected.

After a segment break is made for a word, the segment that is used as the stem must be selected. The following rule is used by Hafer and Weiss.

If (first segment occurs in <= 12 words in database)

stem = first segment

else

stem = second segment.

The above rule illustrates that if a segment of more than 12 words exist in the text and it is being analyzed, then that text is probably referred to as a prefix. Hafer and Weiss observed that multiple prefixes does not appear in the English language and hence the selection of the first or second segment identifies the proper stem.

Q23. What is the likely effect of stemming and lemmatization on?

(i) Vocabulary size: Increase, decrease, unpredictable

(ii) Precision: Increase, decrease, unpredictable

(iii) Recall: Increase, decrease, unpredictable.

Answer : Dec.-14(R09), Q2(a)

(a) Stemming

For answer refer Unit-II, Q19, Topic: Stemming Algorithms (First Paragraph Only).

(i) Vocabulary

The effect of stemming on vocabulary decreases. For example, laugh, laughing, laughed, laughs can be reduced to laugh.

(ii) Precision

The effect of stemming on precision is unpredictable.

(iii) Recall

The effect of stemming on recall increase.

(b) Lemmatization

Lemmatization refers to process of considering words with respect to their morphological analysis. For this, it need the information regarding dictionaries.

(i) Vocabulary

The effect of lemmatization on vocabulary decreases.

(ii) Precision

The effect of lemmatization on precision is unpredictable.

(iii) Recall

The effect of lemmatization on recall increases.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Q24. How do stopping and stemming reduce the size of an inverted index?

Answer : (Model Paper-III, Q5(a) | Dec.-14(R09), Q2(b))

Stopping Method

Stopping method make use of stop word list which drastically decreases the size of index there by decreasing the overall search time. This list is basically a collection of keywords which are not indexed and are considered as irrelevant to index. Therefore, such list are not considered during search time because of its frequency and semantic use. The advantages of deleting such keywords is that it reduces the complexity of storing and accessing the structure. Stopping method is generally preferred for eliminating such words since it is very tedious to enlist those words individually.

Stemming Method

Stemming method is a simplified form of morphologic analysis that perform word truncation. For example, the word bring, bringing, brought are stem to bring. If more stemming is applied the more irrelevant items are retrieved. However, the issues which arrises because of this truncation include.

1. Confusion may encounter with different words.
2. Not all the users may understand the steamer word.

Q25. What is an Error Rate Relative to Truncation (ERRT)? How can its value be computed?

Answer :

Frakes studied various stemming processes and observed that some authors failed in reporting test statistics and sample sizes. This leads to interpretation problems and the test sample sizes were so small that their results were difficult to be questionable. Based on this, Frakes deduced the following three conclusions,

(i) Stemming affects retrieval (recall) but does not cause any harm. The retrieval effectiveness of different full stemmers are different from the exceptions of the Hafer and Weiss stemmer.

(ii) Stemming has a similar capability as manual conflation.

(iii) Stemming depends upon the nature of the word.

Paice defined a stemming performance measure called Error Rate Relative to Truncation (ERRT) to measure the stemmers impact on recall. This measure is used for comparing stemming algorithms. The approach is based upon the division of terms into "concept groups". This division is done as per the semantic and morphological relationship between them. Though a non-perfect stemmer algorithm is applied to the concept groups, they still have multiple stems rather than a single stem. This results in an error and the state of it is reflected in the Understemming Index (UI). It is possible that a similar stem may be present in multiple groups. This error state is also displayed in the Overstemming Index (OI). The worst case of the stemming algorithm is that the words are stemmed through truncated word lengths and the words less than the desired length are not truncated. Thus, the UI and OI values can be computed. The best case of stemming algorithm is that UI and OI are made equal to zero.

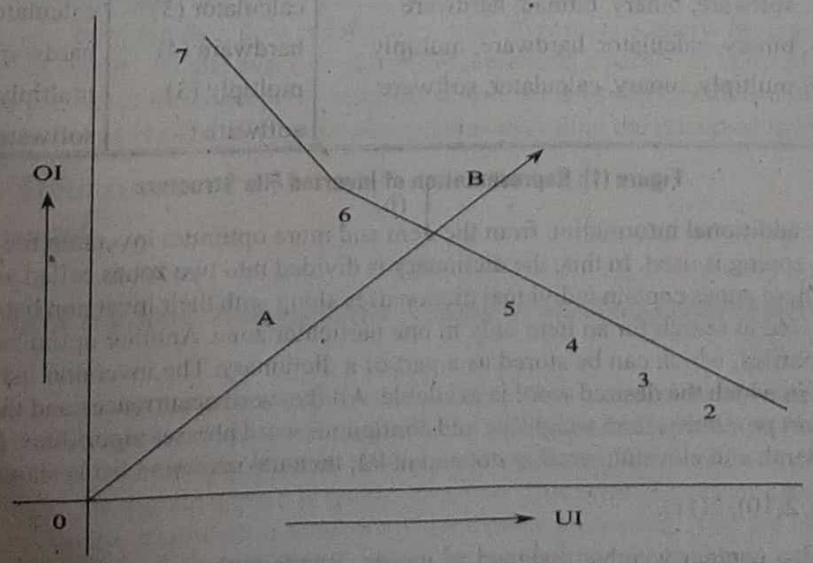


Figure: Computation of ERRT Value

From the above figure, the value of ERRT is computed by calculating the distance from the origin to the stemmer's coordinates (UI, OI) divided by the distance from the origin to its intersection with the truncation line. Stemmers that lie closer to the origin will possess lower UI and OI values that specify a shorter distance value.

The ERRT values are biased with the initial grouping of terms and the larger values of ERRT appear with loose grouping of test terms. For a particular test term, the UI value of the Paice algorithm was less than the Porters algorithm whereas the OI value of the Paice algorithm was greater than the Porters algorithm. As a result, the ERRT value of Porter was greater than Paice. Thus, Paice algorithm is a heavier stemmer than the Porters algorithm. But the comparisons between the stemmers can be neglected. To eliminate impressions and to provide common comparisons between the stemmers, the stemmer evaluation requires additional work by creating a baseline for future developments.

Based upon the comparisons by Frakes and Paice, it is clearly understood that stemming as a generalization of processing tokens has the potential to improve recall. But, after experimentations, stemming is not clearly proved to improve recall rather it helps in reducing precision. The impact of stemming on precision can be reduced by ranking items according to the terms in the query, classification and selection of terms. In case of large databases, stemming should not be seen as a relevant compression technique for storage rather it is used for the reduction of dictionary sizes. Hence, there is a decrease in the processing time for every search term in the query.

2.2.3 Inverted File Structure

Q26. Explain the inverted file structure in detail.

Answer :

Model Paper-I, Q5(b)

Inverted file structure is a common data structure that is used in database management as well as in information retrieval systems. The inverted file structures are divided into the basic files,

- Document file
- Inversion lists or Posting files
- Dictionary.

The inverted file is called so because it has an ability to store an inversion of documents. A unique numerical identifier is given to every document in the system. Hence, the documents are represented as : document #1, document #2, document #3 and so on. For a particular word, dictionary helps in locating the inversion list. The dictionary which is stored in the system is referred to as a sorted list of all distinct words or processing tokens and also as a pointer to the location of inversion list. Dictionaries also help in storing other information such as the length of inversion list. This information is used in query optimization.

(a) Documents	(b) Dictionary	(c) Inversion Lists
Document #1, hardware, software, multiply, bitmap	binary (3)	binary – 3, 4, 5
Document #2, bitmap, software, hardware, calculator	bitmap (3)	binary – 1, 2, 3
Document #3, software, binary, bitmap, hardware	calculator (3)	calculator – 2, 4, 5
Document #4, binary, calculator, hardware, multiply	hardware (4)	hardware – 1, 2, 3, 4
Document #5, multiply, binary, calculator, software	multiply (3)	multiply – 1, 4, 5
	software (4)	software – 1, 2, 3, 5

Figure (1): Representation of Inverted File Structure

To increase precision, additional information from the item and more optimum inversion file structure has been provided. Consider an example, where zoning is used. In this, the dictionary is divided into two zones called as the "Abstract zone" and the "Main Body zone". Each of these zones contain individual dictionaries along with their inversion lists. Due to which, the overhead increases when a user is allowed to search for an item only in one particular zone. Another optimization lists when inversion list consists of only one or two entries, which can be stored as a part of a dictionary. The inversion list contains a unique document identifier for each document in which the desired word is available. All the word occurrences and their positions are stored in the inversion list inorder to support proximity, term weighting and contiguous word phrases algorithms. Hence, if the word "multiply" occurs as the fifth, seventh, tenth and eleventh word in document #2, then the inversion list is shown as,

Multiply – 2(5), 2(7), 2(10), 2(11).

The inversion lists also contain weights assigned to words. Words with unique properties are stored in their relevant dictionaries, so that a fine internal manipulation and representation can be done.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

While performing a search operation, the inversion lists in the query are positioned and are applied with a proper logic for manipulation between these lists. The result obtained is a list of items that satisfy the query. If some systems support ranking functionality, then the inversion lists in them is rearranged according to that ranking order. The unique number of documents helps in retrieving the documents from the document file.

From figure (1), the inversion lists for "multiply" and "software" are (1, 4, 5) and (1, 2, 3, 5) respectively. These two lists are logically added by an AND operation, thereby resulting in the final hit list (1, 5).

Instead of using a dictionary, B-trees are used to locate the inversion list. Inversion lists are placed at the leaf nodes of the B-tree and are referred to by higher level pointers. The following figure shows the representation of a B-tree of order n by using the words that appeared in figure (1).

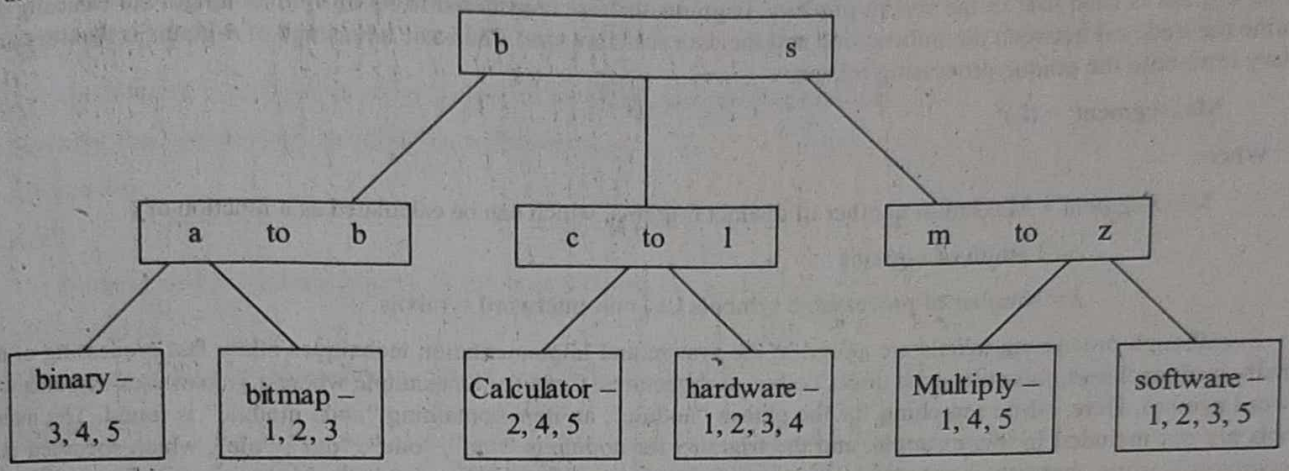


Figure (2): Representation of B-tree Inversion Lists

A B-tree of order n is defined as follows,

(i) A root node should contain 2 and $2n$ keys.

(ii) All the internal nodes should contain n and $2n$ keys.

(iii) The keys are stored in ascending order.

(iv) All leaf nodes must be present at the same level or differ by one level at the most.

B-trees can be used as an efficient inverted file storage structure in which heavy data updates can be done.

Certain items in the information systems are infrequent whenever they are modified after their development. Thus, many commercial system take advantage of this by freezing these kind of items in databases of document files, thereby generating a new structure. These databases are archived and are used for user's query. Only the latest databases are required to be searched, since queries need only the recent information. As older items are deleted and modified, the archived databases are needed to be backed up, thereby saving significant overhead. A new inverted database has a significant overhead because it requires additional words in its dictionary and inversion lists. The inversion file list structure provides an optimal performance in searching large databases. This optimality arises from the data flow reduction in order to resolve a query. Inversion file structures are used for storing concepts and their relationships, so that every inversion list represents a particular concept. Thus, inversion lists contain a list of all the items that are present in the concept. In addition, it also contains the locations and weights of the item so that a finer resolution of concepts can be maintained. This information helps in finding the relationships between the concepts.

2.2.4 N-gram Data Structures

Q27. Describe N-gram model. Oct./Nov.-20(R16), Q3(b)

OR

Explain the first use of n-grams. How n-grams is used in spelling error-detection and correction. Model Paper-II, Q5(b)

Answer :

N-grams are the data structures that are used in information systems. They are considered as a special technique for stemming. They are not concerned about the semantics of the word. Instead, they depend upon the fixed consecutive series of 'n' characters. The data structure is converted into overlapping n-grams so that a searchable database can be created. Some systems allow the usage of interword symbols like #, for n-gram when n is greater than two. This symbol replaces a blank, period, semicolon, colon in the n-gram set. The n-grams that are created are a sequence of searchable processing. Also the similar n-grams can be created a number of times from a particular word. Examples of n-grams for n = 2, 3 and 5 are shown in figure (a) for the word "New world".

For n = 2, Bigrams with no interword symbols	ne ew wo or rl ld
For n = 3, Trigrams with no interword symbols	new wor orl rld
For n = 3, Trigrams with interword symbol #	#ne new ew# #wo wor orl rld ld#
For n = 3, Pentagrams with interword symbol #	#new# #worl world orld#

Figure (a): Bigrams, Trigrams and Pentagrams for the Word "New World"

As shown in figure (a), an n-gram is a fixed length of continuous data. *N*-grams contain the overlapping symbol segments for defining the searchable processing tokens, which have logical linkages to all the items in the query. In some cases, the infrequently occurring *n*-gram is used first in the search process. Trigrams that are considered to be of optimal length are basically used to determine the trade off between the information and the data structure size. The basic advantage of *n*-grams is that they assign a boundary limit onto the unique processing tokens.

$$\text{Maxsegment}_c = (\lambda)^c$$

Where,

MaxSegment = Maximum number of distinct n-grams, which can be calculated as a function of *c*.

c = Length of *n*-grams

λ = Number of processable symbols i.e., non-interword symbols.

Eventhough processing tokens are stored in the system and implementation techniques allow fast processing of tokens on small-sized machines, false hits exist under certain architectures. Consider an example where a system uses trigrams with no interword symbol. Here, while searching for the phrase "nodule", an item containing "node module" is found. The interword symbols are not included in this example, and the trigrams for nodule is "nod", "odu", "dul", "ule", which specified that the consecutively starting character is not same. This type of error occurrence is infrequent in the longer n-grams because there exist a lot of information in the word fragment. The disadvantage with the longer n-gram is that, full word data structures are provided frequently because many words exist within a single n-gram. Another disadvantage of n-grams is large-sized inversion lists, which increase the number of processing tokens. The average length of a word in English is between six and seven characters. The use of trigrams increase the size of processing tokens by a factor of five as shown in figure (a), even if there are no interword symbols.

As the processing tokens are associated with a finite limit boundary, the mapping of an item to an n-gram data structure and in query processing can be achieved by optimized implementation techniques. *N*-grams do not deal with the representation of concepts and their relationships and have no semantic meaning and syntactic information because it only deals with the searchable processing tokens. But, the combination of all terms of *n*-grams can be used for word indexing, recall and precision with an improvement in its performance. Vector representations of *n*-grams are used in determining the closeness of all items in query.

Uses

The following are the two uses of *N*-grams,

1. Role of N-grams in Conflation

N-grams (bigrams) are used as a special technique for conflating terms. Since it does not follow the definition of stemming it does create only the word fragment rather than semantic meaning of words. *N*-gram data structures maps longer words into shorter fragments.

2. Detection and Checking of Errors

Another use of n-grams is that it detects and checks the errors in spelling a word. The probability of occurrence of n-grams is determined in the English vocabulary and if it occurs infrequently, then n-grams is considered as a potential erroneous word. According to Damerau, the following are the four categories of spelling errors for a word phrase "memory".

S.No.	Categories of Error	Example
1.	Insertion of a single character	memmory
2.	Deletion of a single character	meory
3.	Substitution of a single character	memary
4.	Transposition of two adjacent characters	meomry

The above classification scheme provides a way for identifying the errors in a word. Thus, potential input errors are detected for correction by the information retrieval systems. In these systems, trigrams are used for manipulating the index length and for text compression.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Q28. Use the 2-gram index and 3-gram index for processing the following wildcard queries Tol* and Rea*. Is "Tool" result for the wildcard query Tol*. If the answers yes, solve this problem.

Dec.-14(R09), Q2(c)

Answer :

Given that, wildcard query is,

To1* and Rea*.

(i) **To1***

The 2-gram index for wildcard query $ To1* is as follows,

$t, TO and OL

It can be inferred from the above 2-gram index that TOOL is a result of wild card To1*.

Similarly, the 3-gram index for $To1* is as follows,

$To and To1

(ii) **Rea***

The 2-gram index for wildcard query $Rea* is as follows,

$R, Re and ea.

The 3-gram index for wildcard query $Rea* is as follows,

$Re and Rea.

2.2.5 PAT Data Structure

Q29. What is a PAT data structure? Discuss with an example.

Nov./Dec.-12(R09), Q2(b)

Answer :

The use of *n*-grams with interword symbols is similar to the indexing of continuous "*n*" character tokens. The input data stream can be addressed by a technique called PAT tree data structure. The name PAT comes from PATRICIA (Practical Algorithm to Retrieve Information Coded in Alphanumerics) trees, which are used to search text, images and applications in genetic databases. The continuous input data is transformed into a searchable data structure that consists of substrings.

PAT trees are created from the given input string, where each position in it points to a substring which consists of all the input text. All substrings in PAT trees are unique i.e., the location of each character and the length of substrings is different. This is used in parallel processor and hardware text search machines. A substring can begin at any point in the text and the length is measured from its starting location upto the end of the text. Possibly, the length of a substring can be more than the length of the input string. This is done by adding null characters to the substring. Thus, these substrings are referred to as sistrings or semi-infinite strings. Figure (a) describe some sistrings for an input text,

Text : *X and Y are two different characters*

Sistring 1	:	X and Y are two different characters
Sistring 3	:	and Y are two different characters
Sistring 12	:	two different characters
Sistring 24	:	nt characters
Sistring 34	:	ers.

Figure (a): Some Examples of Sistrings

A PAT tree is considered as an unbalanced binary digital tree, which is described by the sistrings. Each bit of the sistring decide the branching patterns in such a way that the left branch contain 0's and the right branch contain 1's. Each node in the tree is used to specify which bit determines the branching or the skip from the root node. This is useful where the levels of tree do not require branching.

A PAT tree of input size '*m*' consists of "*m*" leaf nodes and atmost "*m* – 1" internal nodes. The leaf nodes in the PAT tree contains the key values stored in it. Additional constraints can also be added at the leaf nodes. The sistrings that occur immediately after an interword symbol can be limited. Suppose, if the binary representation of "*d*" is 100, "*a*" is 001 "*t*" is 111 and "*a*" is 001, then the complete word "data" produces the input 100001111001. Figure (b) shows an example in generating a PAT tree,

INPUT	100001111001
sistring 1	100001111001
sistring 2	00001111001
sistring 3	0001111001
sistring 4	001111001
sistring 5	01111001
sistring 6	1111001
sistring 7	111001
sistring 8	11001

Figure (b): Sistrings for Input 100001111001

Using the sistrings from figure (b), the full PAT binary tree is generated, which is shown in the figure (c) below,

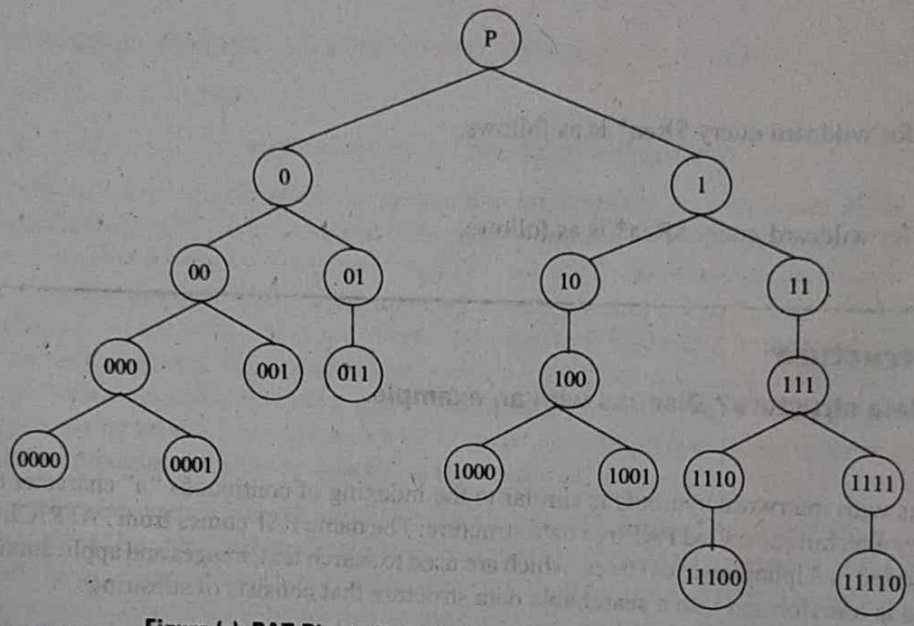


Figure (c): PAT Binary Tree for Input Text "100001111001"

Since any of the substrings can start with 0 or 1, the first level of the tree has nodes 0 and 1. "0" followed by 0 or 1 forms a substring of the input, so they become the child nodes of "0". "01" nodes followed by 1 forms a substring but "01" followed by 0 is not a substring of the given input. Thus, 011 becomes the only child of "01" node. Similarly, all the possible substrings are obtained. As the input string length is 12, its PAT tree can have only 11 (i.e., 12 – 1) high-level (internal) nodes.

As figure (c) is a complex PAT tree, the internal nodes of it can be skipped. Thus, a compact tree can be generated which is shown in the figure (d),

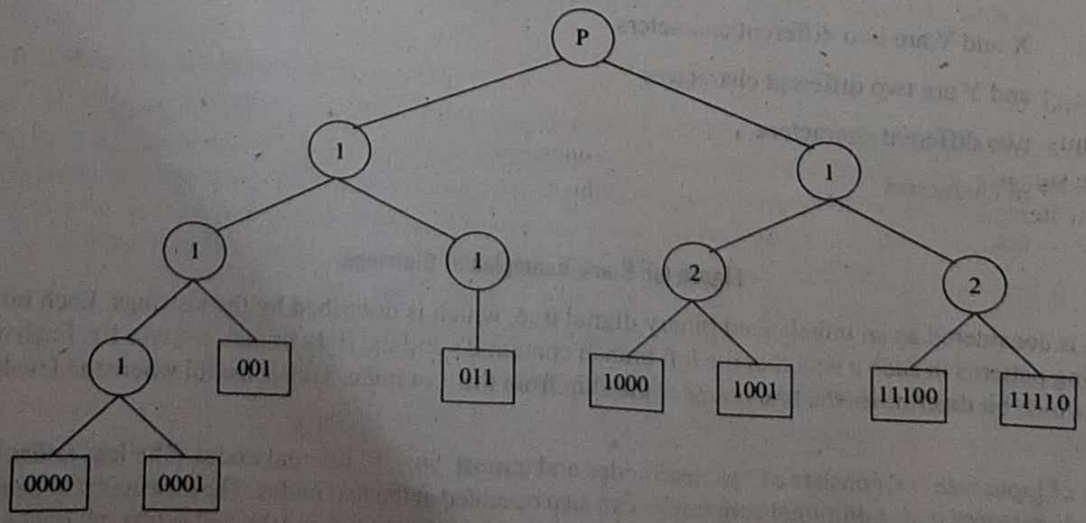


Figure (d): A Compact PAT Tree of Figure (c)

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Figure (d) describes that all the intermediate nodes are skipped until leaf nodes are found. Thus, PAT trees are generated for prefix searches because each subtree contains a prefix node, which defines all the sistrings to satisfy the prefix searching criteria.

PAT trees are more accurate and efficient than the other searchable data structures like inversion lists and signature files. If searching a string is the basic objective of PAT trees, then they provide an alternative structure, which stores the input text so that the string manipulation is supported. This alternative structure does not have the capability of storing concepts and their relationship that are related to a particular item.

Q30. Describe cataloging and indexing. Discuss about automatic indexing and PHT data structure.

Answer : Nov.-13(R09), Q2

Cataloging and Indexing

For answer refer Unit-II, Q12.

Automatic Indexing

For answer refer Unit-II, Q15.

PAT Data Structure

For answer refer Unit-II, Q29.

Q31. Compare the N-gram data structure and PAT data structure.

Answer : May/June-13(R09), Q2(a)

N-gram Data Structure

For answer refer Unit-II, Q27, (Exclude Topic: Users).

PAT Data Structure

For answer refer Unit-II, Q29.

2.2.6 Signature File Structure

Q32. What is signature file structure and explain how it is useful in IRS.

Answer : (Model Paper-III, Q5(b) | Oct./Nov.-20(R16), Q4(b))

Signature File Structure

The main objective of a signature file structure is to perform a fast test so that the unnecessary items that are not related to a query are removed. The items that fulfill the test criteria can be determined by applying some other search algorithm for removing false hits. The text that designates the items is represented in a compressed form so that the fast test can be easily performed. Since signature file structure is compressed and unordered, it requires very less space when compared to inverted file structure and new items that are essential can be added at the end of the structure rather than updating the significant inversion list. As items are deleted very rarely from information databases, it is typical to mark those items as deleted. Signature file search is a linear search of the compressed items, which produces a response time arranged in accordance with the file size.

Surrogate Signature File Structure

To create the surrogate signature search file, a superimposed coding technique is needed. The coding basically depends upon the words that maps into a "word signature". Typically, a word signature refers to a fixed length code in which a fixed number of bits are set to "1". A hash function is used to determine the bits that are set to one. Then the ORed operation is performed on the existing word signatures to create an item signature. The signatures with maximum number of 1's must be avoided and an item with multiple words is divided into blocks. The figure below shows the block size of five words, the code length of 16 bits and the bits that are set to "1" are five,

Text : *Welcome to Sia Educational Group*

Word	Signature			
Welcome	0110	0110	0000	0001
to	0000	0000	1110	1001
Sia	0101	0010	0100	1000
Educational	0111	0100	1000	0000
Group	0110	0100	0110	0000
Block Signature	0111	0110	1110	1001

Figure: Superimposed Coding

The words in the text are mapped into their respective signatures. The signature search is completed by matching the bits that are specified by the words in the text.

The signature file structure consists of a signature block and each row of this file is associated with a pointer that points to the original text block. The design goal of a signature file structure is to trade off the size of data structure rather than the density of final signature code. The probability of occurrence of collision in hashing two unique words to the same value is reduced by longer code lengths. Each code with only some bits decrease the effect of a code word pattern which is present in the final block signature rather than the item. Consider an example, where the word "find" has a signature as : 0100 0110 0000 1001, which do not match the block signature in figure. Due to which, false hits occur. If the final data structure is compressed, then the number of bits per word is one. Thus, a binary coded vector method is adopted, where each position of the bit designates the presence of a word in the item. This method requires a longer code length and ensures that there exist no false hits until hashing between the words is not done to the same value.

The search time of the signature file is O(N) and to minimize this search time, the signature matrix is divided horizontally. One of the traditional methods of signature file search is to hash the block signature to a particular slot. Suppose a query contain words which are less than the words present in a block, then the block maps to the possible number of slots instead of mapping to only a particular slot. As the number of slots decrease exponentially, the number of words in the item increases. Another approach of the signature file is to map the block signatures into an index sequential file where the initial '*m*' bits are identified as the index to the signature block. Other techniques include two level signatures and B-tree structures.

Another implementation of signature file structure is to divide and search the signature matrix vertically (column wise) by neglecting all those columns that are not identified by the hash function. This type of division is known as vertical partitioning. An inverted file structure is used to store the signature file matrix. A significant overhead arises when additional 1's are added at the end of each inverted column, which designates a signature in the new item.

Signature files provide a solution to store and locate the information in various situations. They have been applied as databases with minimum frequency, medium size databases, WORM devices, distributed and parallel processing machines.

2.2.7 Hypertext and XML Data Structures

Q33. Write short notes on hypertext data structures.

Answer : May/June-13(R09), Q8(c)

Hypertext Data Structure

Basically, Internet is a global information network, which introduced a new storage data structure called hypertext to display information on a WEB page. This structure is used largely on the Internet and is different from the other traditional storage data structures. Certain languages such as Hypertext Mark-up Language (HTML) and Extensible Markup Language (XML) are used to store hypertext. The hypertext uses an imbedded pointer to refer one item to the other. Every individual item is referred to as **node** and the reference pointer to that node is called a **link**. All the referenced items or the original items may be similar or have different data types.

For example, HTML specified the internal structure of the item. Typically, a document is associated with the text and HTML tags. These tags define how to display the document. Tags are represented by less than and greater than symbols (examples : <head>, <body>, <P>, <h$_1$>, etc.). The HTML tag consisting of hypertext linkages is represented as : <a href = ...#NAME/ a> where <a> and </a> are the start and end of the anchor tags to denote the text so that the user can activate. "href" refers to the hypertext reference, which contains either the name of the address (Uniform Resource Locator – URL) of a file. "#NAME" specifies a destination point where the item is to be reached. The URL consists of the following three components,

(a) The access method that retrieves the item by the client.

(b) The Internet address for designating where the item is stored.

(c) The address of the item.

The URL for the HTML specification is defined as : http://incl.chicago.ua/hypertext/WWW/MarkUp/HTML.html. Here, "HTTP" is the Hypertext Transfer Protocol used for retrieving the item in HTML. "incl.chicago.ua" is the destination point in which "incl" is the name of the machine at CHICAGO with "ua" being United States of America, and "/hypertext/ WWW/MarkUp/HTML.html" specifies where the file HTML. html is found. The following example shows a part of the HTML document.

Example of a Part of HTML

```
<CENTER>
    <H1>
        Using Hyperlink Access keys
    </H1>
        You might be interested in these two browsers
```

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

```
<p>
    <A HREF = "http://www.microsoft.com/
    ie" ACCESSKEY = "I">
        Internet Explorer
</A>
<BR>
    <A HREF = "http:www.netscape.com"
    ACCESSKEY = "N">
        Netscape Navigator
        </A>
</CENTER>
```

From the above example, it is clearly described that an item can have multiple hypertext linkages through which sequential reading of the item can be continued by skipping over the linkages. This is similar to the user's decision where he can decide whether to continue reading or skip upon reaching a footnote. Sometimes, hypertext is also known as a "generalized footnote".

Usually, the hypertext references are also used to include other information such as graphics, audio, video etc. If the hypertext is a part of the item, then the referenced file is frequently located at the same physical location. If some other users create certain items, then they are usually located at the other physical sites. If some items are deleted or moved, then there is no way of updating other items that refer to them. The major problem in hypertext linkage is linkage integrity.

Dynamic HTML exists with the Netscape Navigator 4.0 and Internet Explorer 4.0. It is a collection of HTML tags, style sheets and programming so that a more animated WEB page can be created, which provides quick response to the user's request. The features that support dynamic HTML are an object-oriented views of WEB page, cascading style sheets, programming etc. The Document Object Model (DOM) defines the object-oriented views. This DOM is called as the dynamic HTML object model and HTML object model by Microsoft and Netscape browsers, respectively. Consider an example, where a page consists of heading that can be named and the text style and color attributes can be applied to that name by including a script on the page. A style sheet specifies the default style features of a document. Dynamic HTML uses style sheets in a cascading way. Due to which, new changes can be made to the appearance of the existing page. An alternative way of style sheets is "layering", which is used to display the content of a page in a different way i.e., in the form of sections. The existing HTML capabilities are equipped with additional features so that new data structures can be addressed. Netscape browser also allows the use of dynamic fonts in the Web page. Due to this, the font choice is independent of what the browser provides. Dynamic HTML is defined differently by both Microsoft and Netscape. Thus, it differs in definition and functioning.

Q34. Give brief introduction on XML and features of XML.

Answer :

XML

Extensible Markup Language (XML) is a simple markup language that describes the structure of a document. It has been derived from the Standard Generalized Markup Language (SGML) for large-scale electronic publishing. It is also used in data exchange on web.

Unlike HTML and other presentation languages, XML does not display data but carry and store data. Thus, XML is easy to implement.

XML defines a set of rules with the help of which other markup languages can be created. Thus, it may be referred as "Meta-markup". Once a language is created using XML, it can be manipulated by any number of applications that are designed for a specific set of requirements.

XML is a W3C recommendation. The W3C is responsible for releasing the XML versions. The latest version of XML is XML 2.0.

It is defined by DTD (Data Type Description) and is not constrained to 50 attributes and 70 defined tags in a single DTD for HTML. Any tag can be created by the user for describing and manipulating their structure. A suite of XML tags is redeveloped as HTML by W3C. An example of XML tagging is as follows,

```
<Company> Sia publishers </Company>
<City> Hyderabad </City>
<State> Telangana </State>
<Product> Books </Product>
```

An RDF i.e., Resource description format is being developed by W3C for specifying the properties of web resources like documents and images. It offers the Platform for Internet Content Selection (PICS) to combine the labels of filtering material.

X-Link i.e., XML linking and X-point i.e., XML pointer language specifications are used to define the Hypertext links for XML. Using this, distinction for various links to the locations is allowed within the document and out of the document. This also allows application to know if the link is a link to other document which is an extension of another document or is just a reference of an item. This helps to determine the requirements that are retrieved inorder to define the total items needed to be indexed.

XML also includes XML style sheet linking definition for defining, displaying and cascading style sheets. The designers are allowed to limit what is displayed to the user and also allows the expansion to the entire item if required.

Features of XML

1. XML files are text files. They can be managed by all text editors.
2. It is extensible. This specifies only the structural rules of tags, allowing the users to create their own tags or to use the existing tags.

3. It is a platform independent language.
4. It is very simple as that of HTML. It has very few syntax rules.
5. It is extremely suitable for document storage and processing, both online and offline.
6. It supports unicode. This allows communication of any information written in any human language.
7. It is a subset of SGML and is used to structure data as well as documents.
8. It can represent common data structures such as records, lists and trees.
9. It has strict syntax and parsing requirements. This allows the parsing algorithms to be extremely simple, consistent and efficient.
10. It provides a communication gateway between applications on same as well as different systems.

2.2.8 Hidden Markov Models

Q35. Explain in detail about Hidden Markov Models (HMM).

Answer :

Hidden markov model is applied in speech recognition and gesture recognition technique. It defines multiple parameters whose values are used for identifying efficient training patterns for known category. Later, the model having highest posterior probability is used to categorized the test pattern. It is been applied since 20 years for solving problems in speech recognition for information retrieval search and leads to good results. The first practical and comprehensive descriptions of HMM was written by Dr. Lawrence Rabiner. The process of HMM can be understood better by defining a discrete markov process.

Consider an example of a markov model of stock market which consists of three states. The following will be one of the states which is observed at closing of the market.

s1 : Market reduced

s2 : No change in market

s3 : Increase in market value.

A state transition matrix defines the movement among the states with state transitions as follows,

$$A = \{a_{i,j}\} = \begin{matrix} .5 & .3 & .4 \\ .1 & .6 & .3 \\ .6 & .7 & .5 \end{matrix}$$

Given that the market fell on state-1, then the suggests that the probability of no change in the market is .1. It then allows questions like the probability of increase in market value in next 4 days. It is equivalent to the sequence = {s3, s3, s3, s3, s1}. To simplify the model, assume that the current state depends on the last state i.e,, (discrete, first order, markov chain). It is then calculated by the following formula.

$$\begin{aligned} P(SEQ) &= P[s3, s3, s3, s1] \\ &= P[s3] * P[s3/s3] * P[s3/s3] * P[s1/s3] \\ &= s3(init) * a_{3,3} * a_{3,3} * a_{3,3} * a_{1,3} \\ &= (1.0) * (.5) * (.5) * (.4) \\ &= .05 \end{aligned}$$

Where SEQ = sequence

In the above equation, the probability of initial state of s3 is given as s3(init) = 1.

Consider the following graph,

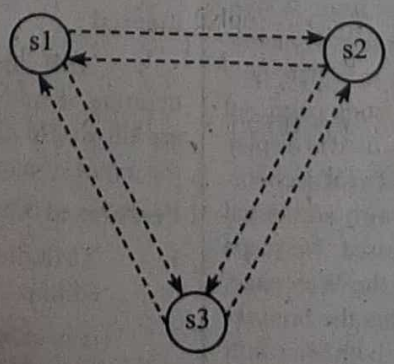


Figure: Graph Model

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

In the above graph, the directed lines represent the probabilities of state transitions $a_{i,j}$ there exists an implicit loop from each state and every state is an observable event. This model is usually applied to solve the problems on speech recognition, but it is too restrictive to be applied. Thus, a probability function is allowed to be associated to increase the flexibility of the state. Hence, it is called hidden markov model. The model is named from the fact that there are exists two stochastic processes with the underlying stochastic process being hidden, it can be determined by the observations that are originated from other stochastic process. Hence, the system has a series of results as input. But the system do not have an idea about the markov model and its states that are used for generating results. Thus, the process of HMM analyzes which model explains the observed results in the best possible way.

A formal definition of a discrete hidden markov model is defined by Schauble and Mittendory as follows,

(i) $s = \{s_0, \ldots\ldots s_{n-1}\}$ denotes set of finite states where s_0 denotes the initial state and n denotes the number of states. All the states are interconnected in such a way that any state can reach any state.

(ii) $v = \{v_0, \ldots\ldots v_{m-1}\}$ denotes set of output symbols. It represents the physical output which is being modeled on the system.

(iii) $A = s \times s$ is a probability transition matrix wher $a_{i,j}$ denotes the probability of transitioning from state i to j in such a way that,

$$\sum_{j=0}^{n-1} a_{i,j} = 1 \ \forall\ i = 0 \ldots\ldots n-1.$$

In this matrix, each value is positive and lies between 0 and 1 and each state which is reached from other state will have the value of non-zero.

(iv) $B = s \times v$ denotes probability output matrix where $b_{j,k}$ denotes the function defining the probability and

$$\sum_{k=0}^{m-1} b_{i,k} = 1 \ \forall\ j = 0 \ldots\ldots n-1.$$

(v) The initial state distribution.

At each state transition, the hidden markov model generates output symbol. The probability of next state given current state is the transition probability. And, the probability generated by the output upon arriving at next state is called output probability.

The above quoted HMM definition is used as a generator of possible sequence of probabilities and their outputs or a proper HMM can be modelled from given output sequence. The total HMM specification requires the following.

(i) Three probability measures for state transitions.

(ii) Output symbols

(iii) Specification of states

(iv) Initial states and

(v) Output probability functions.

The distributions are denotes as A, B and π and are defined as,

$$\lambda = (A, B, \pi)$$

The main issue that is associated with HMM is how to calculate the probability of observed sequence of outputs in the given HMM model. Another approach is to analyze which number of models are required to be chosen when given a set of observed outputs. It effects uncovering the hidden part of the model. Applying the optimality criteria for selecting the states is another typical approach. But there exists various algorithms for selection. If a model is selected, then the user expect corresponds to the output and analyzing the state of sequences explains the best possible output. The tuning of λ model for maximizing the probability of the output sequence is the final issue. It is known at training sequence and is essential to permit the models to adapt a problem for being solved.

UNIT 3

Automatic Indexing, Document and Term Clustering

Syllabus

Automatic Indexing: Classes of Automatic Indexing, Statistical Indexing, Natural Language, Concept Indexing, Hypertext Linkages.

Document and Term Clustering: Introduction to Clustering, Thesaurus Generation, Item Clustering, Hierarchy of Clusters.

Learning Objectives

- Different Classes of Automatic Indexing
- Various Methods of Statistical Indexing
- Concept of Natural Language Processing
- Representation of Concept Indexing and Hypertext Linkages
- Different Types of Clustering i.e., Document, Term and Item Clustering
- Thesaurus Generation
- Grouping Hierarchy of Clusters.

Introduction

Automatic indexing can be defined as an indexing method wherein, the information is obtained from the document by evaluating it. This extracted information is stored in a index. Thus, automatic indexing is a method used for representing a document based on index. The index terms includes single word, lengthy phases and there combinations. An indexing technique that determines the significance of a document using a number is called statistical indexing. This number is obtained based on the number of times, an event has occurred in a particular document.

Clustering means 'Grouping' of similar objects that belong to the same subject. Its goal is to assist the user in easily finding out the location where the information is stored. This gave rise to indexing schemes for organizing objects in libraries and standards of electronic indexes used. The generation of thesaurus, i.e., 'Treasure' in Latin, was originated by clustering different words. A thesaurus stores the antonyms and synonyms for different words as well as dictionary stores their definitions within it. Clustering is actually used to group all the identical objects into a 'class' that holds a more general title. The terms 'Cluster' and 'Class' have the same meaning.

PART-A SHORT QUESTIONS WITH SOLUTIONS

Q1. What is automatic indexing?

Answer : Model Paper-III, Q1(e)

Automatic indexing can be defined as an indexing method wherein, the information is obtained from the document by evaluating it. This extracted information is stored in a index.

It is one of the indexing methods used for the representation of a document based on the index terms. Index terms include any of the following,

(a) Single words

(b) Lengthy phrases

(c) Combination of single words and lengthy phrases.

Q2. List the steps of data flow in information process.

Answer :

The various steps of data flow in information process are,

Step 1: Standardization of document input.

Step 2: Zoning of a document.

Step 3: Identification of index terms.

Step 4: Implementation of STOP algorithms.

Step 5: Characterisation of index terms.

Step 6: Implementation of stemming algorithm.

Q3. Write short notes on statistical indexing.

Answer : Model Paper-I, Q1(e)

An indexing technique that determines the significance of a document using a number is called statistical indexing. This number is obtained based on the number of times, an event has occurred within a particular document. Previously, statistical indexing was implemented by the systems in a different way. The significance of a document was calculated by performing a boolean operation on it.

There are two methods of statistical indexing,

1. Probabilistic Statistical Indexing (PSI)
2. Vector Statistical Indexing (VSI).

Q4. What is,

(i) Word Frequency (WF)

(ii) Total Frequency (TF)

(iii) Document Frequency (DF).

Answer :

(i) Word Frequency (WF)

This frequency can be defined as how many times a particular index term (word) has occurred in a document.

(ii) Total Frequency (TF)

It indicates how many times a specific word has occurred in a database.

(iii) Document Frequency (DF)

This denotes an index term which is held by a distinctive number of documents in the database. It means an index term is held by a unique number of documents. Document frequency can also be known as Item Frequency (IF).

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Q5. Define inverse document frequency.

Model Paper-I, Q1(f)

Answer :

Inverse Document Frequency (IDF) is a method that focuses on determining the frequency of particular word in database. This policy is implemented by a weighing algorithm known as Inverse document frequency algorithm. This algorithm suggest that, if weight of term is more then the rate of occurrence should be less and vice versa.

Q6. In the vector space retrieval model, if a query has just one term, which term weighting heuristic will be ineffective. Use no more than two sentences to explain why?

Dec.-14(R09), Q3(b)

Answer :

In vector space retrieval model, if a query has just single term Inverse Document Frequency (IDF) will be ineffective. This is because all the document will have similar IDF. In IDF, weight of a particular word in a document is calculated depending on how many other documents of database contain that word.

Q7. What is clustering?

Model Paper-II, Q1(e)

Answer :

Clustering means 'Grouping' of similar objects that belong to the same subject. Its goal is to assist the user in easily finding out the location where the information is stored. This gave rise to indexing schemes for organizing objects in libraries and standards of electronic indexes used. The generation of thesaurus, i.e., 'Treasure' in Latin, was originated by clustering different words. A thesaurus stores the antonyms and synonyms for different words as well as dictionary stores their definitions within it. Clustering is actually used to group all the identical objects into a 'class' that holds a more general title. The terms 'Cluster' and 'Class' have the same meaning.

Q8. Prove that a term could not be found in multiple clusters when using the single link technique.

Answer :

Algorithm for single link clustering is given below,

1. Select term T_x that does not belong to the existing class and place it in a new class.
2. All the other terms that are associated with the term T_x are placed in the same class.
3. Whenever a term is placed in a class, perform step 2.
4. When step 2 does not result in the determination of a new term, go to step 1.

In the first step of this algorithm, only those terms are selected and placed in a new class, which do not belong to any other class. Thus, if a term is present in a class, it can't be selected again for placing it in another class. A class contain terms that are similar to any other term in the same class.

Thus, multiple clusters cannot have the same term within them, when clustering is performed using single-link technique.

Q9. Write short notes on,

(i) KWOC

(ii) KWIC

(iii) KWAC.

Model Paper-III, Q1(f)

Answer :

(i) KWOC

KWOC is referred as Key Word Out of Context and is also called as concordance.

(ii) KWIC

KWIC is referred as Key Word In Context. It selects a relevant term from its phrase context and displays it.

(iii) KWAC

KWAC is referred as Key Word And Context. It displays the keywords along with their contexts.

Q10. What is manual clustering?

Answer :

(Model Paper-II, Q1(f) | Dec.-19(R16), Q1(e))

Manual clustering enables a user to generate a thesaurus by carrying out the following steps,

1. A domain for clustering is determined as it reduces ambiguities that may occur because of the homographs and will thus help in focusing the creator.
2. From the existing thesaurus, a starting point for a set of potential words can be generated that can be included in the new thesaurus. For this, the concordances from items are selected. Here, a concordance corresponds to an alphabetical listing of words that is taken from a set of items including the frequency of occurrence of those words along with the references of items that holds these words.
3. Manual thesaurus construction mainly depends on the selection of a group of words that are to be included in the new thesaurus. Thus, those words that are not related to the domain of thesaurus and those words whose frequency of occurrence is very high should not be included in the new thesaurus. Also, the words that doesn't contain any information value must be discarded.
4. After selecting the terms, clustering of those terms is carried out, depending on the guidelines of the word relationship and the interpretation of the relationship strength.
5. Thesaurus thus obtained, will be treated with several quality assurance reviews. This is done by other editors before finalizing the thesaurus.

Q11. Discuss briefly about hypertext linkages in IRS.

Answer :

Dec.-19(R16), Q1(f)

One of the most commonly used methods for retrieval of information is hypertext with linkages. Consider an example wherein you want to know about frequently used mail sites. Hence, when a user opens any search engine and type the search statement. Based on the statement, various sites will be displayed. These are nothing but hypertext or hyperlinked documents. Once a user click on any one of these links, you will enter the home page of that site. For example, www.yahoo.com. Suppose, a user want to know how to create a mail account in yahoo, there will be a sublink regarding this. When it is clicked, a registration form will be displayed. If a user want to enquire about games in yahoo, a separate sublink will be there. Hypertext linkages not only help you to find relevant information about a particular document, but also additional information about it. The above process is the combination of two techniques called information retrieval through manual indexing and information retrieval through intelligent agents or web crawlers. The drawbacks of these techniques are that they are exclusively used for searching process but not for indexing purposes.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

PART-B ESSAY QUESTIONS WITH SOLUTIONS

3.1 AUTOMATIC INDEXING

3.1.1 Classes of Automatic Indexing

Q12. Define automatic indexing. Explain with the help of a diagram how data flow occurs in an information processing system.

OR

Discuss the different classes of automatic indexing.

(Model Paper-I, Q6 | Nov./Dec.-12(R09), Q3)

Answer :

Automatic Indexing

Automatic indexing can be defined as an indexing method wherein, the information is obtained from the document by evaluating it. This extracted information is stored in a index.

It is one of the indexing methods used for the representation of a document based on the index terms. Index terms include any of the following,

(a) Single words

(b) Lengthy phrases

(c) Combination of single words and lengthy phrases.

The index terms are provided by a vocabulary or an index language. The speciality of automatic indexing is that, it enables the user to decide the appropriate index terms to be used for document representation based on an algorithm.

The algorithm follows three principles,

1. Categorization of index terms into two parts,
   (i) Relational and grammatical index terms (words)
   (ii) Content-bearing index terms (words).
2. Usage of content-bearing index terms since they occur more frequently in a document.
3. Differentiating the document on the basis of its occurrence throughout the document collection.

Data Flow in Information Processing System

This process includes different steps.

Step 1 : Standardization of Document Input

It involves,

(a) Finding, the format or language in which, the document has been written and

(b) Converting this format into a format supported by the system.

A system that uses only unicode format must convert the language (used for writing a document) into its equivalent unicoded form.

The advantage of translating the languages into standard format is the ease of displaying all the languages onto a single browser and the use of single search system to perform the search process.

Step 2 : Zoning of a Document

Once the document is standardised, it is zoned i.e., it is the sub-division of a document into different zones or fields. Through this, a user can get the information he wants regarding the document.

For example, consider a book. Generally, the book contains,

(i) Title of the book

(ii) Author of the book

(iii) Synopsis.

These are nothing but zones of the book. Hence, if the user wants to know any information about a particular book, he can type any one of the above zones or two zones at the same time. For instance, title and author of the book.

Step 3 : Identification of Index Terms

In this step, index terms that were implemented by the user to search particular information are identified. Index terms as mentioned earlier are words or phrases. Words are determined based on the following classes of symbols,

(a) **Valid Word Symbols:** Valid word symbols comprise of alphabetical and alphanumeric characters.

Example: It a user want to know about Java Server Pages, you will type "JSP".

(b) **Interword Symbols:** Inter-word symbols include special symbols like underscore, semicolons, periods etc.

Example: It a user need information regarding OOPs concept in Java so you may type "OOPs, Java".

(c) **Special Processing Symbols:** Special processing symbols includes symbols that require additional processing such as hyphen.

Example: If a user want to link two independent words such as 'Large-business men' then hyphen must be placed properly to avoid any interpretation as 'large males that run business'.

Step 4 : Implementation of "STOP" Algorithms

Once the index terms have been identified, special type of algorithms are used by the system known as STOP list/algorithm. These algorithms are mainly used to remove any irrelevant index terms. They follow Ziph's rank-frequency law. It states that,

$$\boxed{F*R=C}$$

Where,

F = Number of times the index term has occurred (frequency)

R = In what order, the index term has occurred (rank)

C = Constant.

Some of the examples of the above algorithms are as follows.

Examples

1. Stop index term that is alphanumeric
2. Stop index term that is case sensitive.

Step 5 : Characterisation of Index Terms

Characterisation is the process used by systems to avoid any confusion that results from a specific index term. For instance, usage of upper and lower case characters while searching.

Examples

1. Suppose a user want to know about famous biker rally in Texas, he must type ROT (Republic of Texas), the system will display the desired information. However, the word ROT has different meanings like,

   (i) It is a name of a river in Germany.

   (ii) ROT is also a kind of a vector operation in maths.

   Other examples of characterisation include displaying the information of a particular index term through parts of speech.

2. Suppose the user has typed pat. The meaning of this index term will be displayed on the basis of parts of a speech like,

   (a) In terms of noun and a verb, pat is defined as touch or a clap.

   (b) In terms of adjective, it means simplistic, unconvincing etc.

Step 6 : Implementation of Stemming Algorithms

This step is performed after the characterisation of index terms. Stemming is the process used for the quick retrieval of any index term by eliminating its suffixes or prefixes. This reduced word is called "stem". E.g.: for the words like connection, connecting, the stem is connect.

The speciality of stemming is that the original meaning of the particular word will not change with the use of a stem.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Consider a few examples,

Suppose a user requires information about computers, so he types comput. Not only this, the other words are also displayed like computing, computational which more or less are related to computers.

Popularly used stemming algorithms are,

(a) Porter stemming algorithm

(b) Dictionary look up algorithm

(c) Lovins stemming algorithm.

Step 7 : Generation of Data Structure or Database

The data structure is the storage of information of different index terms. This database is internally created with the help of STOP and stemming algorithms.

Consider a practical example, in which a user need information regarding an ATM. He/she must open google and type ATM and click search, they then see the hyperlinks of ATM. This is what the user observes. The internal process of this is as follows,

(i) Once the user has typed a word to be searched or made a query, the query will be transferred to the database containing the information of ATM.

(ii) This database will create the hyperlinks based on the query, which will be displayed to the user.

The diagrammatic representation of the above steps is as follows,

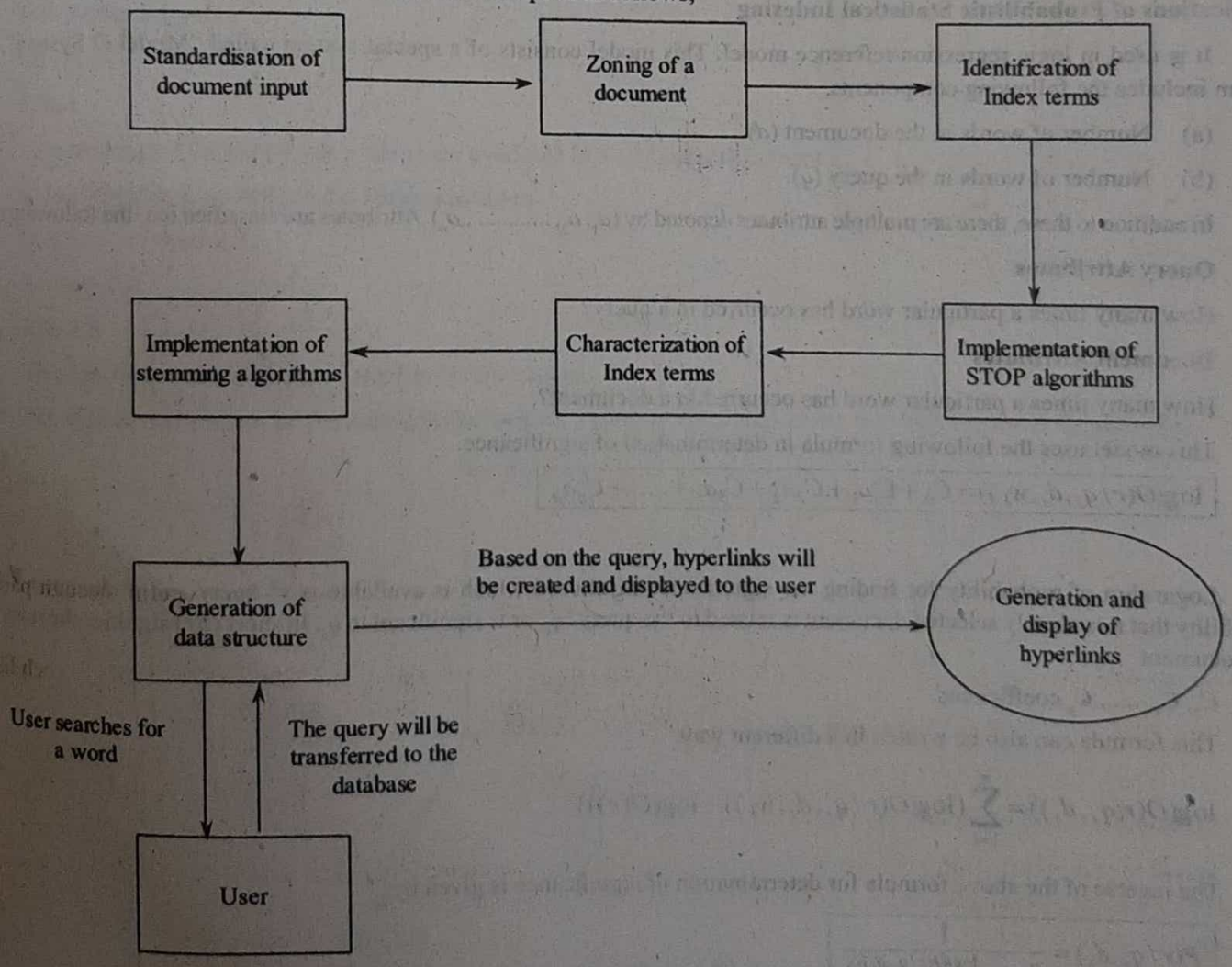


Figure: Data Flow in Information Processing System

3.1.2 Statistical Indexing

Q13. Discuss statistical type of indexing method.

Answer :

Statistical Indexing

An indexing technique that determines the significance of a document using a number is called statistical indexing. This number is obtained based on the number of times, an event has occurred with in a particular document. Previously, statistical indexing was implemented by the systems in a different way. The significance of a document was calculated by performing a boolean operation on it.

There are two methods of statistical indexing,

1. Probabilistic Statistical Indexing (PSI)
2. Vector Statistical Indexing (VSI).

1. Probabilistic Statistical Indexing (PSI)

This method is also known as probabilistic weighting approach. In this method, significance of a document is estimated on the basis of probability value. In other words, documents are ranked based on the probability value.

PSI follows the law of probabilistic ranking. It states that, "the document is displayed to the user by the system based on its significance". The significance is determined through various probabilities. Suppose, a user has made a query regarding information of a particular index term in a document to the system through a search engine. For instance, if the user has typed "What is Java", this query is then transferred to the database containing the desired information. Since, there will be many documents containing java information. The database internally decides which documents are more specific to the user's query and ranks them according to their significance and will therefore be displayed by the system to the user.

Applications of Probabilistic Statistical Indexing

It is used in logic regression/reference model. This model consists of a special system called "Model O System". This system includes the following components,

(a) Number of words in the document (d)

(b) Number of words in the query (q).

In addition to these, there are multiple attributes denoted by ($a_1, a_2, \ldots\ldots\ldots\ldots, a_n$). Attributes are classified into the following types,

(i) Query Attributes

How many times a particular word has occurred in a query?

(ii) Document Attributes

How many times a particular word has occurred in a document?

This model uses the following formula in determination of significance.

$$\boxed{\log(O(r/q_s, d_t, w_k)) = C_0 + C_1a_1 + C_2a_2 + C_3a_3 + \ldots\ldots + C_na_n}$$

log O

Logarithm of probability for finding the significance of a word which is available is s^{th} query and t^{th} document $O(r)$. Probability that a randomly selected document is related to the query 'q_s' or is significant to q_s. In short $O(r)$ signifies the relevance of a document

$C_0, C_1, \ldots\ldots, C_n$ coefficients.

This formula can also be written in a different way,

$$\log(O(r/q_s, d_t)) = \sum_{k=1}^{q} (\log(O(r/q_s, d_t, w_k)) - \log(O(r)))$$

The inverse of the above formula for determination of significance is given by,

$$\boxed{P(r/q_s, d_t) = \frac{1}{(1 + e^{-\log(O(r/q_s, d_t))})}}$$

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

2. **Vector Statistical Indexing (VSI)**

In this method, significance of a document is estimated using a vector. This approach was implemented by SMART system. The vector is nothing but a collection of one-dimensional index term values. VSI is also called as vector weighting method.

Vector weighting method defines two types of techniques for representation of index term values in a document.

(i) Binary technique

(ii) Weighted technique.

(i) **Binary Technique**

In this technique, index term values are represented in the form of ones and zeros.

'1' indicates the availability of a particular index term in the document.

'0' means the non-availability of a particular index term in the document.

(ii) **Weighted Technique**

Here, index term values are represented as real numbers or weights like 3.0, 3.2, 2.0, 0.4, 4.5 etc. These real numbers are used to demonstrate rank of an index term in a document.

An example to demonstrate binary and weighted representation of index term in a document. Consider the following topics,

(a) Java server pages

(b) Servlets

(c) EJB.

These topics can be represented as index terms with the above two techniques.

1. **Using Binary Technique for Representation**

Java Server Pages-1

Servlets-1

EJB-1

This means that the above index terms are available in a particular document.

2. **Using Weighted Technique for Representation**

Java Server Pages-2.0

Servlets-1.0

EJB-3.0

This means that the index term servlets is ranked first.

The above example can be illustrated in the form of a three-dimensional vector which is shown in figure below.

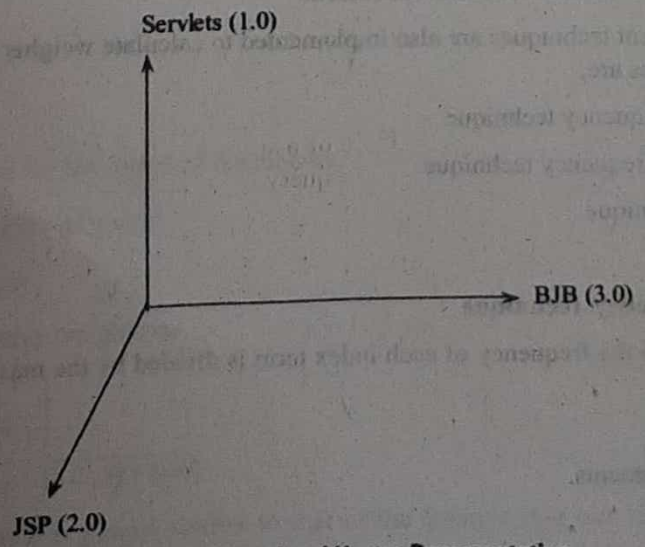


Figure: Three-dimensional Vector Representation

Q14. Outline the simple term frequency algorithm.

Answer : Oct./Nov.-20(R16), Q6(a)

Simple Term Frequency Algorithm

This is one of the algorithms employed by vector statistical indexing for estimation of weights in a particular document. The following information is required to compute weights,

1. **Word Frequency (WF)**

This frequency can be defined as how many times a particular index term (word) has occurred in a document.

2. **Total Frequency (TF)**

It indicates how many times a specific word has occurred in a database.

3. **Document Frequency (DF)**

This denotes an index term which is held by a distinctive number of documents in the database. It means an index term is held by a unique number of documents. Document frequency can also be known as Item Frequency (IF).

This algorithm specifies that the weight of a particular word must be equal to its frequency (rate of occurrence). For example, in a document, a word "microinstruction has been used 25 times", therefore, weight of microinstruction should be 25.

An interesting point to be noted is that the occurrence of a word depends on the length of a document. If the document is short, the probability of occurrence of an index term in a document is less. On the other hand, if the document is lengthy, chances of a particular word occurring in the document is high.

The objective of the algorithm is to reduce the problems among the different documents and words in a database that occurs through normalization. Normalization is the process that describes how a document is arranged in a systematic manner.

In involves,

(a) Standardization of document

(b) Identification of index terms

(c) Characterization of index terms.

The formula derived for computation of weights was,

$$\frac{\left[\dfrac{(1+\log(WF))}{(1+\log(\text{avg}(WF))}\right]}{[(1-sl)*pt+sl*(\text{Number of distinct words})]}$$

Where,

sl – Slope

pt – Pivot.

Both slope and pivot are constant values of a document.

In addition to this, different techniques are also implemented to calculate weights or to determine word frequency. Some of the commonly used techniques are,

(i) Maximum word frequency technique

(ii) Logarithmic word frequency technique

(iii) Normalization technique

(iv) Pivot technique.

(i) Maximum Word Frequency Technique

It is a technique in which the frequency of each index term is divided by the maximum rate of occurrence of that index term in a particular document.

Advantage

Feasible for lengthy documents.

Disadvantage

Not suitable for short documents because the value obtained is too small and insignificant.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

(ii) **Logarithmic Word Frequency Technique**

In this technique, two values are used,

(a) Log value of the frequency of the word

(b) Constant value of the word frequency.

The purpose of log frequency value is to normalize the documents especially when frequency of different words keep on changing because of the document size.

This value also uses a special function called COSINE function. This function is obtained by the following two steps,

1. Representation of document index as a vector.
2. Division of words with the length of the document index (vector).

The result obtained through division gives a vector length of value one.

(iii) **Normalization Technique**

This method is used to normalize the document into equal length. To do this, a correction factor is used which maps COSINE function with another function called normalized function. This function can be explained with the help of a following formula,

$$\boxed{(sl)*(On)+K=Nn}$$

Where,

sl – Slope

On – Old normalized process

Nn – New normalized process.

The value of K is obtained by using two other values,

(a) Pivot value of a new normalized process

(b) Pivot value of an old normalized process.

Note

Both these values should be equal.

Substituting the value of 'K' in the above formula leads to another function called pivoted function which is given by the formula,

$$\boxed{(sl)*(On)+(1.0-sl)*(pt)=p.f}$$

Where,

$p.f$ – Pivoted function.

(iv) **Pivot Technique**

This technique is especially used for two types of documents,

(a) Documents with more number of words

(b) Documents of shorter length.

The formula used here to determine weights is,

$$\frac{\left[\dfrac{(1+\log(WF))}{(1+\log(\text{avg}(WF))}\right]}{[(sl)*\text{Number of distinct terms}+(1-sl)*(pt)]}$$

If observed carefully, this formula is almost similar to that of the formula that was implemented in early days. Only difference is that some values are interchanged.

Q15. Explain the term 'inverse document frequency' with an example.

Answer :

Inverse document frequency is a method that focuses on determining the frequency of a particular word in the database. By using the concept of indexing. It is easy to differentiate the meaning of a particular word in a document from other documents in the database. For instance, meaning of a word "system" will be different in one document and different for other documents. In other words, the information of that word will be useful for one document, but not for other documents. For example, there are 10 documents in a database. System information will be relevant for one document and irrelevant for the remaining documents. From the above statement, it can be concluded that the rate of occurrence of a particular word in the document should not be directly proportional to the weight of that word. This policy is implemented by a weighting algorithm known as IDF (Inverse Document Frequency) algorithm. It means, if the weight is more, the rate of occurrence should be less and vice-versa. This algorithm also uses the concept vector statistical indexing method. In IDF algorithm, weight of a particular word in a document is calculated depending on how many other documents of the database contain that word.

The formula employed by this algorithm for calculating weights is,

$$\boxed{W_{ab} = WF_{ab} * [(\log_2(N) - \log_2(DF_b + 1)]}$$

(or)

$$\boxed{WF_{ab} * [(\log_2(N) - \log_2(DF_b + 1))] = W_{ab}}$$

Where,

W_{ab} – Weight of word "*b*" in document "*a*".

WF_{ab} – Rate of occurrence of word "*b*" in document "*a*" (word frequency).

DF_b – Denotes how many documents contain word "*b*" (document frequency).

N – Number of documents in the database.

The main working of this formula is based on the log computation.

An example to demonstrate this algorithm is,

(i) Consider a word "architecture" is available in 256 documents.

(ii) Consider a word "organisation" is available in 2048 documents.

(iii) Consider a word "computer" is found in 32 documents.

(iv) Consider a word "network" is available in 128 documents.

(v) Consider a word "device" exists in 64 documents.

Calculate the weight of each word if there are 4096 documents in the database and word frequency for each word is 8, 16, 20, 2 and 10 respectively.

Solution for above question the answer is,

$N = 4096$

D.F. for architecture = 256

D.F. for organisation = 2048

D.F. for computer = 32

D.F. for network = 128

D.F. for device = 64

W.F. for architecture = 8

W.F. for organisation = 16

W.F. for computer = 20

W.F. for network = 2

W.F. for device = 10

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Applying the above formula, we get,

$$W_{architecture} = 8 * (\log_2(4096) - \log_2(256) + 1)$$
$$= 8 * (12 - 8 + 1)$$
$$= 40$$

$$W_{organisation} = 16 * (\log_2(4096) - \log_2(2048) + 1)$$
$$= 16 * (12 - 11 + 1)$$
$$= 32$$

$$W_{computer} = 20 * (\log_2(4096) - \log_2(32) + 1)$$
$$= 20 * (12 - 5 + 1)$$
$$= 160$$

$$W_{network} = 2 * (\log_2(4096) - \log_2(128) + 1)$$
$$= 2 * (12 - 7 + 1)$$
$$= 12$$

$$W_{device} = 10 * (\log_2(4096) - \log_2(64) + 1)$$
$$= 10 * (12 - 6 + 1)$$
$$= 70$$

Q16. Write the importance of signal weight.

Answer : **(Model Paper-II, Q6(a) | Oct./Nov.-20(R16), Q6(b))**

Signal Weighting

The drawback of inverse document frequency algorithm is that it does not specify which documents are relevant for a particular word. In short, rate of occurrence of a particular word is not properly distributed among multiple documents. As a result, documents are not properly ranked.

Consider an example where there are 5 documents - *D*1, *D*2, *D*3, *D*4, *D*5 and two words *W*1 and *W*2. Let the documents be denoted as *M*, *N*, *O*, *P*, *Q* respectively and *W*1 and *W*2 as "assembly" and "language" which occurs frequently among these documents.

This example can be illustrated with the help of a tabular representation.

M (D1)	Assembly (W1)	12 times (rate of occurrence)	Language (W2)	10 times (rate of occurrence)
N (D2)	Assembly (W1)	12 times (rate of occurrence)	Language (W2)	15 times (rate of occurrence)
O (D3)	Assembly (W1)	12 times (rate of occurrence)	Language (W2)	20 times (rate of occurrence)
P (D4)	Assembly (W1)	12 times (rate of occurrence)	Language (W2)	8 times (rate of occurrence)
Q (D5)	Assembly (W1)	12 times (rate of occurrence)	Language (W2)	7 times (rate of occurrence)
		60		60

From the above table, it can be seen that the rate of occurrence of *W*1 for each *D*1, *D*2, *D*3, *D*4, *D*5 is 12 which gives a total of 60. Similarly, rate of occurrence of *W*2 for each *D*1, *D*2, *D*3, *D*4, *D*5, is 10, 15, 20, 8, 7 respectively that also results in total of 60. This means that both 'assembly' and 'language' have occurred for a total of 60 times in five documents. Here, it is difficult to identify which document is significant for a word 'assembly'. This process of determining document's significance is called 'precision'. A special algorithm is used to determine precision called Shannon's information theory algorithm. In this algorithm, the probability of occurrence of a particular document should not be directly proportional to the information available in a word or an object or a word in an event (document) that occurs less frequently and has more information than the word that occurs more frequently in a document.

The value of information content of a word in a document can be calculated by using the following formula,

$$info = -\log_2(d_w)$$

Where,

d_w – Probability of occurrence of a word '*w*' in a particular document '*d*'.

Consider an example,

Find the information content value of a word in a document if its probability of occurrence is 20%.

Solution for above question is,

$$d_w = 20\% = 20/100 = 0.20$$

$$\text{Information} = -\log_2(d_w)$$
$$= -\log_2(0.20)$$
$$= -(-2)$$
$$= 2$$

The above formula is applicable only for one document. If there is more than one document, then the formula is,

$$\text{Avg_information} = -\sum_{l=1}^{n} d_{wl} \log_2(d_{wl})$$

Note

'l' can be denoted as a document.

This formula helps in identification of those words whose rate of occurrence is maximum in a document. Here d_{wl} is the ratio of how many times a word has occurred in a document to the total number of document occurrences in the database. Symbolically, this can be represented as,

$$d_{wl} = \frac{Wf_{jl}}{tf_l}$$

The above avg_information formula finds only the information content value. In order to make this formula applicable for computation of weights, a special weighting algorithm called signal weighting algorithm is used. The formula used by this algorithm to calculate weights is given by,

$$\boxed{sig_l = \log_2(tf) - \text{Avg_information}}$$

Where,

sig_l is the weighting factor.

or

$$\boxed{W_{jl} = Wf_{jl} * sig_l}$$

Where,

W_{jl} – Weight, W_f – Word frequency.

or

$$\boxed{W_{jl} = Wf_{jl} * [\log_2(tf_l) - \sum_{j=1}^{n} \frac{Wf_{jl}}{tf_l} \log_2\left(\frac{Wf_l}{tf_l}\right)]} \qquad \left[\because d_{wl} = \frac{Wf_{jl}}{tf_l}\right]$$

Using the same example of "Assembly" and "Language" as above, the signal (weighting factor) for "Assembly" and "Language" is computed by,

$$sig_{\text{assembly}} = \log_2(60) - \left[5 * \frac{12}{60} . \log_2\left(\frac{12}{60}\right)\right]$$

$$= 6 - \left[5 * \left\{\frac{1}{5} . \log_2\left(\frac{1}{5}\right)\right\}\right]$$

$$= 6 - [5 * \{0.2 \times -2\}]$$

$$= 6 - [5 * -0.4]$$

$$= 6 - [-2]$$

$$= 6 + 2$$

$$= 8$$

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

$$sig_{language} = \log_2(60) - \left[\frac{10}{60}\log_2\left(\frac{10}{60}\right) + \frac{15}{60}\log_2\left(\frac{15}{60}\right)\right] + \frac{20}{60}\log_2\left(\frac{20}{60}\right) + \frac{8}{60}\log_2\left(\frac{8}{60}\right) + \frac{7}{60}\log_2\left(\frac{7}{60}\right)$$

$$= 6 - \left[\frac{10}{60}\times -2 + \frac{15}{60}\times -2 + \frac{20}{60}\times -1 + \frac{8}{60}\times -3 + \frac{7}{60}\times -3\right]$$

$$= 6 - \left[\frac{-20-30-20-24-21}{60}\right]$$

$$= 6 - \left[\frac{-115}{60}\right]$$

$$= 6 + \frac{115}{60} = 7.9$$

$$= 8 \text{ (approx).}$$

Q17. Explain about discrimination value.

Answer :

Discrimination Value

This algorithm is used to determine which documents are significant based on the discriminated value of a particular word in a document. To do this, the database must be able to differentiate different documents because if all documents are similar to each other, then it is difficult to identify the relevant documents.

The formula used by this algorithm is as follows,

$$\boxed{\text{disc}_j = avg_simi_j - avg_simi}$$

Where,

avg_simi – Denotes the average similarity among the documents in the database.

avg_simi_j – Indicates the average similarity among those documents in the database from where the word "*j*" is deleted.

$\text{disc}_j \rightarrow$ Discriminated value of a word "*j*"

The discriminated value can be either of the following,

(a) Positive

(b) Negative.

If the value is positive, it means that more documents are similar to each other as the word *j* is removed from all the existing documents.

If the value is negative, lesser documents are similar to each other with the exclusion of the word "*j*".

In order to calculate weights, the above formula can be modified as,

$$\boxed{W_{jl} = Wf_{jl} * \text{disc}_l}$$

Q18. What problems are encountered in using weighting schemes and the vector model?

Answer :

Limitations or Drawbacks of Weighting Algorithms

Most of the time weighting algorithms, especially inverse document frequency and signal algorithms, calculate weights on the basis of values of the two factors,

(i) Document frequency

(ii) Total frequency.

The problem with these factor values is that, they are obtained only when the rate of occurrence of words distribution in the database.

Since, the databases are normally not static in nature. This means that there is,

1. Updation of old documents and
2. Insertion of new documents.

Therefore, it is obvious that frequency distribution of different words will keep changing. In other words, distribution is nonuniform. This in turn leads to a change in the above two factor values used by IDF and signal algorithms.

To overcome this problem, three methods are used.

First Method

Calculation of weights without thinking about the differences in the values of the above two factors. In short, expenses required to maintain a dynamically changing database can be reduced. This is because weights of a particular word are calculated by using those factor values that vary from time to time according to the frequent updations done to the database. As a result, database can be rebuilt or recomputed regularly through its frequent updations.

Second Method

Usage of a constant aggregate value to handle the changes that occur in the values of a document and total frequency. As long as the changes do not affect the final weight value, the constant value can be used. If it is found that changes are occurring frequently and affecting the final weight value, then the fixed aggregate value is not feasible. To handle such changes, a new updated aggregate value should be implemented.

Third Method

Calculation of only those weights of a particular word that can,

(a) Rank the documents properly

(b) Retrieve those documents that are specific to the query of these two methods, this method is more reliable.

Drawbacks of Vector Model

The problems with vector model are almost similar to the limitations of weighting algorithms. For instance, dynamic nature of the information database.

An interesting point to be noted is that each and every word in a particular document can also be treated as a separate topic. Consider a document '*d*' in which there are two words "computer" and "instructions". Here, "computer" can be viewed as a separate topic and "instructions" can be viewed as another one.

Vector model is feasible if there is only one topic in a specific document. This means vector model cannot be implemented when document has more than one topic. This can be explained with the help of an example.

Example

Document	Words
D1	Elephants, Location, Thailand, Tigers, reside, India

From the above document, it can be seen that there are more than one topic like "Elephants", "Location", "Thailand" "Tigers", "reside", "India". By using vector model, it will be difficult to make out which word can be associated with another word.

To avoid this confusion, a special mechanism called 'precoordination' is used. In this method, a word can be linked with another word with the help of keywords like "Modifier", "Action", "Subject" etc.

By applying precoordination, we get,

Subject : Elephants
Action : Located
Modifier : in Thailand
Subject : Tigers
Action : resides
Modifier : in India

Unfortunately, the above mechanism cannot be used in vector model.

Other than this, in vector model, 'proximity searching' is not possible. Proximity searching means determining what should be the appropriate distance range from one word to another word for information searching. The lesser the range the easier will be the search for information i.e., the search process is easier if the range is small.

The syntax of proximity is,

Word1 within "*M*" "Units" of word2

Where,

M – Denotes the range of distance between the two words.

Units – Indicates where the range should be applied.

Example

User wants to search "computer" within zero (*M*) paragraph (units) of "design".

The system will display those documents that will be having 'computer' and 'design' in the same paragraph.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Q19. Describe Bayesian model.

Answer : **(Model Paper-III, Q6(a) | Oct./Nov.-20(R16), Q5(b))**

Bayesian Approach

The problems of vector model can be reduced by implementing a model known as Bayesian model. This model follows conditional probability principle. Consider there are two documents d_1 and d_2 and two events e_1 and e_2, where e_1 is the event occurrence in d_1 and e_2 is the event occurrence in d_2. The above principle states that occurrence of e_1 is possible only if e_2 has already occurred. Symbolically, it can be represented as,

if e_2 then $p(e_1)$.

This law is mainly used for two purposes.

(i) Determination of document relevance through searching process.

(ii) Determination of weights of a particular word.

To implement the searching process, a special formula known as a Bayesian formula is used which is given by,

$P(R/D, Q)$

Where, D– Document

Q – Query

R – Relevance.

To determine weights of a specific word, a network called Bayesian network is used. This network consists of two layers,

(a) Topic(t) layer

(b) Word(w) layer.

There are a maximum of 'm' topics and each topic has a maximum of 'n' words.

The diagrammatic representation of the Bayesian network is given below.

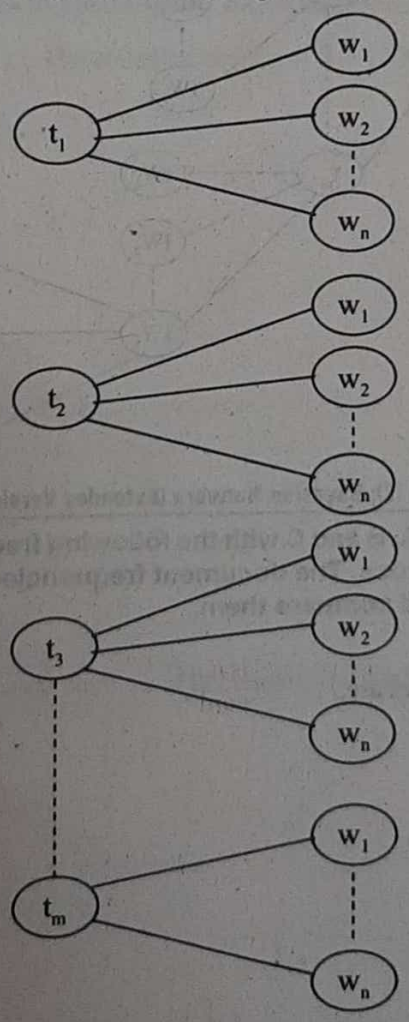


Figure (1): Bayesian Network

Based on the above network, a new theorem was derived called "Binary Independence Theorem". This theorem states that the topics or the words in a document are not dependent on each other. In short, one topic is different from another topic. Unfortunately, this theorem is not perfect. There are some cases wherein a document consists of two words that are related to each other. For example, consider a document which consist of words "computer" and "science". It can be said that this particular document deals with contributions of computer in the field of science. There is a possibility that these two words are not related to each other.

To overcome such problem, two different methods are used,

1. To assume that words or topics are always dependent on each other and also the relevance of a document can be determined easily despite the errors that might have occurred with the use of binary independence theorem.
2. Expand the Bayesian network. This means you must add few more layers in addition to the given layers,
   (a) IS (Independent Subjects (topic)) layer
   (b) IW (Independent Words) layer.

Both the above layers are used to reduce the discrepancies that might have been raised due to the dependency among the words or topics.

The diagrammatic representation of the extended version of Bayesian network is given in figure (2).

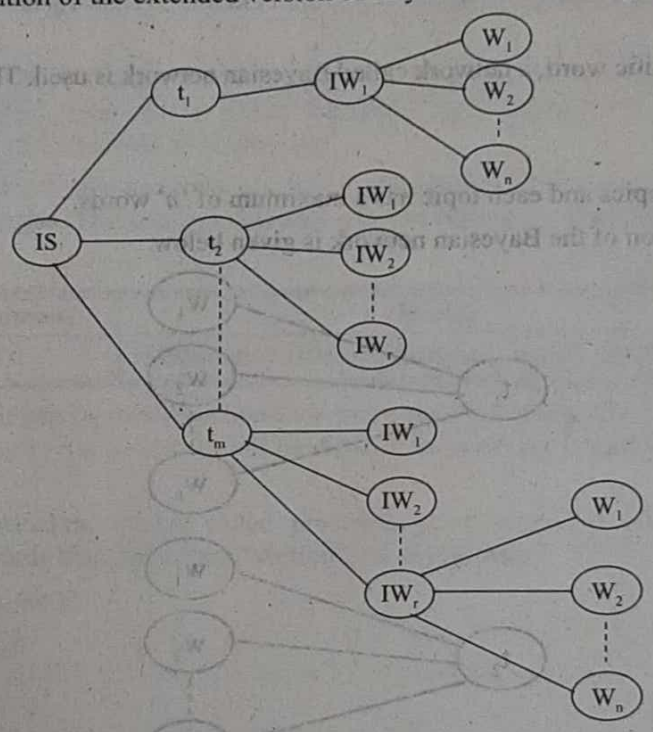


Figure (2): Bayesian Network (Extended Version)

Q20. Given a document with the terms, A, B and C with the following frequencies A:3, B: 2, C: 1. The document belongs to a collection of 10,000 docs. The document frequencies are: A: 50, B: 1300, C: 250. Compute the normalized tf and the tf-idf and compare them.

Answer : Dec.-14(R09), Q3(a)

Given, terms along with there frequencies are,

A : 3

B : 2

C : 1

Number of document = 10,000

Frequencies of document are,

A : 50

B : 1300

C : 250

The normalized tf and idf is calculated in following manner.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

(i) **tf and idf of Term A**

$$tf = \frac{\text{Frequency of term } B}{\text{Maximum terms in document}} = \frac{3}{3}$$

idf= $\log_2$(Number of documents/Number of documents containing term A in it)

$= \log_2(10000/50)$ [∵ Document frequency of A = 50]

$= \log_2(200)$

$= 7.6$

∴ tf-idf is calculated as,

tf-idf = tf × idf

$= 7.6 \times 1$

$= 7.6.$

(ii) **tf and idf of Term B**

$$tf = \frac{\text{Frequency of term B}}{\text{Maximum termsin document}}$$

$= \frac{2}{3}$

$= 0.66$

idf= $\log_2$(Total number of documents/ Number of documents containing term B in it)

$= \log_2(10000/1300)$ [∵ Document frequency of B = 1300]

$= \log_2(7.6)$

$= 2.9$

∴ tf-idf is calculated as,

tf-idf = tf × idf

$= 2.9 \times 0.66$

$= 1.9.$

(iii) **tf and i df of Term C**

$$tf = \frac{\text{Frequency of term C}}{\text{Maximum termin document}}$$

$= \frac{1}{3}$

$= 0.33$

idf= $\log_2$(Number of documents/ Number of documents containing term C in it).

$= \log_2(10000/250)$

$= \log_2(40)$

$= 5.3.$

∴ tf-id is calculated as,

tf-idf = tf × idf

$= 5.3 \times 0.33$

$= 1.74$

Q21. What are the three major term weighting heuristics in the vector-space retrieval model? Explain in brief.

Answer : Dec.-14(R09), Q3(c)

The three major weighting heuristics terms in the vector space retrieval model are as follows,

(i) Term Frequency (TF)

(ii) Inverse Document Frequency (IDF)

(iii) Document length normalization.

(i) Term Frequency (TF)

For answer refer Unit-III, Q14.

(ii) Inverse Document Frequency (IDF)

For answer refer Unit-III, Q15.

(iii) Document Length Normalization

It is a technique that normalizes the effect of document length on document ranking by setting the term frequency or relevance score accordingly.

3.1.3 Natural Language

Q22. What do you mean by natural language processing? Explain index phrase generation in information systems.

Answer : Model Paper-II, Q6(b)

Natural Language Processing (NLP)

NLP is a method that is mainly used to improve document indexing in two ways,

(a) Semantically

(b) Statistically.

Hence, it will be easier for a user to identify what are the relevant documents based on the query specification or user can get the exact document information on the basis of query given to the system. Also, improvement of document indexing leads to generation of new phrases.

Document indexing can be improved semantically with the help of the following procedure.

To treat all the words in a document as a single entity and process them.

Another way in which phrases can be generated is based on statistics, by using a method known as proximity searching. This can be explained with the help of an example, consider a phrase "Computer Programming". Now apply a special proximity operator known as adjacency (ADJ). For instance,

"Computer ADJ Programming".

Due to this operator, the phrase "Computer Programming" seems similar to the another phrase "Programming Computer".

In addition to generation of phrases, NLP system also plays an important role in specifying how relationships can be determined among different concepts (topics/word phrases) in a particular document.

To do this, a special system is used by them called as 'DR-LINK' system. This system defines several types of relationships like semantic, general, statistical etc. The determination of each relationship is explained through various steps.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Step 1

Mapping of concepts in a document to special subject codes occur in step 1. These codes are specified in LDOCE (Longman's Dictionary Of Common English). In LDOCE, theoretical (priori) as well as statistical relationships are established among the word phrases (concepts) in a document with the help of a process called disambiguation. Also through disambiguation, subject codes are arranged in sequential order. As a result, it is easier to identify which word phrase in a document is more appropriate for a particular subject code.

Step 2

Once the statistical as well as theoretical relationships have been identified, DR-LINK system determines the general relationships among the concepts in a document by using a special tool known as text structure. Using this tool, a particular concept in a document can be partitioned into various fields for easy understanding.

Step 3

Here, DR-LINK system identifies semantic relationships through another relationship called connector relationship. Consider an example of a document in which there are two concepts – "bad performances" and "lack of team spirit". For ordering these two concepts or topics, the use of connector relationship "as a result of" is critical. Hence, one can easily make out whether bad performances was the reason for lack of team spirit or vice-versa.

Step 4

This is the final step. In this step, DR-LINK system assigns final weight values to the above mentioned relationships.

Index Phrase Generation (IPG)

Another method used in NLP for creation of phrases is IPG mechanism. In this method, phrases are designed in the form of indexes.

In order to implement this method, a special algorithm was used called Salton algorithm. This algorithm used a special factor called "Cohesion" factor for creation of phrases. The formula for this factor is given by,

$$\boxed{\text{norm_factor} * \left(\frac{\text{pair_freq}_{a,b}}{\text{ft}_a * \text{tf}_b}\right) = \text{coh}_{a,b}}$$

or

$$\boxed{\text{Size_factor} * \left(\frac{\text{pair_freq}_{a,b}}{\text{tf}_a * \text{tf}_b}\right) = \text{coh}_{a,b}}$$

Here, cohesion is denoted by coh.

norm_factor – Used for normalization on the basis of vocabulary size.

$\text{pair_freq}_{a,b}$ – Rate of co-occurrence (it can be defined by using adjacency, word or sentence proximity) of the pair of terms '*a*' and '*b*' in the same document.

Slight modifications have been made in this algorithm. These modifications are explained in the form of following rules,

First Rule: The above term pair or any other pair of terms in a document should include non-stop words that are attached to each other are referred to as a phrase.

Second Rule: The above term pair or any other pair of terms should not be available in less than 25 documents.

Third Rule: Normalization process can be implemented only if the division operation is performed successfully on the length of subvector of a single word.

Fourth Rule: A variation of SMART system (single-term algorithm) should be used by phrase weighting.

Natural language approach has an advantage over statistical approach in the creation or identification of phrases. Given a phrase, statistical approach can only divide the particular phrase into two sub phrases, whereas, by using natural language approach, the phrase can be partitioned into more than two sub phrases.

Examples

Consider few examples.

1. "Object-oriented Programming Languages"

By using statistical approach, we get,

(i) "Object-oriented Programming"

(ii) "Programming Languages".

By using natural language approach, the phrases obtained are,

(i) "Object-oriented Language"

(ii) "Programming Language"

(iii) "Object-oriented Programming Language".

2. "Information Retrieval Systems"

With the help of statistical approach, the phrases obtained are,

(i) "Information Retrieval"

(ii) "Retrieval Systems".

By using natural language approach, the phrases acquired are,

(i) "Information System"

(ii) "Retrieval System"

(iii) "Information Retrieval System".

Q23. Write short notes on Tagged Text Parser (TTP).

Answer :

Tagged Text Parser (TTP)

Phrases can also be determined in a lexical format by using a part of speech tags. These tags can be divided into various classes. The classes along with the examples are given below.

CLASSES		EXAMPLES
1.	Singular nouns	action, bag, activity, conclusion, accent etc.
2.	Plural nouns	cattle, bags, activities, borders, actions, etc.
3.	Determiners	a, an, the
4.	Adjective	confused, delightful, beautiful, amazing, etc.
5.	Present tense participle	multitasking, multithreading, multiprogramming etc.
6.	Present tense verb	catch, accept, beat, be, derive, cook, etc.
7.	Prepositions	above, beyond, in front of, next to, in, for, between, etc.

Using the above parts of speech tags, TTP represents a particular sentence in the form of a tree structure known as TTP parse tree.

These type of trees consist of a special type of pairs called header modifier.

The headers describe the main abstract. The modifiers are used to indicate what things are required to form the abstract.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Example

An example to demonstrate a TTP parse tree representation of any sentence.

Sentence 1

Roger Federer has been a consistent tennis player for a long time.

[Consistent] – adj

[has] – verb

[been] – verb

[tennis] – noun

[player] – noun

[a] – determiner

[for] – preposition

[Roger Federer] – noun

[a] – determiner

[long] – verb

[time] – noun

In order to know whether header modifier needs to be indexed or not, a special type of value called "Informal Contribution" (IC) is used. The formula for determining this value is,

$$\frac{F_{a,c}}{n_a + d_a - 1} = IC(a,[a,c])$$

Where,

n_a – Denotes the number of pairs in which the term 'a' has occurred in the same position as is positioned in (a, c)

d_a – Indicates those different words that are paired with the term 'a'

$f_{a,c}$ – Rate of occurrence of the terms (a, c) in the database.

3.1.4 Concept Indexing

Q24. Explain in detail about concept indexing.

Answer : **(Model Paper-III, Q6(b) | Dec.-19(R16), Q6)**

Concept Indexing

In natural language processing method, concepts in a document can be represented in the following manner,

(i) Semantically

(ii) Statistically

(iii) Generally.

Whereas in concept indexing, concepts are represented in a more orthodox format. Consider an example of a document which consists of a word "Computer". It is related to different concepts like,

(a) Electronic device

(b) Storage device

(c) Processing device

(d) Programming.

But, the disadvantage of this approach is that it is difficult to make out which concept is more appropriate for "Computer". To overcome this, special type of algorithms are implemented by this indexing method called neural network algorithms. In these algorithms, each concept of a term is assigned different weight values. Based on these values, determination of which concept is related to a given term is easier. Also the given word in a document is represented in the form of a vector on the basis of weights. This vector is known as concept vector.

Using the above example to demonstrate this process.

Word	Related Concepts	Weight Values (Assumed)
Computer	Electronic device	0.50
	Storage device	0.35
	Processing	0.12
	Programming	0.11

Thus, vector representation for computer is (0.50, 0.35, 0.12, 0.11). From the above weights, it can be determined that electronic device and storage device are more closely related to computer.

One of the methods that implement neural network algorithm is LSI approach. LSI stands for "Latent Semantic Indexing". LSI analyses the relationship among the documents and the terms by providing concepts that are related to documents and the terms.

This type of document representation is done with the help of a term-document matrix, where all the words are represented in terms of rows and documents in terms of columns. An interesting point to be noted is that, this matrix is formed by multiplication of three sub-matrices. Let this matrix be denoted by 'O' and the three sub-matrices be denoted by A, B, C.

Therefore, 'O' matrix is computed by,

$$O = A * B * C$$

Where, if the matrix 'O' consists of M rows and N columns ($M \times N$), then

(i) A consists of M rows and R columns ($M \times R$)

(ii) B consists of R rows and N columns ($R \times N$)

(iii) C consists of R rows and R columns ($R \times R$)

Hence, it can also be written as,

$$M \times N = (M * R) * (R * R) * (R * N)$$

R is the rank of O. The above matrix formula is feasible if B has smaller values. Suppose B has larger values, then the revised formula is,

$$O = A_n * B_n * C_n$$

Note

Both A and B are orthogonal matrices.

Q25. What is meant by automatic indexing? Discuss about statistical indexing and concept indexing.

May/June-13(R09), Q3

OR

Complete statistical indexing and concept indexing.

(Refer Only Topics: Statistical Indexing, Concept Indexing)

Nov.-13(R09), Q3

Answer :

Automatic Indexing

For answer refer Unit-III, Q12.

Statistical Indexing

For answer refer Unit-III, Q13.

Concept Indexing

For answer refer Unit-III, Q24.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

3.1.5 Hypertext Linkages

Q26. Explain briefly about Hypertext linkages.

Answer : **Model Paper-I, Q7(a)**

Hypertext Linkages

One of the most commonly used methods for retrieval of information is hypertext with linkages. Consider an example wherein you want to know about frequently used mail sites. Hence, when a user opens any search engine and type the search statement. Based on the statement, various sites will be displayed. These are nothing but hypertext or hyperlinked documents. Once a user click on any one of these links, you will enter the home page of that site. For example, www.yahoo.com. Suppose, a user want to know how to create a mail account in yahoo, there will be a sublink regarding this. When it is clicked, a registration form will be displayed. If a user want to enquire about games in yahoo, a separate sublink will be there. Hypertext linkages not only help you to find relevant information about a particular document, but also additional information about it. The above process is the combination of two techniques called information retrieval through manual indexing and information retrieval through intelligent agents or web crawlers. The drawbacks of these techniques are that they are exclusively used for searching process but not for indexing purposes.

Besides this, the other technique known as information retrieval through automatic indexing. The disadvantage of this method is that no linkage facility is available. In short, no sublinks are there. All the information will be displayed on one site.

Examples of the sites that implement automatic indexing,

(i) www.altavista.com

(ii) www.lycos.com.

To overcome the above drawbacks, an index algorithm is implemented that searches for documents with linkages and also used for indexing purpose.

The working of this algorithm is based on the following method,

Inclusion of those words in the database or index of the hyperlinked document that have stronger weight values.

The formula for determining such concepts is given by,

$$\boxed{(\alpha * wt_{a,b} + \beta * wt_{c,d}) * (\gamma * \text{link}_{a,c}) = wt_{a,b,c,d}}$$

Where,

α, β and γ are normalization factors.

$\text{link}_{a,c}$ – Specifies the strength of the link. For example, weak link, strong link etc.

$wt_{a,b,c,d}$ – Indicates the final weight value of word '*b*' in document '*a*' and of word '*d*' in document '*c*'.

3.2 DOCUMENT AND TERM CLUSTERING

3.2.1 Introduction to Clustering

Q27. Define clustering. Explain the process of clustering in information system and its types.

Model Paper-II, Q7(a)

OR

Explain document clustering with example. **Oct./Nov.-20(R16), Q5(a)**

(Refer Only Topic: Document Clustering)

OR

Describe the concept of term clustering. Discuss in detail about item clustering. **May/June-13(R09), Q4**

(Refer Only Topics: Term Clustering, Item Clustering)

OR

What is term clustering? Explain item clustering with suitable examples.

(Refer Only Topic: Term Clustering)

Answer : Nov.-13(R09), Q4

Clustering

Clustering means 'Grouping' of similar objects that belong to the same subject. Its goal is to assist the user in easily finding out the location where the information is stored. This gave rise to indexing schemes for organizing objects in libraries and standards of electronic indexes used. The generation of thesaurus, i.e., 'Treasure' in Latin, was originated by clustering different words. A thesaurus stores the antonyms and synonyms for different words as well as dictionary stores their definitions within it. Clustering is actually used to group all the identical objects into a 'class' that holds a more general title. The terms 'Cluster' and 'Class' have the same meaning.

Process of Clustering in Information System

The process of clustering in information systems has the following steps,

1. A domain for the cluster is defined. This corresponds to the determination of the scope of thesaurus, while creating a thesaurus. When a set of objects that need to be clustered is determined, it is called document clustering. Defining a cluster domain is very important because it specifies those objects that can be used in the process of clustering so as to reduce the errors.
2. After the identification of domain the object's attributes that are to be clustered are determined. During thesaurus generation, only those words that belong to objects and those used in the process of clustering are determined. In the same way, document clustering process revolves around the set of zones in the items that help in determining the resemblance.
3. When two or more attributes of objects are occurred together which are supposed to be in the same class, then the strength of the relationship they share is determined. With respect to thesaurus generation, it corresponds to the determination of synonymous words along with the determination of their relationship strengths. With respect to documents, it corresponds to defining a resemblance function which is derived from the words that occur together and gives the amount of resemblance between two items.
4. When the above three steps are performed, the determination of the complete set of objects along with their relationship strengths will be fulfilled. In this final step, a class to which the items are to be assigned is determined. This is done with the help of some algorithm.

Types of Clustering

Clustering are of three types. They are as follows,

1. Document clustering
2. Term clustering
3. Item clustering.

1. Document Clustering

Document clustering is a type of clustering done on textual documents. It is also known as text clustering. It is an application of cluster analysis that focuses on organizing the document automatically, extracting the searched topics and quick retrieval of information and filtering. Moreover, Document Clustering is considered as a grouping process that sets the documents into cluster of similar documents. Its search technique helps in retrieving items that are similar to the search result.

Example

Consider a web search engine. A web search engine is basically used for retrieving multiple search pages in response to the generated query by eliminating irrelevant information. Here, document clustering method automatically groups the retrieved information as generates a list of meaningful words/sentences/phrases.

2. Term Clustering

Term clustering is a type of clustering done on similar terms. This clustering is done to create statistical thesaurus. It enables a user to expand the search by retrieving the similar terms based on the search query. This technique provides completeness based on correlated terms. Here, clusters are created by passing data at once. Term clustering can be done either manually (or) automatically. Term clustering done automatically is referred to as automatic term clustering.

Automatic Term Clustering

For answer refer Unit-III, Q33.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

3. **Item Clustering**

Item clustering is a clustering technique similar to term clustering of thesauri generation. Item clustering can also be done either manually or automatically. Manual item clustering can be done based on library or filling system. In this, the user can generate the item based on the category to which it belongs. Here, each item is assigned to a single category. Later indexing is performed on the set of clusters to retrieve the items that are physically stored in the database.

Automatic item clustering generates clustering of items automatically. In this, technique, similarity between two documents depends on two items that have common terms and items. Hence, similarity function is performed based on rows on items in a vector matrix. Thus, the similarity equation is given as,

$$S(\text{Item x, Item y}) = \Sigma(\text{Term}_{x,z})\ (\text{Term}_{y,z})$$

Example

For answer refer Unit-III, Q40.

Q28. Give the guidelines for determining the items and characteristics of a class.

Answer :

The guidelines for determining the items and characteristics of a class are as follows,

1. All the classes should have a well-defined semantic definition. The names of the semantic definitions of different classes can be ambiguous. Thus, the classes are assigned with numbers in some systems.

Example

Consider a class called "Computer" which is a cluster of items. A user can think that, this class contains main memory items, which are actually present in the "hardware" class.

2. All the classes should have the same size. Classes are generally used in the query expansion or the expansion of the resultant set of items that have been retrieved. These two purposes cannot be fulfilled by a class that holds about 90 percent of the objects. It also consider the usefulness of other classes that are distributed among 10 percent of the leftover objects.

3. A class should not contain an object that dominates its own class.

4. During the creation of an object the decision about the assignment of an object to a single or multiple classes is taken. There is a trade-off between two capabilities of object semantics that is, specificity and partitioning. When the language ambiguity is considered, it will be preferred to assign an object to multiple classes. When this is done, the creation and maintenance of objects will become more complex.

Q29. Which of the additional decisions can be incorporated in an automatic statistical thesaurus construction program? Describe how they would be implemented and the risks with their implementation.

Answer :

The additional decisions incorporated in an automatic statistical thesaurus construction program are,

1. **Word Coordination Method**

It describes whether or not to cluster the phrases along with the individual terms.

2. **Word Relationships**

There is a possibility of having different kinds of relationships between words if a human interface is involved in generating thesaurus.

There are three different types of relationships. They are,

(i) Equivalence relationship

(ii) Hierarchical relationship

(iii) Non-hierarchical relationship.

(i) Equivalence Relationship

This type of relationship is the most commonly used to denotes synonyms. While creating a thesaurus, care must be taken in defining a synonym as there exists few terms that are different but have some sort of similarity. For instance, the terms such as "print" and "photograph" can be specified as synonyms even though lithography is also included in printing. Some terms that have the same 'role' but different meanings are also included. For instance, the words such as 'moron' and 'genius' are synonyms in the "intellectual capability" class as they define the intellectual capability but have different meanings.

(ii) Hierarchical Relationship

This type of relationship is also very common. In this type of relationship, a class name is a more general term and holds the entries that are examples of that general term. For instance, a "Computer" is a class containing entries such as "Pentium", "Microprocessor" etc.

(iii) Non-hierarchical Relationship

Other kinds of relationships such as "object" – "attribute" relationships are included in non-hierarchical relationship. For instance "employee" and "job-title" can be included in "object"-"attribute" relationship.

There can be other relationship between words such as,

1. Parts-wholes
2. Collocation
3. Paradigmatic
4. Synonymy
5. Taxonomy
6. Antonymy
7. Contrasted words
8. Child-of
9. Parent-of
10. Part-of.

3. Homograph Resolution

A word that has several meanings which are completely different from each other is called a "homograph". For instances, the word "light" can mean "brightness" and it can also mean "not being too heavy". A homograph cannot be eliminated by providing it a unique meaning. Thus, the user will have to interact with the system to obtain the desired meaning. This can be done by allowing the user to enter more than two search terms.

4. Vocabulary Constraints

It contains the normalization and specificity instructions of the vocabulary. Normalization will limit the thesaurus to contain only the stems rather than the complete words. A set of words can be eliminated or general words can be used to denote class identifiers with the help of specificity.

When the terms are clustered properly recall is increased thereby assisting the users. But, whenever recall is increased, precision will be decreased. This should not happen, because the human processing of the retrieved items cannot be managed. Thus, relevant measure should be taken to stop a decrease in precision whenever there is an increase in recall.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Q30. What is the difference between document clustering and term clustering?

Answer :

(Model Paper-III, Q7(a) | Nov./Dec.-12(R09), Q4(a))

	Document Clustering		Term Clustering
1.	Document clustering is similar to arrangement of books in traditional libraries wherein books are grouped together according to subjects.	1.	Term clustering is similar to a thesaurus which provides synonyms and antonyms for every word.
2.	It expands search by using documents similar to the documents used for retrieval.	2.	It expands search by using the terms that were used in the query.
3.	A document may or may not belong to multiple clusters.	3.	A term belongs to multiple clusters.
4.	It is performed by focusing on the specific zones within the retrieved items. Examples: Titles in articles.	4.	It is performed by determining the specific words of the objects used in clustering.
5.	The domain of the clusters determines set of items to be clustered.	5.	The domain of the clusters determines the scope of the objects to be clustered.
6.	The strength of the relationship is determined by defining similarities found based on the word co-occurrence.	6.	The strength of the relationship is determined by synonym words.

Q31. Will the Clustering process always come to the same final set of Clusters, no matter what the starting Clusters? Explain your answer.

Answer :

Dec.-19(R16), Q7

The clustering process may or may not produce the same final set of clusters regardless of the different initial clusters. Consider an example process of k-means clustering.

k-means Clustering

The *k*-means is an iterative clustering algorithm in which objects are moved among sets of clusters until the desired set is achieved. It is the most popular and commonly used method. The algorithm is built on the concept of user specified input parameter (*k*). A set of '*n*' objects are divided into '*k*' clusters by the algorithm. A high degree of similarity among elements in clusters is obtained, while a high degree of dissimilarity among elements in different clusters is achieved simultaneously. The cluster's centroid gives the measure of cluster's similarity.

Working of k-means Algorithm

Algorithm's Input

The number of desired clusters are denoted by *k*

- A data set containing '*n*' objects is denoted by '*D*'.

Algorithm's Output

k : A set consisting of *k* clusters.

Procedure

Step 1

Initially select '*k*' objects randomly from *D*, as initial cluster centers.

Step 2

Depending upon the distance between the object and the cluster mean, each remaining object is assigned to the cluster to which it is most similar or near.

Step 3

Calculate new mean value of the object for each cluster.

Step 4

The step (3) is repeated and the process iterates until the criterion function converges. The resulting 'k' clusters are compact and separate. k-means method, typically, uses the square-error criterion function, which is expressed as,

$$S = \sum_{a=1}^{k} \sum_{e \in C_a} |e - \text{Mean}_a|^2 \text{ and}$$

$$\text{Mean}_a = \frac{1}{\text{Mean}} \sum_{j=1}^{\text{Mean}} n_{aj}$$

Where,

$S =$ The sum of the square error for all objects in the data set.

$e =$ The point in space that represents the given object.

$\text{Mean}_a =$ The mean of cluster C_a.

Example

Consider a triangular space containing a group of numbered objects. Let, $k = 2$ be the numbers of clusters desired by the user. The algorithm proceeds as follows.

Randomly choose two objects as initial cluster centers marked as ✡ in the figures with mean values as $m_1 = 2$ and $m_2 = 4$. And find the Euclidean distance between the mean and the objects to classify objects into two clusters as shown in figure (1).

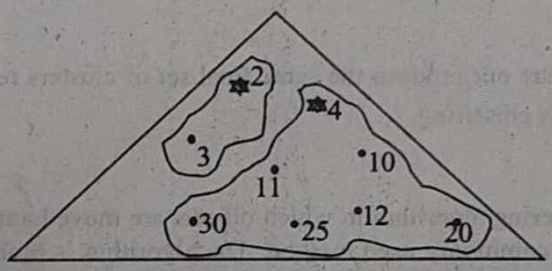


Figure (1): Initial Partitioning with $m_1 = 2$ and $m_2 = 4$

Now, evaluate new mean for the resulting clusters and again partition the objects is done on the basis of euclidean distance as shown below,

$$m_1 = \frac{2+3}{2} = 2.5$$

$$m_2 = \frac{4+10+11+12+20+25+30}{7} = 16.$$

$m_1 = 2.5$ and $m_2 = 16$

The resulting cluster is shown in figure (2).

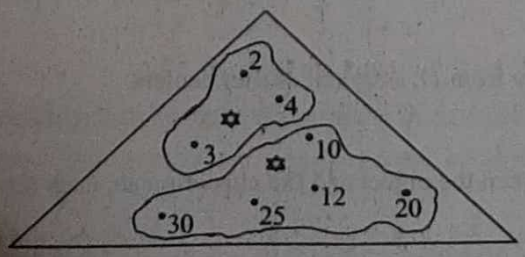


Figure (2): New Clusters Formed with $m_1 = 2.5$ and $m_2 = 16$

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

The process iterates for the successive values of the mean, to enhance the partitioning of the clusters. At last there is no redistribution of objects in any cluster and the process terminates when the means become identical and converge. The following figures depict the iterative relocation.

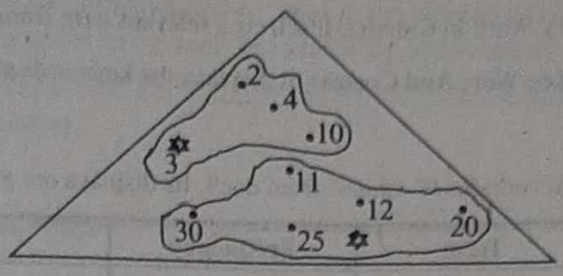


Figure (3): Clusters with $m_1 = 3$ and $m_2 = 18$

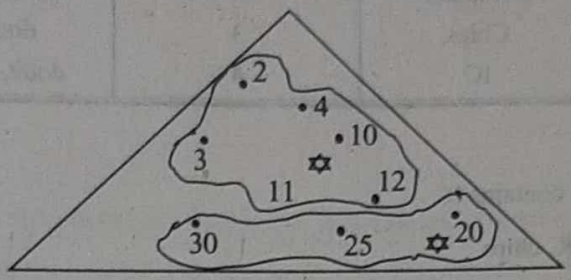


Figure (4): Clusters with $m_1 = 4.75$ and $m_2 = 19.6$

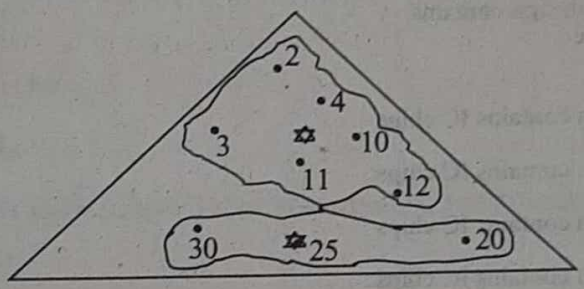


Figure (5): Final Clusters are Returned by the Clustering Process with $m_1 = 7$ and $m_2 = 25$

Finally, the correct partitioning of the numbered objects into two clusters is achieved as desired by the users.

Thus, from the above example it can be concluded that clustering process may or may not come to the same final set of clusters.

3.2.2 Thesaurus Generation

Q32. Explain in detail about the manual clustering method.

Answer :

Manual clustering enables a user to generate a thesaurus by carrying out the following steps,

1. A domain for clustering is determined as it reduces ambiguities that may occur because of the homographs and will thus help in focusing the creator.
2. From the existing thesaurus, a starting point for a set of potential words can be generated that can be included in the new thesaurus. For this, the concordances from items are selected. Here, a concordance corresponds to an alphabetical listing of words that is taken from a set of items including the frequency of occurrence of those words along with the references of items that holds these words.
3. Manual thesaurus construction mainly depends on the selection of a group of words that are to be included in the new thesaurus. Thus, those words that are not related to the domain of thesaurus and those words whose frequency of occurrence is very high should not be included in the new thesaurus. Also, the words that doesn't contain any information value must be discarded.

Useful words can be determined using tools such as KWOC, KWIC and KWAC, when the concordances are used.

- **KWOC:** KWOC is referred as Key Word Out of Context and is also called as concordance.
- **KWIC:** KWIC is referred as Key Word In Context. It selects a relevant term from its phrase context and displays it.
- **KWAC:** KWAC is referred as Key Word And Context. It displays the keywords along with their contexts.

Example

Let the phrase "Computer design contains IC Chips" from doc9. Its displays are given below,

KWOC	Term	Frequency	Ids of items
	Design	2	doc5, doc9
	Computer	4	doc9, doc3, doc1, doc6
	Chips	3	doc1, doc9, doc11
	IC	4	doc9, doc7, doc2, doc6

KWIC

Chips	Computer design contains IC
Computer	Design contains IC chips
Design	Contains IC chips/computer
IC	Chips/computer design contains.

KWAC

Chips	Computer design contains IC chips
Computer	Computer design contains IC chips
Design	Computer design contains IC chips
Memory	Computer design contains IC chips.

KWIC contains a "/" character to denote the termination of the phrase. To find out the meaning of homographs, KWIC and KWAC are used. As, the term "Chips" could mean either the wood chips or the IC chips, KWOC does not hold or display any information that will help in resolving the ambiguity. Whereas, the displays of KWIC and KWAC helps the editor of thesaurus to read the sentence fragment of the term and find out this meaning.

4. After selecting the terms, clustering of those terms is carried out, depending on the guidelines of the word relationship and the interpretation of the relationship strength.
5. Thesaurus thus obtained, will be treated with several quality assurance reviews. This is done by other editors before finalizing the thesaurus.

Q33. Explain about automatic term clustering.

Answer : Model Paper-I, Q7(b)

A statistical thesauri can be created by generating term clusters automatically with the help of several techniques. The basic concept employed by automatic term technique that, when two terms occur frequently in the same item then there is a very high possibility that they have the same concept. The technique provides completeness based on which the terms are correlated. Depending on the completeness of the correlation, the overhead of time and computation changes in the creation of clusters. There will be an overhead of $O(n^2)$ when the relationship strength between each of the combinations of 'n' distinct words is computed by a process whose degree of completeness is very high. There are several additional techniques as well, that start with an arbitrary set of clusters and iterate on each of the terms that are assigned to those clusters. In the simplest case, clusters are created by passing the data only once. If very large number of clusters are created, then a hierarchy of clusters can be formed by making the initial clusters as the starting point for generating clusters which are more abstract.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

The automatic generation of a thesaurus depends on a set of items that corresponds to a vocabulary that is to be included in the thesaurus. The first step of domain determination for the thesaurus is actually the selection of these set of items. These items contain processing tokens which form the attributes that are used in the creation of clusters. The implementation of remaining steps may differ depending on the algorithm being used. The clustering of documents using automated method depends on the polythetic clustering. In this type of clustering, a set of words and phrases are used to define a cluster.

Q34. Write short notes on following,

(i) Term-term matrix

(ii) Term-term relationship matrix

(iii) Clique technique

(iv) Single-link technique

(v) Star technique.

Answer :

(i) **Term-Term Matrix**

Term-term matrix is a type of term-document matrix that specifies the frequency of terms that occurs within a document. This matrix is similar to the mathematical matrix in which rows specifies collection of documents and columns describes the terms. Term-term matrix for a vector model is used for calculating the similarities between each term. The formula to determine similarities is as follows,

$$S(Term_x, Term_y) = \Sigma(Term_{z,x})\,(Term_{z,y})$$

(ii) **Term-Term Relationship Matrix**

Term-Term relationship matrix allows representation of document term relationship. This matrix defines set of documents that are associated with indexed terms.

(iii) **Clique Technique**

It is a clustering technique that focuses on determining those terms that share similar classes.

For remaining answer refer Unit-III, Q37.

(iv) **Single Link Technique**

It is a clustering technique in which, terms that are similar with in a class are selected are added.

Algorithm for single link clustering is given below,

1. Select term T_x that does not belong to the existing class and place it in a new class.
2. All the other terms that are associated with the term T_x are placed in the same class.
3. Whenever a term is placed in a class, perform step 2.
4. When step 2 does not result in the determination of a new term, go to step 1.

In the first step of this algorithm, only those terms are selected and placed in a new class, which do not belong to any other class. Thus, if a term is present in a class, it can't be selected again for placing it in another class. A class contain terms that are similar to any other term in the same class.

Thus, multiple clusters cannot have the same term within them, when clustering is performed using single-link technique.

(v) **Star Technique**

In this clustering technique, a term is selected and placed in a class containing related terms. Here, the term which does not belongs to a class are selected as new speed for next star as smallest term.

Q35. Given the following term-term matrix.

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9
T_1		14	9	0	3	0	12	0	16
T_2	14		0	6	4	0	14	0	11
T_3	9	0		12	7	4	1	0	14
T_4	0	6	12		3	0	14	9	8
T_5	3	4	7	3		12	6	16	0
T_6	0	0	4	0	12		9	2	9
T_7	12	14	1	14	6	9		0	12
T_8	0	0	0	9	16	2	0		8
T_9	16	11	14	8	0	9	12	8	

Table: Term-term Matrix

(a) Determine the term relationship matrix using a threshold of 10 or higher.

(b) Determine the clusters using the clique technique.

(c) Determine the clusters using the single link technique.

(d) Determine the clusters using the star technique where the term selected for the new seed for the next star is the smallest number term nor already part of a class.

Answer :

(a) Determination of Term-relationship Matrix using a Threshold of 10 or Higher

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9
T_1		1	0	0	0	0	1	0	1
T_2	1		0	0	0	0	1	0	1
T_3	0	0		1	0	0	0	0	1
T_4	0	0	1		0	0	1	0	0
T_5	0	0	0	0		1	0	1	0
T_6	0	0	0	0	1		0	0	0
T_7	1	1	0	1	0	0		0	1
T_8	0	0	0	0	1	0	0		0
T_9	1	1	1	0	0	0	1	0	

Table: Term-relationship Matrix

As the given threshold value is 10, the two terms are said to be similar only if they have a similarity value of 10 or greater, between them. The resulting matrix contains '1', if they have a similarity value as 10 or greater and it contains '0' if they have similarity value, less than 10. This matrix is called as term-relationship matrix.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Determination of Clusters using Clique Technique

(b) 1. Using the above term-relationship matrix, determine those terms that can share same class (i.e., terms that have threshold as '1' with respect to a given term)

Class1 : T_1, T_2, T_7, T_9

Class 2 : T_2, T_1, T_7, T_9

Class3 : T_3, T_4, T_9

Class4 : T_3, T_4, T_7

Class5 : T_5, T_6, T_8

Class6 : T_5, T_6

Class7 : T_1, T_2, T_4, T_7, T_9

Class8 : T_5, T_8

Class9 : T_1, T_2, T_3, T_7, T_9

2. Classes that are subsets of other classes must be deleted. The resulting classes are,

Class1 : T_3, T_4, T_9

Class2 : T_3, T_4, T_7

Class3 : T_5, T_6, T_8

Class4 : T_1, T_2, T_4, T_7, T_9

Class5 : T_1, T_2, T_3, T_7, T_9

3. Now, the validations should be performed to check whether each term in a given class is within the threshold of all other terms present in the same class.

Consider Class1, that contain terms T_3, T_4 and T_9. As T_4 and T_9 cannot come in a single class T_3 and T_4 are placed in a single class.

Thus, one of the clusters formed as,

$\boxed{\text{Cluster1}: T_3, T_4}$

And terms T_3, and T_9 are placed in a separate class. Thus, the cluster formed is,

$\boxed{\text{Cluster 2}: T_3, T_9}$

Consider Class5, that contains terms T_3, T_4 and T_7. As T_3 and T_7 cannot come in a single class, and T_3 and T_4 are already placed in a separate cluster, T_4 and T_7 can also be placed in a separate cluster.

$\boxed{\text{Cluster 3}: T_4, T_7}$

Consider Class3, that contains terms T_5, T_6 and T_8. As T_6 and T_8 cannot come in a single class, the terms T_5 and T_6 are placed in a separate class.

$\boxed{\text{Cluster 4}: T_5, T_6}$

The terms T_5 and T_8 are placed in a separate class,

$\boxed{\text{Cluster 5}: T_5, T_8}$

Consider Class4, that contains terms T_1, T_2, T_4, T_7 and T_9. The terms T_1, T_2, T_7 and T_9 can come in a single class.

$\boxed{\text{Cluster 6}: T_1, T_2, T_7, T_9}$

And terms T_4 come only with T_7 in a class which is already formed.

Consider Class5, that contains terms T_1, T_2, T_3, T_7 and T_9. As T_1, T_2 and T_7 can come in a single class, its cluster can be formed. But it will be a subset of cluster 6. Thus, a new cluster will not be formed.

Also, T_3, can come only with T_9, So they can be placed in a single cluster, but it is already formed.

(c) Determination of Clusters Using Single Link Technique

In this technique, if any of the terms present in a class is similar to any other terms selected, then it can be added to the same class. It should be noted that no two classes will have the same term.

Thus, using the above term-relationship matrix the cluster formed are,

$\boxed{\text{Cluster 1}: T_1, T_2, T_7, T_9, T_4, T_3}$

As T_1, T_2, T_7, and T_9, are within the threshold, they are selected. But T_4 is similar to T_7 and T_3 is similar to T_9, so they are also placed in the same class.

$\boxed{\text{Cluster 2}: T_5, T_6, T_8}$

As T_5, T_6, T_8 are within the threshold, they are placed in a single cluster.

(d) Determination of Cluster Using the Star Technique

In this technique, a term is selected and placed in the class containing related terms.

Here, the term does not belongs to a class are selected as new seed for the next star as the smallest number term.

To start with, T_1 is selected and is placed in cluster1. T_1 is similar to T_2, T_7 and T_9 so they all are placed in cluster1.

$\boxed{\text{Cluster1}: T_1, T_2, T_7, T_9}$

Now, T_3 is selected as a new seed because T_3 is not included in cluster1. As the term T_3 is similar to T_4 and T_9, they all are placed in a new cluster *viz.*, cluster2.

$\boxed{\text{Cluster 2}: T_3, T_4, T_9}$

Finally, T_5 is selected as a new seed because it is not a part of any cluster. T_5 is similar to T_6 and T_8, so they all are placed in a new cluster *viz.*, cluster3.

$\boxed{\text{Cluster3}: T_5, T_6, T_8}$

Q36. Given the following vector model (matrix).

	$Term_1$	$Term_2$	$Term_3$	$Term_4$	$Term_5$	$Term_6$	$Term_7$	$Term_8$
Item1	0	4	0	0	0	2	1	3
Item2	3	1	4	3	1	2	0	1
Item3	3	0	0	0	3	0	3	0
Item4	0	1	0	3	0	0	2	0
Item5	2	2	2	3	1	4	0	2

Table: Vector Model

(a) Determine the term-term matrix.

(b) Determine the term-relationship matrix.

(c) Determine the clusters using the clique technique.

(d) Determine the clusters using the single link technique.

(e) Determine the clusters using the star technique.

Answer :

(a) Determination of Term-term Matrix

Term-term matrix for the given vector model is determined by calculating the similarities between each term using the following formula.

$$\boxed{S(Term_x, Term_y) = \Sigma(Term_{z,x})(Term_{z,y})}$$

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Where 'z' can be any one of the items. Using this formula, similarity between $Term_1$ and $Term_2$ is determined as follows,

$$S(Term_1, Term_2) = (Term_1, Item1)(Term_2, Item1) + (Term_1, Item2)(Term_2, Item2) + (Term_1, Item3)(Term_2, Item3) + (Term_1, Item4)(Term_2, Item4) + (Term_1, Item5)(Term_2, Item5)$$
$$= (0 \times 4) + (3 \times 1) + (3 \times 0) + (0 \times 1) + (2 \times 2)$$
$$= 3 + 4$$
$$= 7$$

Similarly, the similarities between each of the eight terms are calculated and are placed in an $n \times n$ matrix, called as term-term matrix. Here, 'n' corresponds to the total number of columns or terms in the vector model. This matrix does not have any values along its diagonal as it denotes the similarity of a word with itself.

The term-term matrix for the given vector model is shown in the table (1).

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
T_1		7	16	15	14	14	9	7
T_2	7		8	12	3	18	6	17
T_3	16	8		18	6	16	0	8
T_4	15	12	18		6	18	6	9
T_5	14	3	6	6		6	9	3
T_6	14	18	16	18	6		2	16
T_7	9	6	0	6	9	2		3
T_8	7	17	8	9	3	16	3	

Table (1): Term-term Matrix

(b) **Determination of Term-relationship Matrix**

Term-relationship matrix for the above term-term matrix is given in table (2). It uses a threshold of 10.

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
T_1		0	1	1	1	1	0	0
T_2	0		0	1	0	1	0	1
T_3	1	0		1	0	1	0	0
T_4	1	1	1		0	1	0	0
T_5	1	0	0	0		0	0	0
T_6	1	1	1	1	0		0	1
T_7	0	0	0	0	0	0		0
T_8	0	1	0	0	0	1	0	

Table (2): Term-relationship Matrix

As the given threshold value is 10, then the two terms are said to be similar only if they have a similarity value of 10 or greater than 10 between them. The resulting matrix contains '1', if they have a similarity value as 10 or greater and it contains '0', if they have similarity value less than 10. This matrix is called as term-relationship matrix.

(c) **Determination of Clusters using Clique Technique**

1. Using the above term relationship matrix, determine those terms that can share same class (i.e., those terms that have threshold as '1' with respect to a given term).

Class1 : T_1, T_3, T_4, T_5, T_6
Class2 : T_2, T_4, T_6, T_8
Class3 : T_3, T_1, T_4, T_6
Class4 : T_1, T_2, T_3, T_4, T_6
Class5 : $T_1, T_5,$
Class6 : $T_1, T_2, T_3, T_4, T_6, T_8,$
Class7 : $T_7,$
Class8 : $T_2, T_6, T_8,$

2. Classes that are subsets of other classes must be deleted. The resulting classes are,

   Class1 : T_1, T_3, T_4, T_5, T_6

   Class2 : $T_1, T_2, T_3, T_4, T_6, T_8$

   Class3 : T_7

3. Now, the validations should be performed to check whether each term in a given class is within the threshold of all other terms present in the same class.

Consider class1, it contain terms T_1, T_3, T_4, T_5 and T_6. Terms (T_3, T_5), (T_4, T_5) and (T_5, T_6) cannot come together in a single class as they are not similar. So, T_5 being common in all these pairs, is removed from the class. Thus, the cluster formed is,

$\boxed{\text{Cluster1}: T_1, T_3, T_4, T_6}$

The term T_5 is similar to T_1, thus, these terms can be placed in a separate cluster.

$\boxed{\text{Cluster2}: T_1, T_5}$

Consider Class 2. It contains terms T_1, T_2, T_3, T_4 T_6 and T_8. The pairs of dissimilar terms that cannot be placed in a single class are (T_1, T_2), (T_1, T_8), (T_2, T_3), (T_3, T_8), (T_4, T_8). Thus, the clusters that are possible from class2 are,

1 : T_1, T_3, T_4, T_6 (Already formed)

2 : T_2, T_4, T_6

3 : T_2, T_6, T_8

Thus, the two new clusters formed are,

$\boxed{\text{Cluster3}: T_2, T_4, T_6}$ And $\boxed{\text{Cluster4}: T_2, T_6, T_8}$

Consider Class3, as there is only one term T_7 which is not similar to any of the other terms, it is placed in a separate cluster, cluster5 : T_7.

(d) Determination of Cluster using Single Link Technique

In this technique, if any of the terms present in a class is similar to any other term selected, then it can be added to the same class. It should be noted that, no two classes will have the same term.

Thus, using the above term-relationship matrix, the clusters formed are,

$\boxed{\text{Cluster1}: T_1, T_3, T_4, T_5, T_6, T_2, T_8}$

As T_1, T_3, T_4, T_5, T_6, are within the threshold, they are selected. In addition to that, T_6 is similar to both T_2 and T_8, so they are also placed in the same class.

$\boxed{\text{Cluster2}: T_7}$

As T_7 is the only term which is not included in cluster 1, it is placed in a separate cluster.

(e) Determination of Clusters Using the Star Technique

Here, the term selected as new seed for the next star is the smallest number term which is not already part of a class.

To start with, T_1 is selected and is placed in cluster 1. T_1 is similar to T_3, T_4, T_5 and T_6 so they are placed in cluster 1.

$\boxed{\text{Cluster1}: T_1, T_3, T_4, T_5, T_6}$

Now, T_2, is selected as a new seed because it is not included in cluster 1. Terms, T_4, T_6 and T_8 are similar to T_2, so they all are placed in a new cluster, cluster 2,

$\boxed{\text{Cluster2}: T_2, T_4, T_6, T_8}$

Then, T_7 is selected as a new seed as it is not included either in cluster 1, or cluster 2. No other term is similar to term T_7, thus, it is placed in a separate cluster.

$\boxed{\text{Cluster3}: T_7}$

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Q37. Write clique's algorithm for creating clusters.

Answer :

Clique's Algorithm

0. Let $j = 1$
1. Select $Term_j$ from all the available terms and place it in a new class.
2. Begin with $Term_z$ where $p = z = j + 1$
3. Validate if $Term_z$ is within the threshold of all terms present in the class under consideration.
4. If not, Let $z = z + 1$
5. If $z > n$ (where 'n' is the total number of words)

   Then $p = p + 1$

   if $p = n$ then go to 6

   else $z = p$

   Create a new class which contains $Term_j$ in it go to 3

   else go to 3
6. If the current class contains only the $Term_j$ and other classes also contain $Term_j$ then delete the current class

   else $j = j + 1$
7. If $j = n + 1$ then go to 8

   else go to 1
8. Discard all those classes which are the subsets or duplicates of other classes.

Q38. Explain existing terms clustering with an example.

Dec.-14(R09), Q4(b)

Answer :

Clustering Using Existing Clusters

This method makes use of existing clusters, inorder to easily determine new cluster by performing minimal similarity calculations. The terms that are initially assigned to the clusters are reevaluated by performing validations on each term to be assigned to a cluster. This process is carried out until there is minimal movement of terms between clusters. The number of calculations to find out the similarity is reduced by calculating a centroid for each of the clusters. In physics, centroid means the center of mass of a group of objects and in vectors it is the average of all the vectors present in a cluster.

Centroids, in this method are a point in the N-dimensional space, where N corresponds to the total number of items. The centroids produced after the initial assignment of terms are not associated with the terms in the final cluster. The similarity is calculated between the existing terms and the centroids of the clusters. The cluster that has the highest similarity is selected and the term is placed in it. This process continues until there is no change in cluster assignment. The calculations performed in this process are of the order $O(n)$.

The terms and centroids can be graphically represented as follows,

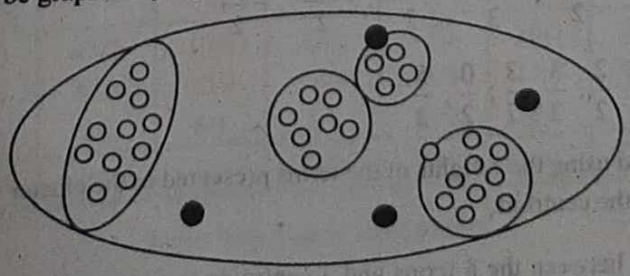


Figure (1): Centroids after Initial Assignment of Terms in Clusters

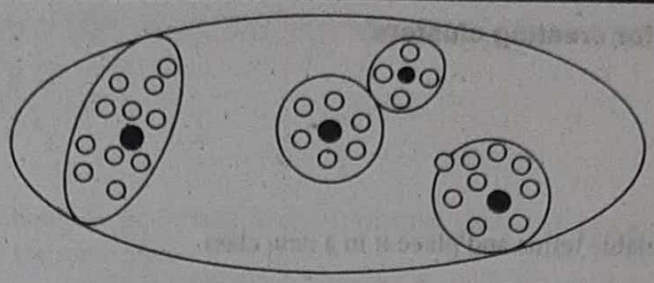


Figure (2): Centroids after Reassignment of Terms

The coloured circles in the above figures are the centroids of the classes. The ovals shown in both the figures are the clusters formed. After each iteration the similarity between the terms and the clusters is determined and cluster that has the highest similarity will be assigned to the term. These iterations are carried out until the stability is achieved. In each of these iterations new centroids are found that are more closer to a centroid that is related to an ideal cluster.

Example

Consider the following vector example,

	$Term_1$	$Term_2$	$Term_3$	$Term_4$	$Term_5$	$Term_6$	$Term_7$	$Term_8$
Item1	0	4	0	0	0	2	1	3
Item2	3	1	4	3	1	2	0	1
Item3	3	0	0	0	3	0	3	0
Item4	0	1	0	3	0	0	2	0
Item5	2	2	2	3	1	4	0	2

Table: Vector Model

Suppose Class1 has terms $Term_1$ and $Term_2$ and Class2 has terms Terms3 and Term4 and Class3 has terms Term5 and Term6.

Thus, the centroids produced for all these classes are,

$$\text{Class1} = (Term_1, Term_2) = \frac{0+4}{2}, \frac{3+1}{2}, \frac{3+0}{2}, \frac{0+1}{2}, \frac{2+2}{2}$$

$$= \frac{4}{2}, \frac{4}{2}, \frac{3}{2}, \frac{1}{2}, \frac{4}{2}$$

$$\text{Class2} = (Term_3, Term_4) = \frac{0+0}{2}, \frac{4+3}{2}, \frac{0+0}{2}, \frac{0+3}{2}, \frac{2+3}{2}$$

$$= \frac{0}{2}, \frac{7}{2}, \frac{0}{2}, \frac{3}{2}, \frac{5}{2}$$

$$\text{Class3} = (Term_5, Term_6) = \frac{0+2}{2}, \frac{1+2}{2}, \frac{3+0}{2}, \frac{0+0}{2}, \frac{1+4}{2}$$

$$= \frac{2}{2}, \frac{3}{2}, \frac{3}{2}, \frac{0}{2}, \frac{5}{2}$$

These centroids are calculated using the weights of the terms presented in the cluster for the items present in the database. The average of those weights gives the centroid.

Now, measure the similarity between the 8 terms and 3 centroids to assign similarity weights and assign terms to new classes which have the highest similarity. This assignment of new class is indicated in the row 'Assign' of the following table.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

	$Term_1$	$Term_2$	$Term_3$	$Term_4$	$Term_5$	$Term_6$	$Term_7$	$Term_8$
Class1	29/2	29/2	24/2	27/2	17/2	32/2	15/2	24/2
Class2	31/2	20/2	38/2	45/2	12/2	34/2	6/2	17/2
Class3	28/2	21/2	22/2	24/2	17/2	30/2	11/2	19/2
Assign	Class2	Class1	Class2	Class2	Class3	Class2	Class1	Class1

Table: Assigning Classes after an Iteration

The similarity between Term1 and Class1 is calculated using the weights in the above vector model for Term1 and the centroids calculated for Class1. It is shown below,

$$= (\frac{4}{2} \times 0) + (\frac{4}{2} \times 3) + (\frac{3}{2} \times 3) + (\frac{1}{2} \times 0) + (\frac{4}{2} \times 2)$$

$$= 0 + \frac{12}{2} + \frac{9}{2} + 0 + \frac{8}{2} = \frac{29}{2}$$

Similarly, all the calculations are performed and are placed in the above table.

As the Class2 contains the highest similarity for Term1 (i.e., $\frac{31}{2}$ is greater than 29/2 and 28/2), the Term1 is assigned to Class2. Similarly all the terms are assigned to those classes that have the highest similarity. The Term5 contains two classes that have the same similarity (i.e., Class1 and Class3 with similarity as 17/2). Class3 is assigned to Term5 as all the terms of Class1 have weights that are greater than 20.

Now, the new centroids are calculated as follows,

$$\text{Class1} = (Term_2, Term_7, Term_8)$$

$$= \frac{(4+1+3)}{3}, \frac{(1+0+1)}{3}, \frac{(0+3+0)}{3}, \frac{(1+2+0)}{3}, \frac{(2+0+2)}{3}$$

$$= \frac{8}{3}, \frac{2}{3}, \frac{3}{3}, \frac{3}{3}, \frac{4}{3}$$

$$\text{Similarly, Class2} = (Term_1, Term_3, Term_4, Term_6)$$

$$= \frac{2}{4}, \frac{12}{4}, \frac{3}{4}, \frac{3}{4}, \frac{11}{4}$$

$$\text{Class3} = (Term_5)$$

$$= \frac{0}{1}, \frac{1}{1}, \frac{3}{1}, \frac{0}{1}, \frac{1}{1}$$

Then, new cluster assignments are carried out using the above classes.

	$Term_1$	$Term_2$	$Term_3$	$Term_4$	$Term_5$	$Term_6$	$Term_7$	$Term_8$
Class1	23/3	45/3	16/3	27/3	15/3	36/3	23/3	34/3
Class2	67/4	45/4	70/4	78/4	33/4	17/4	17/4	40/4
Class3	12/1	3/1	6/1	6/1	72/1	9/1	9/1	3/1
Assign	Class2	Class1	Class2	Class2	Class3	Class2	Class3	Class1

Table: New Cluster Assignments

All the terms have been assigned the same classes as before, except the Term7. Because, it is more strongly related to Class3 instead of Class1.

Q39. Write about one pass assignment technique. Model Paper-II, Q7(b)

OR

Explain any one technique for term clustering with help of an example.

Answer : Nov./Dec.-12(R09), Q4(b)

One Pass Assignments

One pass assignment is a technique that assigns terms to classes only in one pass. Initially, the first term is assigned to a new class. Then, the additional terms are compared with the centroids of classes that are already existing and a threshold is selected. When an item is greater than threshold, it will be assigned to a class which is the most similar one. As the class is changed, a new centroid should be calculated. A term will be assigned to a new class if its similarity with all the existing centroids is less than the threshold. This process is said to be completed when all the items are assigned to classes and no item is left to be assigned to a class.

Example

Consider the following system which is a term-term matrix that considers a threshold of 10.

	$Term_1$	$Term_2$	$Term_3$	$Term_4$	$Term_5$	$Term_6$	$Term_7$	$Term_8$
$Term_1$		7	16	15	14	14	9	7
$Term_2$	7		8	12	3	18	6	17
$Term_3$	16	8		18	6	16	0	8
$Term_4$	15	12	18		6	18	6	9
$Term_5$	14	3	6	6		6	9	3
$Term_6$	14	18	16	18	6		2	16
$Term_7$	9	6	0	6	9	2		3
$Term_8$	7	17	8	9	3	16	3	

Table: Term-term Matrix

Using the above matrix and employing a threshold of 10, the following classes are generated

Class1 = $Term_1$, $Term_3$, $Term_4$

Class2 = $Term_2$, $Term_6$, $Term_8$

Class3 = $Term_5$

Class4 = $Term_7$

During this process of one-pass, the centroid values used are,

$$\text{Class1} = (Term_1, Term_3)$$

$$= \frac{(0+0)}{2}, \frac{(3+4)}{2}, \frac{(3+0)}{2}, \frac{(0+0)}{2}, \frac{(2+2)}{2}$$

$$= \frac{0}{2}, \frac{7}{2}, \frac{3}{2}, \frac{0}{2}, \frac{4}{2}$$

$$\text{Class1} = (Term_1, Term_3, Term_4)$$

$$= \frac{(0+0+0)}{3}, \frac{(3+4+3)}{3}, \frac{(3+0+0)}{3}, \frac{(0+0+3)}{3}, \frac{(2+2+3)}{3}$$

$$= 0, \frac{10}{3}, \frac{3}{3}, \frac{3}{3}, \frac{7}{3}$$

$$\text{Class2} = (Term_2, Term_6)$$

$$= \frac{4+2}{2}, \frac{1+2}{2}, \frac{0+0}{2}, \frac{1+0}{2}, \frac{2+4}{2}$$

$$= \frac{6}{2}, \frac{3}{2}, 0, \frac{1}{2}, \frac{6}{2}$$

The computation required to produce clustered classes is very less and is of the order $O(n)$. But the classes produced are not optimum. If the order of analyzing items is changed then the classes produced may also change.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

3.2.3 Item Clustering

Q40. Given the following vector model (matrix):

(a) Determine the Item-item matrix

(b) Determine the Item-relationship matrix.

(c) Determine the clusters using the Clique technique

(d) Determine the clusters using the single link technique

(e) Determine the clusters using the star techniques.

Answer :

	$Term_1$	$Term_2$	$Term_3$	$Term_4$	$Term_5$	$Term_6$	$Term_7$	$Term_8$
Item1	0	4	0	0	0	2	1	3
Item2	3	1	4	3	1	2	0	1
Item3	3	0	0	0	3	0	3	0
Item4	0	1	0	3	0	0	2	0
Item5	2	2	2	3	1	4	0	2

Table (1): Vector Model (Matrix)

(a) Determination of Item-item Matrix

Item-item matrix for the given vector model is determined by calculating the similarities between each item, using the following formula,

$$S(\text{Item } x, \text{Item } y) = \sum (\text{Term}_{xz})(\text{Term}_{yz})$$

Where, z can be any one of the terms.

Using this formula, similarity between Item1 and Item2 is,

$$S(\text{Item1}, \text{Item2}) = (\text{Term}_1, \text{Item2})(\text{Term}_1, \text{Item2}) + (\text{Term}_2, \text{Item1})(\text{Term}_2, \text{Item2}) + \ldots + (\text{Term}_8, \text{Item1})(\text{Term}_8, \text{Item2})$$

$$= (0 \times 3) + (4 \times 1) + (0 \times 4) + (0 \times 3) + (0 \times 1) + (2 \times 2) + (1 \times 0) + (3 \times 1)$$

$$= 0 + 4 + 0 + 0 + 0 + 4 + 0 + 3$$

$$= 11.$$

Similarly, the similarities between each of the five items is calculated and is placed in an $n \times n$ matrix called as item-item matrix. Here, 'n' corresponds to the total number of columns or items in the vector modes. This matrix does not have any value along its diagonal as it denotes the similarity of an item with itself. The item-item matrix for the given vector model is shown in table (2),

	Item1	Item2	Item3	Item4	Item5
Item1		11	3	6	22
Item2	11		12	10	36
Item3	3	12		6	9
Item4	6	10	6		11
Item5	22	36	9	11	

Table (2): Item-item Matrix

(b) Determination of Item-relationship Matrix

Item-relationship matrix for the above item-item matrix is given in table (3). It uses a threshold of 10.

	Item1	Item2	Item3	Item4	Item5
Item1		1	0	0	1
Item2	1		1	1	1
Item3	0	1		0	0
Item4	0	1	0		1
Item5	1	1	0	1	

Table (3): Item Relationship Matrix

As the threshold value is considered to be 10, the two items are said to be similar if they have a similarity value between them as 10 or greater than 10. The resulting matrix contain '1' if they have a similarity value greater than or equal to 10 and its contains '0' if they have a similarity value less than 10. This matrix is called as item-relationship matrix.

(c) Determination of Clusters Using Clique Technique

Using the above item-relationship matrix, determine those items that can share same class (i.e., those items that have threshold as '1' with respect to a given item).

Class1: I_1, I_2, I_5

Class2: I_2, I_3

Class3: I_1, I_2, I_3, I_4, I_5, (Not possible)

Class4: I_4, I_2, I_5

Class5: I_1, I_2, I_4, I_5 (Subset of Class3)

Thus, the possible classes are,

Class1 = I_1, I_2, I_5

Class2 = I_2, I_3

Class3 = I_2, I_4, I_5

(d) Determination of Clusters Using Single Link Technique

In this technique, if any of the items present in a class is similar to any other selected item then it can be added to the same class. It should be noted that no two classes have the same item.

Thus, using the above item-relationship matrix, the cluster formed is,

Class1 = I_1, I_2, I_5, I_3, I_4

As I_1, I_2, I_5 are items that are within the threshold they are selected. In addition to that, I_3 is similar to I_2 and I_4 is similar to I_2 and I_5, thus I_3 and I_4 are also placed in the same class.

(e) Determination of Clusters Using the Star Technique

Here, the item selected as new seed for the next star is not already part of a class.

To start with, Item1 is selected and is placed in Class1. I_1 is similar to I_2 and I_5 so they are placed in the same class, Class1.

$\boxed{\text{Class1}: I_1, I_2, I_5}$

Now I_3 is selected as a new seed because it is not included in Class1. As I_2 is the only item which is similar to I_3 it is placed in Class2 that holds I_3.

$\boxed{\text{Class2}: I_3, I_2}$

Then I_4 is selected as a new seed because it is not included in any of the above two classes. As, I_2 and I_5 are similar to I_4, they are placed in the same class, Class3, which holds I_4.

$\boxed{\text{Class3}: I_4, I_2, I_5}$

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

3.2.4 Hierarchy of Clusters

Q41. Explain about cluster hierarchy. **Model Paper-III, Q7(b)**

OR

Signify the role of dendrograms in identifying the alternate paths of growing the database.

(Refer Only Topic: Visualizing the Hierarchy of Clusters using Dendrograms)

Answer :

Hierarchy of Clusters

Hierarchy of clusters is an cluster analysis technique that groups similar objects/terms into a set of clusters. Here, each cluster is different from each other but the objects belonging to the clusters are similar. This technique groups the objects in the form of a tree.

Hierarchy of clusters are grouped into two types.

(i) **Hierarchical Agglomerative Clusters (HAC)**

This type of clustering begins with selection of unclustered items, performs pair-wise similarity measures on them inorder to identify the clusters.

(ii) **Hierarchical Divisive Clusters**

This type of clustering begins with a cluster and partitioning them into smaller clusters.

Objectives of Hierarchy of Clusters

1. Reduces the search overhead
2. Represents the clusters visually within the problem space.
3. Expand the relevant items that are retrieved.

Visualizing the Hierarchy of Clusters using Dendrograms

The dendrograms, together with visual indications on the cluster sizes and the linkage strength in between the clusters helps the user in identifying the alternate paths of browsing the database. The dendrograms also help the user in identifying those clusters that are still to be reviewed but are predicted to have essential items.

Graphical representation of dendogram can be given as,

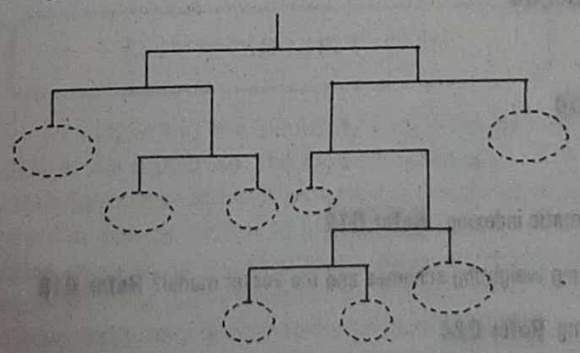


Figure: Graphical Representation of Dendrogram

The dashed ellipses are the clusters and lines connecting them are the paths that are traversed to reach the cluster.

Q42. Discuss the objectives for creating a hierarchy of clusters in information systems. Also list the characteristics of terms in building a concept hierarchy.

Answer :

The following are the objectives for creating a hierarchy of clusters in information systems.

1. Reducing the overhead of search process.
2. Providing information space's visual representation.
3. Enlarging the relevant items retrieved.

1. **Reducing the Overhead of Search Process**

This objective can be achieved by performing top-down searches of different centroids present in the hierarchy of clusters and removing all the irrelevant branches.

2. **Providing Information Space's Visual Representation**

For answer refer Unit-III, Q41.

3. **Enlarging the Relevant Items Retrieved**

This objective can be achieved by allowing the user to see additional items present in the cluster apart from the desired item. The specificity of items can be increased by allowing the user move through the children clusters or by increasing the generality of items that are under review there by moving on to its parent class.

Characteristics of Terms in Building a Concept Hierarchy

(i) Terms should demonstrate topics in the best way.

(ii) Parent terms corresponds to a more general concept compared to its child.

(iii) A child term corresponds to its associated parent's subtopic.

(iv) A directed acyclic graph describes the relationships in a pure hierarchy.

(v) Terms that have several meanings have separate entries for each meaning, in the hierarchy.

IMPORTANT QUESTIONS

SHORT QUESTIONS

1. List the steps of data flow in information process. **Refer Q2**
2. What is,
   (i) Word Frequency (WF)
   (ii) Total Frequency (TF)
   (iii) Document Frequency (DF). **Refer Q4**
3. What is clustering? **Refer Q7**
4. What is manual clustering? **Refer Q10**

ESSAY QUESTIONS

5. Discuss the different classes of automatic indexing. **Refer Q12**
6. What problems are encountered in using weighting schemes and the vector model? **Refer Q18**
7. Explain in detail about concept indexing. **Refer Q24**
8. Describe the concept of term clustering. Discuss in detial about about item clsutering. **Refer Q27**
9. Explain in detail about the manual clustering method. **Refer Q32**
10. Explain existing terms clustering with an example. **Refer Q38**
11. Explain about cluster hierarchy. **Refer Q41**

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

UNIT 4

User Search Techniques and Information Visualization

Look for the Logo on the Title Cover Before You Buy
SIA GROUP

Syllabus

User Search Techniques: Search Statements and Binding, Similarity Measures and Ranking, Relevance Feedback, Selective Dissemination of Information Search, Weighted Searches of Boolean Systems, Searching the INTERNET and Hypertext.

Information Visualization: Introduction to Information Visualization, Cognition and Perception, Information Visualization Technologies.

LEARNING OBJECTIVES

- Concept of Search Statement and Different Levels of Binding
- Different Similarity Measures and Ranking
- Various Terms Associated with Relevance Feedback
- Process of Dissemination of Information Search
- Problem and Solutions in Performing Weighted Searches of Boolean System
- Internet Based Search Capabilities
- Need of Information Visualization
- Cognition and Perception
- Various Technologies for Information Visualization.

INTRODUCTION

Searching is nothing but a way of estimating the similarity that exist between the user specified search statements and the items available in the database. The modern techniques of retrieval systems are based on weight concepts in which every individual item index is assigned a weighted value. The similarity measure can either be applied to complete item or can be limited to just logical fixed sized length passages in the item. In addition to these, similarity function can also be applied to variable length passages which are dependent on information similarity. This variable passage define locality based searching and also extract the information location that exactly matches the search query. If the similarity level of a passage is high, then it is used as a similarity measure. If the similarity measure is limited to passage then precision value is significantly increased with minimum effect on the recall value. When there are less number of terms, then it is difficult to locate shorter passages, which generally comprises of search terms generated from shorter length queries. Once all the items relevant to the user query are identified, then each item should be ranked so that the most relevant item is displayed first. The process of assigning rank based on relevance is referred to as ranking process. The result of similarity measure is a scalar number that exist between the item and the search query.

Information visualization conception is emerging for specifying the way the brain processes and uses mental images. In order to perform visualization, it is very essential to have advent in the field of technology to enhance information retrieval mechanisms. Doyle is one of the earliest researchers in Information visualization that defined the semantic road conception. These maps assisted the user with a view of the entire database. Another purpose of road maps is to display the interrelated items onto a particular theme of semantic. This view provided by road maps can be utilized by user so that he can concentrate on particular part of database semantic.

PART-A SHORT QUESTIONS WITH SOLUTIONS

Q1. Write short notes on search statement. List three levels of binding.

Answer : Model Paper-II, Q1(g)

Search Statement

Search statements are user generated statements wherein a weight is assigned to different concepts present in the statements. The main purpose of generating this information is that it needs statement to define the information that is being looked up in the documents (items). Basically, two methods i.e., Boolean logic or Natural language queries are used for generating search statements.

Levels of Binding

The three levels of binding are,

1. **First Level**

   This level of binding is done to specifying the vocabulary and pervious experience of the users.

2. **Second Level**

   This level of binding is done when passing of statement is completed.

3. **Third Level**

   This level of binding is done when search is applied to particular database.

Q2. What is similarity measure?

Answer : Model Paper-III, Q1(g)

Similarity that exist between the item and the search statement can be estimated by applying different kinds of similarity measures. One of the essential feature of these measures is that outcomes generated after applying the formula increases whenever the similarity between the item is high. The value of similarity formula is equal to zero if similarity does not exist between the items. "Sum of the products" is a kind of similarity measure that is applied in order to identify the items similarity for clustering. It is done using the formula,

$$\boxed{\text{SIM}(I_i, I_j) = \Sigma(T_i, k)(T_j, k)}$$

Q3. Write short note on Ranking algorithms.

Answer : (Model Paper-I, Q1(g) | Dec.-19(R16), Q1(g))

The two ranking algorithms are,

1. **Coarse Grain Ranking**

   This ranking process is dependent on the availability of query terms present in the items. It assigns rank to the item only if the word is available in the item.

2. **Fine Grain Ranking**

   In this ranking process, the estimation of the exact rank of an item is performed. It takes into consideration the physical location of the query terms as well as the location of the related words.

Q4. How would you define an item on internet with respect to a search statement and similarity function?

Answer : Dec.-19(R16), Q1(h)

A user defines an item on the internet with respect to search statement and similarity function in order to get the relevant information of the item.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Searching is nothing but a way of estimating the similarity that exist between the user specified search statements and the items available in the database. The modern techniques of retrieval systems are based on weight concepts in which every individual item index is assigned a weighted value. The similarity function (or) measure can either be applied to complete item or can be limited to just logical fixed sized length passages in the item. In addition to similarity function can also be applied to variable length passages which are dependent on information similarity. This variable passage define locality based searching and also extract the information location that exactly matches the search query. If the similarity level of a passage is high, then it is used as a similarity measure. If the similarity measure is limited to passage then precision value is significantly increased with minimum effect on the recall value. When there are less number of terms, then it is difficult to locate shorter passages, which generally comprises of search terms generated from shorter length queries. Once all the items relevant to the user query are identified, then each item should be ranked so that the most relevant item is displayed first. The process of assigning rank based on relevance is referred to as ranking process. The result of similarity measure is a scalar number that exist between the item and the search query.

Some of the similarity functions/measures that are most oftenly used are as follows,

(i) Jaccard similarity function

(ii) Dice similarity function.

Search statements are user generated statements where in a weight is assigned to different concepts present in the statements. The main purpose of generating this information is to define the information that is being looked up in the documents (items).

Q5. Why do commercial web search engines typically not provide relevance feedback functionality? Give at least 3 reasons.

Answer : **(Model Paper-I, Q1(h) | Dec.-14(R09), Q5(b))**

Commercial web search engines typically not provide relevance feedback functionality because of following reasons.

1. Relevance feedback is difficult for most of the users to understand and use.
2. Relevance feedback needs search session inorder to retrieve a query.
3. Relevance feedback is very expensive as it gets executed on a typical IR engine. This result in serving the queries in a very high amount.

Q6. What is the main advantage of the cosine similarity measure?

Answer : **Dec.-14(R09), Q5(c)**

The advantages of cosine similarity measures are as follows,

(i) It directly manipulates the raw data.

(ii) It answers the queries in an efficient manner.

(iii) It is easy to implement.

(iv) It can implement density based and standard distance based algorithms in a better manner.

(v) It can possibly implement partial matching as it returns the values in the range of 0 to 1.

(vi) It can rank in accordance to the retrieved results of similarity scores.

Q7. What is Dissemination system?

Answer : **Model Paper-III, Q1(h)**

Dissemination systems are the most frequent term used for referring selective dissemination of information. The main purpose of this process is to perform dynamic comparison between items recently received and the existing item statements, which are desired by users. Because of emerging growth of internet technology, the dissemination systems are becoming more prevalent.

Q8. Define information visualization.

Answer : **Model Paper-II, Q1(h)**

Information visualization conception is emerging for specifying the way the brain processes and uses mental images. In order to perform visualization, it is very essential to have advent in the field of technology to enhance information retrieval mechanisms. Doyle is one of the earliest researchers in information visualization that defined the semantic road conception. These maps assisted the user with a view of the entire database. Another purpose of road maps is to display the interrelated items onto a particular theme of semantic. This view provided by road maps can be utilized by user so that he can concentrate on particular part of database semantic.

PART-B ESSAY QUESTIONS WITH SOLUTIONS

4.1 USER SEARCH TECHNIQUES

4.1.1 Search Statements and Binding

Q9. What are search statements and why there are three levels of binding in the creation of a search?

Answer : Model Paper-II, Q8(a)

Search Statements

Search statements are user generated statements wherein a weight is assigned to different concepts present in the statements. The main purpose of generating this information is that it needs statement to define the information that is being looked up in the documents (items). Basically, two methods i.e., Boolean logic or Natural language queries are used for generating search statements.

Three Levels of Binding

First Level

At this stage, binding is done to specify the vocabulary and previous experience of the users. Binding is defined when more generalized form of statements are redefined to more specialized form. The search statements are basically a way of subsetting the entire item space in order to form item clusters that comprises of information required by users.

Second Level

This level of binding is done when parsing of search statement is completed. The parsing is done so that the search statements are utilized by a particular search system. This system is responsible for converting the query into a meta language which is easily understood by the system. Some of the searches include,

(i) Statistical Systems

These systems identify the desired processing tokens and the weights assigned to them depending on their occurrence frequency.

(ii) Natural Language Systems

These system identifies the syntactical and semantics usage of the terms. It does this by employing algorithm like index phrase generation.

(iii) Concept Systems

These systems are responsible for associating the search statements with the set of concepts which are used to index particular items in that statement.

In order to specify the expansion by a thesaurus, parentheses are used in this level of binding.

Third Level

This level of binding is done when the search is applied to a particular database. This type of binding is dependent on the processing token statistics and even on the semantics used in the database. The statistics measure applied relies on the existing information present in the database. Some of the statistical technique are document frequency, total frequency. In concept indexing system, statistical algorithms are applied for identifying the concepts used as basis for indexing. This algorithm is applied against a representative database sample. Natural language system apply corpora independent algorithms for searching the items.

The capability of information retrieval system to locate relevant item is directly proportional to the length of search statement i.e., shorter the length of search statement, the more difficult it is to find the relevant items. The search statement used in selective dissemination process are of longer length comprising of 75 to 100 terms, whereas adhoc search statements consist of 7 terms. The adhoc search statement decreases the effectiveness of many searching techniques.

Example

Search Statement Input	Binding
Find the information on the significance of petrol reserves in Saudi Arabia on the price of petrol.	User search statement by applying user's vocabulary.
Significance, petrol, reserves, Saudi Arabia, price.	Statistical type of system binding and retrieval of processing tokens.
Significance (0.406), petrol (0.702), reserves (0.14), Saudi Arabia (0.54), price (0.18).	Assigning weights as per inverse document frequency algorithm and databases.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

4.1.2 Similarity Measures and Ranking

Q10. Explain about similarity measures and ranking. (Nov.-13(R09), Q5(a) | May/June-13(R09), Q5(a))

OR

'A variety of different similarity measures can be used to calculate the similarity between the item and the search statement'. Justify your answer.

(Refer Only Topic: Similarity Measures)

OR

Identify various similarity measures in IRS.

(Refer Only Topic: Similarity Measures)

Answer : (Model Paper-III, Q8(a) | Oct./Nov.-20(R16), Q7(a))

Similarity Measures

Searching is nothing but a way of estimating the similarity that exist between the user specified search statements and the items available in the database. The modern techniques of retrieval systems are based on weight concepts in which every individual item index is assigned a weighted value. The similarity measure can either be applied to complete item or can be limited to just logical fixed sized length passages in the item. In addition to these, similarity function can also be applied to variable length passages which are dependent on information similarity. This variable passage define locality based searching and also extract the information location that exactly matches the search query. If the similarity level of a passage is high, then it is used as a similarity measure. If the similarity measure is limited to passage then precision value is significantly increased with minimum effect on the recall value. When there are less number of terms, then it is difficult to locate shorter passages, which generally comprises of search terms generated from shorter length queries. Once all the items relevant to the user query are identified, then each item should be ranked so that the most relevant item is displayed first. The process of assigning rank based on relevance is referred to as ranking process. The result of similarity measure is a scalar number that exist between the item and the search query.

Similarity that exist between the item and the search statement can be estimated by applying different kinds of similarity measures. One of the essential feature of these measures is that outcomes generated after applying the formula increases whenever the similarity between the item is high. The value of similarity formula is equal to zero if similarity does not exist between the items. "Sum of the products" is a kind of similarity measure that is applied in order to identify the items similarity for clustering. It is done using the formula,

$$\boxed{\mathrm{SIM}(I_i, I_j) = \Sigma(T_i, k)(T_j, k)}$$

In this formula, sum of the product of terms (T) of two items (I) is calculated, whenever an index is considered as a vector. The same formula can also be used for estimating the similarity between the "item" and a "query". This is done by replacing I_j with Que_j. While determining the similarity between item and query, the difficulty comes in the normalization process. This process considers the variances in the length of items. In order to make the final result to range between 0 and +1, additional normalization is performed.

Robertson and Spark Jones developed a theory for similarity function. According to this "The weight of the terms within the relevant items which are extracted from a query should be adjusted in accordance with the terms knowledge in the weighting process". This theory makes use of comparison between,

1. Number of relevant items and number of non-relevant items.
2. Number of relevant items consisting of particular query term and number of non-relevant items consisting of term used to derive four weighting formula.

Later, Croft expanded this theory by considering the frequency factor of term present in an item. He generated a similarity formula which is,

$$\boxed{\mathrm{SIM}(I_i, Que_j) = \sum_{i=1}^{k}((\alpha + IDF_i) * freq_{i,j})}$$

In the above formula, 'I' represents the item, 'Que' represents the query, 'α' is a constant which is used in tuning, IDF_i is the inverse document frequency for term 'i', and '$freq_{i,j}$' is the frequency which is calculated as,

$$\boxed{freq_{i,j} = \frac{\beta t + (\beta t - 1)\, Tfreq_{i,j}}{\max freq_j}}$$

Where βt is the tuning constant which lies between 0.3 and 0.5 (i.e., $0.3 < \beta t < 0.5$).

$Tfreq_{i,j}$ is the $term_i$ frequency in $item_j$.

$maxfreq_j$ is the maximum frequency of term in $item_j$.

In addition to the above similarity formulas, another formula was used by Salton in "SMART" system. This formula is referred to as "cosine formula", that estimates the cosine angle between item and query vectors. Salton considered the index and search query as n-dimensional vectors. In order to identify the weight assigned to an item in accordance to the search statement, cosine formula is applied for estimating the distance between the item's vector and query vector.

Cosine formula is defined as,

$$\mathrm{SIM}(I_i, Que_j) = \frac{\sum_{k=1}^{n}(I_{i,k} * Que_{j,k})}{\sqrt{\sum_{k=1}^{n}(I_i, k)^2 * \sum_{k=1}^{n}(Que_{j,k})^2}}$$

Where I_{ik} specifies the kth term presented in the weighted vector for item 'i'.

Que_{jk} specifies the kth term present in $Query_j$ when the cosine angle = 1, then it means that the item and query vector are coincident. On the other hand, when the cosine angle = 0, it means that both these vectors are dissimilar and orthogonal. Cosine formula does not consider the vector length while calculating the similarity between the vectors.

Cosine formula was later enhanced by Salton and Buckley. They improved the formula by changing the term factors in the query as,

$$Que_{i,k} = \left(0.5 + \left(\frac{0.5Tfreq_{i,k}}{\max freq_k}\right)\right) IDF_i$$

Where $Tfreq_{ik}$ specifies the term frequency of 'i' in query 'k' and $maxfreq_k$ specifies the maximum frequency of term in query 'k'. IDF_i is inverse document frequency for term 'i'.

Various Similarity Measures

There are two other similarity measures that are most oftenly used.

(i) Jaccard similarity measure

(ii) Dice similarity measure.

In both of these similarity measures, the normalizing factor in the denominator is changed to specify distinguishable data features. The cosine formula does not varies with the number of terms in common and generate less number of similar term, whenever the size of vector is large and when the number of common items is small.

(i) Jaccard Similarity Measure

It is defined as,

$$\boxed{\mathrm{SIM}(I_i, Que_j) = \frac{\sum_{k=1}^{n}(I_{i,k} * Que_{j,k})}{\sum_{k=1}^{n} I_{i,k} + \sum_{k=1}^{n} Que_{j,k} - \sum_{k=1}^{n}(I_{i,k} * Que_{j,k})}}$$

In contrast to cosine similarity measure, the denominator in Jaccard similarity measure is dependent on the number of similar terms i.e., when the number of similar term decreases, the similarity value increases rapidly. The similarity value is in the range –1 to +1.

(ii) Dice Similarity Measure

It is defined as,

$$\boxed{\mathrm{SIM}(I_i, Que_j) = \frac{2 * \sum_{k=1}^{n}(I_{i,k} * Que_{j,k})}{\sum_{k=1}^{n} I_{i,k} + \sum_{k=1}^{n} Que_{j,k}}}$$

Dice similarity measure reduces the complexity of denominator in Jaccard measure. Factor of 2 is added in the numerator. It is similar to cosine measure in the sense that it does not vary with the number of similar terms.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Similarity measures is not only applied to statistical systems but also to natural language system. While implementing this system, weighted values are generated from statistical data together with the natural language processing which is generally stored in the form of indexes (i.e., stored indexes). The similarity measure in natural language system is applied in the similar way as applied to statistical system. But natural language system requires constructs at discourse level in order to carry out additional item filtering apart from similarity measure.

The entire database is returned as a search result when similarity algorithm is applied.

The similarity value for item is either close or equal to zero. Because of this, threshold values are generally assigned to every search process. These values identifies the items available in the resultant hit files generated from the query. In general, threshold refers to,

1. A value that is made equal or greater by the similarity measure.
2. A number that imposes the restrictions on number of items present in the hit file.

When the similarity value is equal to zero, then that case is considered as default.

Example

Let us consider an example to illustrate how the threshold process takes place.

DOC1(0, 0, 1, 2, 3, 0)

DOC2(1, 0, 3, 2, 0, 0, 3)

DOC3(1, 3, 3, 0, 0, 3)

Query(1, 0, 0, 1, 0, 1)

Sum of products similarity measure is used to calculate the similarity between document and query.

SIM(Q, DOC1) = (0 + 0 + 0 + 2 + 0 + 0) = 2

SIM(Q, DOC2) = (1 + 0 + 0 + 2 + 0 + 3) = 6

SIM(Q, DOC3) = (1 + 0 + 0 + 0 + 0 + 3) = 4

If the threshold value is not specified, then all the three documents are treated as hits. If threshold is equal to 5, then DOC2 is the only document returned.

Item Cluster Hierarchy

Searching of terms that are stored in clusters is an important concern in item clustering. These clusters are stored in hierarchical format which are individually represented by the centroid.

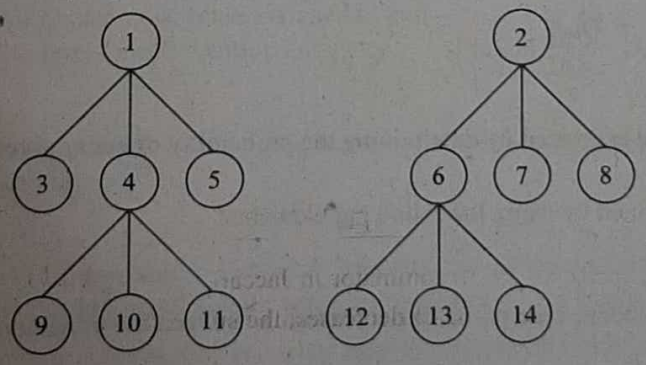


Figure: Hierarchy of Item Cluster

In the above hierarchies, the nodes at the lower level represent an item (3, 9, 10, 11, 5, 12, 13, 14, 7, 8) and the nodes at higher level (1, 4, 2, 6) represents the centroid of the children nodes following them. Top-down approach is followed for performing the search within the hierarchy.

Steps followed for performing search process is,

1. Initially, the query received is compared with the centroid 1 and 2.
2. If the similarity value is greater than threshold values then the query is applied to the node's children otherwise, the respective part of the hierarchy is pruned and searching is not performed on that specific part.
3. The search process is done until the comparison of all the leaf nodes that are not pruned is completed.

The following are the problems that arises because of the centroid, which is the mean of an item collection.

(a) If a continued search process is performed on a query, the resultant mean is not same. However, some specific terms involved in determining the centroid may be approximately same in order to satisfy the search process.

(b) The risks associated with missing items and an increase in the reduced recall increases the standard deviation.

(c) A decrease in similarity computation and recall is observed when centroids are used. However, use of centroids will not cause any effect on precision because it is dependent on the similarity calculations that take place at the leaf level.

Clustering is considered as another similarity measure to similarity search ranking by Hearst and Pedersen. In this, the clustering is applied to complete database. The advantage of clustering is that it gives complete information about the content as well as the structure of database. In contrast, to standard similarity query measure, clustering approach is ineffective as it is restricted to perform the search on the hierarchy that extract less relevant items.

Ranking

For answer refer Unit-IV, Q12.

Q11. Explain the application of Hidden-Markov models in searching.

Answer : **Model Paper-II, Q8(b)**

Hidden-Markov Models

Hidden-Markov Model (HMM) have been marked as a new model for searching textual corpora. In contrast to other searching methods, HMM treats documents as unknown statistical processes. These processes produces the result which is similar to the set of queries. HMM can also be defined by considering the observed result (query) generated by traversing some unknown key (relevant document). This traversing is done using state transition through noisy channel.

HMM method is initiated by applying Bayesian formula to conditional probability.

$$P(\text{DOC is Rel/Que}) = P\left(\frac{Que}{DOC} \text{ is Rel}\right) * \frac{P(DOC \text{ is Rel})}{P(Que)}$$

P(Que) is ignored, if analysis is performed from document point of view.

P(DOC is Rel) is not possible if the textual corpora is of larger size. It does not enhance the query resolution as relevant documents are sensitive to particular queries.

$$\therefore \quad P(\text{DOC is Rel/Que}) = P\left(\frac{Que}{DOC} \text{ is Rel}\right)$$

i.e., $P\left(DOC \text{ is } \frac{Rel}{Que}\right)$ can be estimated by determining the probability of query given document is relevant.

Hidden-Markov model is defined by using the following elements,

1. A set of transition states.
2. A transition matrix that defines the probability of moving between the transition states.
3. A set of resultant symbol.
4. Probability of resultant document symbol given a specific transition state.

Resultant document symbol represents set of all possible queries and the transition state is defined by document file. Therefore, HMM passes through the document state that results in transition. This transition leads to query term related to the new state. State transitions defines the way, words are merged to form documents. Whenever, a query is specified, then the probability of determining the query generation from a document is estimated.

The major drawback of HMM method is the calculation of transition matrix and output for all the documents in the database. Estimation-maximization algorithms can be applied to solve the problem, only if more number of queries database and relevant documents are associated with queries and document present in adequate coverage.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Q12. Explain in detail about the various ranking algorithms used for ranking of output.

Answer :

Ranking Algorithm

Ranking is the process of assigning ranks to the most relevant items. The item which is more relevant is given the highest rank and the one which is least relevant is given the lowest rank. The advantage of ranking process is that the user's overhead in identifying the relevant item is reduced. The user can retrieve the most relevant item based on the rank assigned to it. Ranking systems are very much different from traditional Boolean systems, because the later retrieve the relevant items which are ordered based on their entry date. In many of the modern systems, ranking has become an important feature because of the presence of statistical similarity methods. The other reason for applying ranking approach is that more number of hits which are produced from searching diverse database are generated.

Many of the business-oriented systems make use of heuristic rules. These rules support the system, while ranks are assigned to the relevant items. The usage of factors that require knowledge about the database is not desired by systems for performing ranking and similarity function.

This is because, it is not easy for any system to maintain the current values each time database is modified. RetrievalWare system is one of the best examples of business-oriented system that consolidates efficiency with theoretical concepts. In order to rank the relevant items, Retrieval Ware system initially index every item so as to identify the most relevant items. Later it applies coarse grain ranking and fine grain ranking approaches.

Ranking Algorithms

1. **Coarse Grain Ranking**

This ranking process is dependent on the availability of query terms present in the items. It assigns rank to the item only if the word is available in the item. It can be treated as a weighted formula that is modified based on the following factors,

(i) Completeness

(ii) Contextual evidence and

(iii) Semantic distance.

(i) **Completeness**

Completeness can be defined as,

$$\text{Completeness} = \frac{\text{Number of query terms in an item}}{\text{Number in the query}}$$

This factor sets an upper limit for the rank value assigned to every item. The weight values are factorized into the value only if weights are assigned to terms present within the query.

(ii) **Contextual Evidence**

When the related words that are generated from semantic networks are also present in the item, then there is a possibility of contextual evidence to take place.

(iii) **Semantic Distance**

This factor specifies the level of association that exist between the newly added words and the original query term. Synonymous words increases the additional weight whereas, the antonymous words decreases the weight.

The advantages of coarse grain process is that, it is very easy to estimate the ranking value as physical proximity is not being considered.

2. **Fine Grain Ranking**

In this ranking process, the estimation of the exact rank of an item is performed. It takes into consideration the physical location of the query terms as well as the location of the related words.

It considers these locations by using the following factors,

(i) Proximity

(ii) Completeness

(iii) Contextual evidence

(iv) Semantic distance.

Proximity

The item is said to be relevant, if their exist a close proximity between the query terms and related words. This level of proximity increases at adjacency and decreases as the physical distance increases.

The fine grain rank can be zero, irrespective of the existence of query terms, if the query terms are distributed throughout the longer length item. This ranking algorithm are not preferred by many users for assigning the rank value whenever a hit lists are displayed. This is because, there is a chance that the ranking score created by ranking algorithm can mislead and confuse the user.

Q13. Consider a very small collection C that consists of the following three documents,

d_1 : 'scary green crocodile'

d_2 : 'scary green big'

d_3 : 'small crocodile'

Find a vector representation for the documents. State which two documents are most similar based on the cosine coefficient.

Answer : Nov./Dec.-12(R09), Q5(b)

$$\text{Cosine formula} = \frac{\sum_{k=1}^{n}(DOC_{i,k} * DOC_{j,k})}{\sqrt{\sum_{k=1}^{n}(DOC_{i,k})^2 * \sum^{n}(DOC_{j,k})^2}}$$

	Scary	Green	Crocodile	Big	Small
d_1	1	1	1	0	0
d_2	1	1	0	1	0
d_3	0	0	1	0	1

$$d_1.d_2 = \frac{1.1+1.1+1.0+0.1+0.0}{\sqrt{1^2+1^2+1^2}.\sqrt{1^2+1^2+1^2}}$$

$$= \frac{1+1}{\sqrt{3}.\sqrt{3}}$$

$$= \frac{2}{3}$$

$$= 0.66$$

$$d_2.d_3 = \frac{1.0+1.0+0.1+1.0+0.1}{\sqrt{1^2+1^2+1^1}.\sqrt{1^2+1^2}}$$

$$= 0$$

$$d_1.d_3 = \frac{1.0+1.0+1.1+0.0+0.1}{\sqrt{1^2+1^2+1^2}.\sqrt{1^2+1^2}}$$

$$= \frac{1}{\sqrt{3}.\sqrt{2}}$$

$$= \frac{1}{(1.732)(1.414)}$$

$$= \frac{1}{2.46}$$

$$= 0.40$$

Documents d_1 d_2, d_1 d_3 are most similar based on cosine coefficient.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

4.1.3. Relevance Feedback

Q14. What is relevance feedback? Explain the Rochio method for relevance feedback.

(Model Paper-II, Q9(a) | Nov./Dec.-12(R09), Q5(a))

Answer :

Relevance Feedback

The system based on users input retrieve a set of documents. Users examine these documents and make decisions on the relevance of individual document. Decisions are about whether a document is relevant to query or not. Based on the decision user alters the query to retrieve more documents. The system again retrieves documents based on the user query and a new cycle of user examination begins.

This process of user query, system retrieval, document examination and query alteration is known as relevance feedback.

Rochio Method for Relevance Feedback

For answer refer Unit-IV, Q15, Topic: Rochio Method for Relevance Feedback.

Q15. Discuss in detail about relevance feedback and explain the impact of positive and negative feedback on retrieval strategy.

Dec.-19(R16), Q9

OR

What are some potential ambiguities in use of relevance feedback on hypertext documents? Is the use of positive feedback always better than using a negative feedback to improve a query?

Answer :

Relevance Feedback

For answer refer Unit-IV, Q14, Topic: Relevance Feedback

Potential Ambiguities in Use of Relevance Feedback

The difficulty in locating relevant item is because of the difference between the author's vocabulary and the user's vocabulary. In order to support related search terms many thesauri and semantic networks are used. These networks do this by expanding the search statement entered by the users. By using these networks, the problem still remains, since it is not possible to link the vocabulary used by different authors. Another problem that arises using the thesaurus network is that, it may not include the modern idioms, acronyms. But the advantage of using thesaurus network in interactive system is that, it provides the user with the capability of manually modifying the query which is inefficient or expanding the query automatically. Once the query is processed the user can utilize the relevant item generated by the system. The users can use these items so as to enhance the searches made in the future. Thus using relevant item without considering their ranking order is the major idea used in relevance feedback. The main purpose of using relevant item is to assign the weights again to the existing query terms or to expand the search statement by using new terms.

Rochio Method for Relevance Feedback

In relevance feedback, processing the new query must rely on the old query which is modified. This is done to increase and decrease the value of weights assigned to the terms in relevant and non-relevant items respectively. Using this method, it is possible to modify the original query terms and also to expand new terms generated from relevant item. Relevance feedback is estimated using the formula,

$$\boxed{Que_k = Que_0 + \frac{1}{rel}\sum_{i=1}^{rel} DRel_i - \frac{1}{nrel}\sum_{j=1}^{nrel} DNrel_j}$$

Where 'Que_k' represents the modified vector for the new query; 'Que_0' represents the original query; '*rel*' represents the total number of relevant items; '$DRel_i$' represents the vector used for relevant items; '*nrel*' represents the total number of non-relevant items and '$Dnrel_j$' represents the vector used for non-relevant items.

The modified form of the above formula is,

$$\boxed{Que_k = \alpha Que_0 + \beta\sum_{i=1}^{rel} DRel_i - \gamma\sum_{j=1}^{nrel} DNrel_j}$$

Where, α, β, γ represents the constants related to *rel*, *Nrel* factors.

The term $\beta\sum_{i=1}^{rel} DRel_i$ is termed as positive feedback and $\gamma\sum_{j=1}^{nrel} DNrel_j$ is termed as negative feedback.

Positive Feedback

Positive feedback increases the term value in the query vector. It provides more relevant items which are very much closer to user information requirement, when the query is specified. This type of feedback is responsible for moving the query in order to extract items which are similar to the extracted items. Positive feedback moves closer in the direction of relevant items. This feedback generates significant values in order to produce accurate queries.

Negative Feedback

Negative feedback decreases the term value in the query vector. This type of feedback is useful in some way, but it decreases the effectiveness of a query. In contrast to positive feedback, negative feedback moves in the direction of non-relevant items but does not move nearer to the relevant items.

The major problem that arises while applying relevance feedback algorithms is "what should be done to query terms which are not present in the extracted relevant items?" when these algorithms are applied then the weight of the retrieved relevant items is decreased. The reduction of relative weight is done in accordance to the other terms in the query. However, this is not preferred by user, as they assume that every individual term may be important and may be used in order to improve the future searches.

Most of the information retrieval system comprises of relevance feedback as an important characteristics. When this technique is used for modifying original query, it becomes essential for system to verify if the original terms, of query are present in the modified query. The modified query is given to the user, so that they can readjust the value of weight assigned and check the query again to identify the terms that are newly added. The modification increases the value of similarity measure for relevant item and decrease it for the non-relevant item. If the query specified consist of fewer terms, then there are more chances that the non-relevant terms will be extracted. New terms are added to original query so as to expand the query and to add the term to the user's vocabulary.

TREC conference performed a research on the relevance feedback approach and concluded it by laying out the following advantages,

1. When compared to the results generated manually, the results produced by processing queries using relevance feedback approach are more accurate.
2. Automatic relevance feedback that is dependent on ranking order of items is used, when the user does not specify the query completely. This is referred to as pseudo-relevance feedback or blind feedback as it doesn't require any interference of the judgement a human makes about relevant item. The item which is assigned the highest rank is considered to be the most relevance item. When positive feedback approach is applied to such relevant items, then it is possible to create and execute the query that is expanded by adding new terms. Based on this expanded query, hit file is returned to the user by the system. Blind feedback approach enhances the system performance even though automatic feedback approach is not used.

Q16. Suppose that a user's initial query is "cheap CDs cheap DVDs extremely cheap CDs". The user examines two documents d_1 and d_2. She judges d_1 with the content DCs cheap software cheap CDs relevant and d_2 with content cheap thrills DVDs non relevant. Assume that we are using direct term frequency (with no scaling and no document frequency). There is no need to length-normalize vectors. Using Rocchio relevance feedback what would the revised query vector be after relevance feedback? Assume $\alpha = 1$, $\beta = 0.75$, $\gamma = 0.25$.

Answer : Dec.-14(R09), Q5(a)

Given that,

Initial query (q) = "Cheap CDs cheap DVDs extremely cheap CDs"

d_1 = "CDs software cheap CDs"

d_2 = "Cheap thrills DVDS"

$\alpha = 1,\ \beta = 0.8, \gamma = 0.2$

Rocchio relevance feedback is calculated with the help of following equation.

$$\bar{q}_m = \propto \bar{q}_0 + \beta \frac{1}{|D_r|} \sum_{\bar{d}_j \in D_r} \bar{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\bar{d}_j \in D_{nr}} \bar{d}_j$$

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

The table below shows the occurrence of words like CDS, DVDS, cheap, extremely, software and thrills in initial query, document d_1, and document d_2.

word	q	d_1	d_2	$\alpha q\ \alpha = 1$	$\beta d_1\ \beta = 0.8$	$\gamma d_2 \gamma = 0.2$	$\bar{q}_m$
CDs	2	2	0	$2 \times 1 = 2$	$0.8 \times 2 = 1.6$	$0.2 \times 0 = 0$	$\alpha q + \beta d_1 + \gamma d_2$ $2 + 1.6 - 0 = 3.6$
DVDs	1	0	1	$1 \times 1 = 1$	$0.8 \times 0 = 0$	$0.2 \times 1 = 0.2$	$1 + 0 - 0.2 = 0.8$
Cheap	3	1	1	$3 \times 1 = 3$	$0.8 \times 1 = 0.8$	$0.2 \times 1 = 0.2$	$3 + 0.8 - 0.2 = 3.6$
Extremely	1	0	0	$1 \times 1 = 1$	$0.8 \times 0 = 0$	$0.2 \times 0 = 0$	$1 + 0 - 0 = 1$
Software	0	1	0	$1 \times 0 = 0$	$0.8 \times 1 = 0.8$	$0.2 \times 0 = 0$	$0 + 0.8 - 0 = 0.8$
Thrills	0	0	1	$0 \times 0 = 0$	$0.8 \times 0 = 0$	$0.2 \times 1 = 0.2$	$0 + 0 - 0.2 = 0$ ($\because$ Negative number is not allowed)

4.1.4 Selective Dissemination of Information Search

Q17. Explain in detail about dissemination systems.

Answer :

Dissemination Systems

Dissemination systems are commonly used to refer selective dissemination of information. The main purpose of this process is to perform dynamic comparison between items recently received and the existing item statements, which are desired by users. Because of emerging growth of internet technology, the dissemination systems are becoming more prevalent. Dissemination systems are different from the search systems because of the following reasons,

(i) Dissemination systems are referred to as push systems whereas, search systems are referred to as pull systems.

(ii) In search system, the user himself is responsible for deciding what information is required and passes this query to the information search system that initiate the search process after receiving the query. On the other hand, in dissemination systems the profiles are created by the user, which are automatically compared with the newly added information. If the profile matches with the newly added information then the matching information is sent to a mail file of the requested user in an asynchronous manner.

The query and profile search statements can be linked to each other by considering the time factor which is related to both of the search statements. The search statement is said to be active and disseminating only if the time factor is in the future. On the other hand, if the timer value expires, then the information need specified by the user does not exist. It is returned to the user only if a demand request is issued. The two functions differs in the following aspects of the profiling process.

(i) Dynamic nature of profiling process.

(ii) Size and diversity of search statements.

(iii) Number of searches performed parallely for each item.

Dissemination system doesn't have database associated with it but on the other hand an existing database is available for search system. Because of which, the algorithm followed in dissemination system does not depend on past data or develop a method for calculating terms for their respective formula. Dissemination system is also referred to as binary classification system because of the absence of user feedback, which is required while referring the search statements. It is upto the system to decide whether to accept the item and file it or reject the item.

One of the search statements that include variety of topics are profiles, that generally generalize the necessary user requirement. They concentrate on the present requirement of the user, due to which profiles tend to contain more number of terms when compared to adhoc query. However, when the size of the profiles increases it becomes very difficult for a user with less knowledge to perform modifications.

The following are the examples of system where dissemination of information search is employed.

1. Logicon Message Dissemination System (LMDS).
2. Personal Library Software System (PLSS).
3. Statistical Classification Approach (SCA).

1. Logicon Message Dissemination System (LMDS)

This dissemination system is one of the earliest business-oriented search methods. It was basically designed to search the numerous profiles at higher speed. It solved the problem that arises while treating the profiles as state database and while representing the newly added terms as query. In order to search the profile LMDS utilize the terms present in the item. It does this so as to lookup for the profiles whose logic matches with that of item. Trigraph algorithm is used by the system for identifying those profiles whose logic does not match with the item. The profiles logic which are satisfied are evaluated in-depth so as to ensure that the item is a hit.

2. Personal Library Software System (PLSS)

Personal library software is another system which uses dissemination approach for gathering items that are recently received in the database. After receiving it, execution of user's profile in accordance to the database is carried out on periodic basis. The advantage of using this approach is that retrospective search software is utilized to maximum extent but at the same time, real-time delivery of items is lost. RetrievalWare and InRoute software systems follows the similar approach. In both of these systems, the processing of items is carried out in a searchable form. Because of the static nature, profiles are generally used in determining the terms that are used in every profile. Words present in the items and that belongs to the list are not considered while estimating the similarity value. Therefore, such words are removed from the search structure. After the elimination, the resultant items are compared with each individual profile. Though, RetrievalWare applies statistical algorithm, it does not include any data associated with the current database. If database is not present then it does not have any impact on its similarity measure. The route system like INQUERY system makes use of information related to Inverse Document Frequency (IDF). This information is created for storing and altering information so that they can be used whenever new items arrive in future.

Due to these frequency alterations, the weights of inverse document frequency cannot be stabilized until sufficient number of items have arrived to perform stabilization. For improving the searching capabilities of adhoc queries, relevance feedback approach is used. This approach can also be applied to selective dissemination of information search system. In contrast to adhoc query, the dissemination system approach is a continuous process but in this system it is not possible to store all the past relevance results that are to be used in relevance feedback process.

3. Statistical Classification Approach (SCA)

Dissemination process is used in the systems that uses statistical classification and error minimization methods. These methods are used to identify the decision standard using which selection of items are made for a specific profile. Classification process is used for assigning every individual item to one of the following two classes,

(i) Relevant to user's profile class

(ii) Non-relevant to user's profile class.

The problem of using error minimization approach is that it faces hurdles in the high-dimensionality spaces, which specifies the total number of distinct terms. In these terms, every individual term is treated as distinct dimension. This problem arises because of numerous dimensions used for a realistic training set in order to create parameters associated with error minimization.

Latent Semantic Indexing (LSI) is used to decrease the dimensionality spaces. The prerequisite for initiating this indexing technique is the requirement of training data set together with the associated profile.

The two methods used by Schutze for reducing the dimensionality spaces are,

1. Selection of existing concept vectors set.
2. Creation of new set of vectors of smaller size. These vectors are those in which the original vectors are mapped.

For identifying the essential characteristics a measure referred to as chi-square(χ^2) is used. The items are arranged in a table format. This table specifies the total number of relevant (T_{rel}) and non-relevant (T_{nrel}) items in which a term exist and the total number of relevant items (T_{rel^-}) and non-relevant items (T_{nrel^-}) in which the term does not exist. χ^2 is calculated using the following formula,

$$\chi^2 = \frac{T(T_{rel}T_{nrel^-} - T_{rel^-}T_{nrel})^2}{(T_{rel}+T_{rel^-})(T_{nrel}+T_{nrel^-})(T_{rel}+T_{nrel})(T_{rel^-}+T_{nrel^-})}$$

The items which are determined by profiles are analyzed in detail when compared to the occurrence frequency of terms, chi-square approach is more effective. When the result of χ^2 is high, then the vector specifies whether the item is retrieved or not. The frequency of this vector set depends on the occurrence of the items.

In addition to LSI method, another method which is used for identifying the new reduced concept vector set is a enhanced version of LSI. The modified LSI is different from conventional LSI technique because the former creates distinct ways of representing every item or term within a profile, thereby generating local item space for each item.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Once the analysis is completed, the result is passed to the learning algorithm which is related to the corresponding classification technique. It is very important to define the local region so as to make use of the profile when working with databases of larger sizes. In situation wherein profiles are not defined, then it is not possible to process numerous LSI factors. Instead of storing the LSI factors separately for individual profile it is beneficial to use another method, wherein the result generated from all queries are combined into a single LSI analysis. The advantage of this method is that the number of factors are increased which also increases the processing complexity.

The learning algorithm can be used to perform the classification process, when the identification of reduced concept vector set is completed. The technique which can be compared with Schutze approaches are,

1. Linear discriminate analysis
2. Logistic regression and
3. Neural networks.

1. Linear Discriminate Analysis

The main purpose of this analysis is to detect the feature dependency by using covariance class for every individual item class. Let us consider the data sample generated from two groups consisting of the following elements as,

(a) m_1 and m_2 representing members

(b) $\overline{V_1}$ and $\overline{V_2}$ representing mean vectors

(c) CV_1 and CV_2 representing covariance matrix.

The goal of discrimination analysis is to increase the separation distance between two groups which is attained by,

(i) Increasing the distance between the mean vectors.

(ii) By performing scaling transformation in order to reflect the design structure in the covariance matrix.

Therefore, P must be chosen in such a way that P^* is maximized.

$$P^* = \arg_P \max \frac{P^T(\overline{V}_1 - \overline{V}_2)}{\sqrt{P^T C_P}}$$

The variable 'T' in the above formula specifies the transpose and $(m_1 + m_2 - 2)C = (m_2 - 1)CV_1 + (m_2 - 1)CV_2$ where C represents the Cholesky decomposition.

If the value of C is a positive integer, then $C = R^T$

Let us assume $q = R_p$, then the above formula becomes,

$$P^* = \arg_q \max \frac{q^T R^{T-1}(\overline{V}_1 - \overline{V}_2)}{\sqrt{q^T q}}$$

The P^* is maximized by choosing

$$q \propto R^{T-1}(\overline{V}_1 - \overline{V}_2)$$

$$\Rightarrow \quad P^* = R^{-1}q = C^{-1}(\overline{V}_1 - \overline{V}_2).$$

In order to maximize the separation between the two groups, $y = P^{*T}V$ is used for defining one-dimensional space.

Since linear discriminate analysis is used for generating linear classifier, another approach referred to as "regularized discriminant analysis" is used for producing non-linear classifier. Two shrinkage parameters are used for creating a typical estimator family for set of covariance matrix. This analysis process views combination of pooled and unpooled covariance matrices that are assigned specific weights. The selection of shrinkage parameter values should be done by relying on the cross validation over the training dataset. Incomparable to linear discriminate analysis, regularized discriminant analysis does not make major changes in the classification process.

2. Logistic Regression

Logistic regression is another method of identifying reduced concept vector set. It uses logiclink function for creating binary response variable. It does this by considering a linear combination of at least one predictor variable. The logic link function is expressed as,

$$L(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

In addition to modeling binary response variable, this approach also models variance by using a binomial random variable. The modeling of variance is done by modeling the dependent variable $\log\left(\frac{\pi}{1-\pi}\right)$ in the form of linear combination of independent variable. This is achieved by using the formula,

$L(\pi) = a_i w$, where π specifies the probability of relevance; 'a_i' represents the feature vector for item I, and 'w' specifies the weight vectors that is calculated from the covariance matrix of reduced vector. Another way of estimating the value of 'w' is by using maximum likelihood, Newton's Raphson method. The difference between the logistic regression and Schutze approach is that the latter is restricted to perform the analysis for individual profile rather than considering the information from all the profiles.

3. Neural Networks

A neural network is a large parallel distributed processor madeup of simple processing nodes and has connection from one node to another. These nodes are similar to neuron present inside a human brain and hence it is also called as "neurons" in the network. These nodes forms a network that consist of both input and output cells. The input pattern which is received by the nodes are send forward through the network. If an error is detected, the pattern is send backward by modifying the cell parameters in order to reduce the error and to attain the learning objective. The advantage of neural network is that it is very versatile and can support large number of distribution. The problems that arises while using neural network approach are,

(i) It is possible that neural network can overfit by learning the training dataset features.

(ii) The level of generalization is very low for normal input of items.

The first issue can be overcome by ensuring that validation item set is used along with the training items. The validation set is tested recursively whenever the parameters are adjusted on the learning dataset. If the testing results in more number of errors, then there is a chance for overfitting problem to occur. Because of this, the iteration on training set is performed so as to enhance the parameter value without sacrificing the generalization level.

Linear and Non-linear Networks

Non-linear network consists of hidden blocks which in turn consist of three hidden units. These units are treated as 'feature detectors'. The main purpose of these detector is to calculate the probability of vector available in the input pattern. The level of relevance in the output unit can be improved by sending this probability to the output unit. In linear as well as non-linear architectures, every input unit has an explicit connection with the output unit. Relevance computation is a three-step process,

Step-1: Set the input units activation to the item's representation.

Step-2: Send the activation to the output unit via network.

Step-3: Send the error back (if any) using gradient descent algorithm via network.

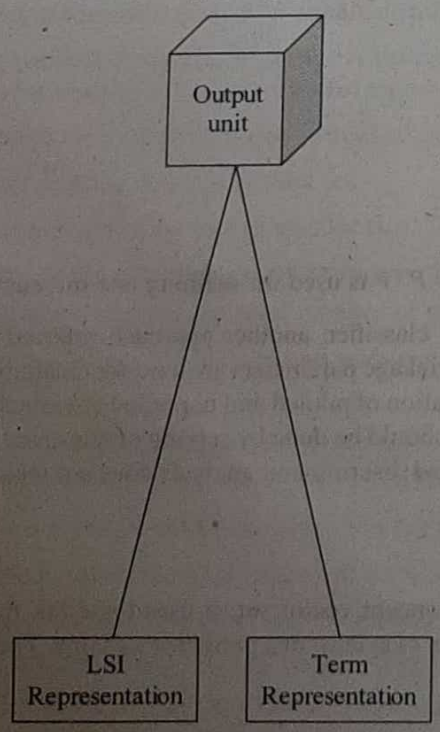


Figure: Linear Neural Network

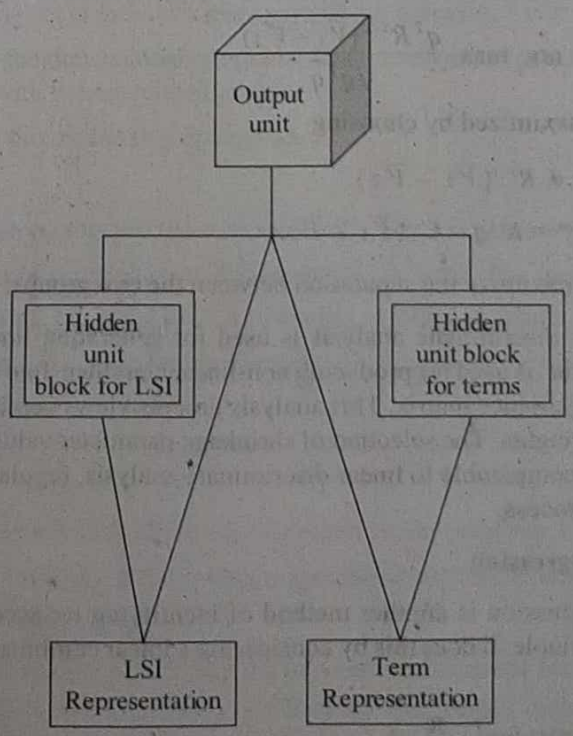


Figure: Non-linear Neural Network

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Activation function for the input and output unit is represented using a sigmoid which is,

$$f(x) = \frac{e^x}{1 - e^x}$$

Back propagation reduces similar error which is minimized by applying logistic regression. The formula for cross-entropy error is,

$$E = -\Sigma e_i \log(\sigma_{i+1} - e_i) \log(1 - \sigma_i)$$

Where, 'e_i' specifies the relevance for item I

'σ_i' specifies the estimated relevance for item 'i'.

$$f(x) \Rightarrow x = \log\left(\frac{f(x)}{1 - f(x)}\right)$$

The above formula is similar to logic link functions. The results obtained after comparing three algorithms with Schutze approach using Tipster test database in contrast to conventional relevance feedback are,

1. Linear classification approach is 10 to 15% better.
2. Dimensionality reduction must be used for utilizing learning algorithm which is dependent on error minimization and statistical classification.
3. LSI is mostly preferred for performing linear discrimination analysis and logistic regression. This is because, there are no methods for avoiding the occurrence of overfitting problem. In situation wherein neural network techniques are used for protecting against overfitting, then it is possible to use less precise method for dimensionality reduction.

Mail files makes it possible to dynamically update the content of the item in an asynchronous manner. Due to this, it becomes sophisticated to extract the result of dissemination in relevant order as desired by the user. Another problem associated with mail files is that, it is very difficult to perform logical reorganization. Dissemination process does not consider a static database while estimating ranking order for an item. This is because it always expand the query by adding new items in order to match user profile with the mail file. Because of these changes, the order of the ranked stored items changes whenever new items are added. This lead to confusion for users, who recognizes any item both by spatial association and by its name. However, it becomes very difficult to retrieve the item when reordering process is changed frequently.

4.1.5 Weighted Searches of Boolean Systems

Q18. Discuss the problems involved in generating queries using boolean logic in weighted index system along with their solutions.

Answer : **Model Paper-I, Q8**

The methods used for generating queries are,

1. Natural language
2. Boolean approach.

The queries generated by natural language have the capability of being represented using numerical and statistical models. These queries are easily used by the similarity measures. Though, it is easy to generate using Boolean logic, the problem occurs when these queries are related to weighted index system. Apart from this problem the other issues are,

(a) Difficulty in knowing the exact working of logic operators, when assigned a weighted value.

(b) Difficulty in knowing the logic of associating weights with the terms in the query.

(c) The operators are restrictive or general when they are interpreted in their normal interpretation.

Salton, Fox and Wu solved this restrictive problem by using "strict-definition" of operators. Usage of this definition will suboptimize the expected level of user's extraction. A problem associated with the strict definition is that the query terms are not assigned ranks, which are generated from a pure Boolean process. By integrating Boolean and weighted systems, the ranking problem can be addressed. Earlier, integration of these systems resulted in the recognition of some ill-defined and unclear features.

Fuzzy set technique was proposed by Fox and Sharat so as to integrate Boolean and weighted systems model. These sets develop the degree of membership conception to individual set.

Degree of membership for "AND" is defined as,

$$DEG_{X \cap Y} = \min(DEG_X, DEG_Y)$$

Degree of membership for "OR" is defined as,

$$DEG_{X \cup Y} = \max(DEG_X, DEG_Y)$$

In both definitions X, Y represents the query terms in an item, DEG specifies the degree of membership. MMM model (i.e., mixed min and max) treats the similarity that exist between the query and item is a linear combination of minimum and maximum weights of item. A similarity formula was proposed by Fox for both AND and OR operation.

$$SIM(Que_{OR}, DOC) = C_{OR_1} * \max(DOC_1, DOC_2, ..., DOC_k) + C_{OR_2} * \min(DOC_1, DOC_2, ..., DOC_k)$$

$$SIM(Que_{AND}, DOC) = C_{AND_1} * \min(DOC_1, DOC_2, ..., DOC_k) + C_{AND_2} * \max(DOC_1, DOC_2, ..., DOC_k)$$

Where C_{OR_1}, C_{OR_2}, C_{AND_1} and C_{AND_2} represents the weighting coefficients for OR and AND operation respectively. A research by Lee and Fox showed that the best performance can be obtained if $0.5 < C_{AND_1} < 0.8$ and $C_{OR_1} > 0.2$.

Paice expanded the MMM approach by taking into account every weight of the item against the maximum/minimum approach. According to Paice, the similarity measure is estimated using the formula,

$$SIM(Que, DOC) = \frac{\sum_{i=1}^{k} w^{i-1} V_i}{\sum_{i=1}^{k} w^{i-1}}$$

In the above formula, 'V_i' specifies terms which are viewed in increasing and decreasing order for AND, OR queries respectively. The variable 'w' indicates the weighting coefficients. For best performance, Lee and Fox suggested that the value of $w = 1.0$ for AND queries, $w = 0.7$ for OR queries.

The drawback of Paice approach is that more computation is required as the values need to be arranged and stored in increasing or decreasing order.

P-norm Model

P-norm model not only allows to have terms in the items, but also enables the query terms to have weighted value. It is similar to cosine similarity measure as it takes into consideration the membership values (V_{n_1}, V_{n_2}, ..., V_{n_k}). These values are assumed to represent the coordinates in an "k"-dimensional space. For "AND" query, the best point is the unit vector where all the values of $V_i = 1$. On the other hand, for OR query the origin is the worst possibility. Therefore, in an OR query the items which are given the highest rank will have maximum distance from origin and will have minimum distance from the unit vector.

The generalized form of queries are,

$$Que_{OR} = (N_1, n_1) \text{ OR } (N_2, n_2) \text{ OR } ..... (N_k, n_k)$$

$$Que_{AND} = (N_1, n_1) \text{ AND } (N_2, n_2) \text{ AND } ... (N_k, n_k)$$

Both the operators have strictness value assigned to them. These value ranges from 1 to ∞ where '∞' signifies the strict definition of AND, OR, NOT operators. The 'n_i' specifies weights assigned to the query term. The similarity formulas between queries and items, if a strictness value is assigned to a parameter (S) are,

$$SIM(Que_{OR}, DOC) = \sqrt[S]{\frac{n_1^S V_{N_1}^S + n_2^S V_{N_2}^S + ... + n_k^S V_{N_k}^S}{n_1^S + n_2^S + ... + n_k^S}}$$

$$SIM(Que_{AND}, DOC) = 1 - \sqrt[S]{\frac{n_1^S (1 - V_{N_1})^S + n_2^S (1 - V_{N_2})^S + ... + n_k^S (1 - V_{N_k})^S}{n_1^S + n_2^S + ... + n_k^S}}$$

$$SIM(Que_{NOT}, DOC) = 1 - SIM(Que, DOC)$$

Salton Approach

In this approach, extra intuition is made under the assumption that no weights are assigned to the terms within the indexes while considering the issues of integrating Boolean and weighted query terms. The goal of Salton approach is to carryout the normal Boolean operations and refine the output by applying weighting methods.

The salton approach is modified in order to define the search results. The output obtained after performing the normal Boolean operation are,

1. **X OR Y:** According to this, the items containing either term X or term Y or both are extracted.
2. **X AND Y:** According to this result, the items containing both terms X and Y are extracted.
3. **X NOT Y:** This means the item that contains term X, but not term Y is extracted.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

If terms are assigned weights between the values 0.0 to 1.0, then it is possible to interpret those terms depending on the importance the users are placing on individual term. If the term is assigned a weight 1.0, then it means that it refers to strict interpretation of Boolean query. If the assigned weight value is equal to 0.0, then it means that the user has very less importance on that term. Because of this, the term with weight equal to 0.0 have no impact on the extracted dataset. Therefore,

X_1 OR Y_0 ⇒ The item set that contains term X, should be returned.

X_1 AND Y_0 ⇒ It also returns the item set that contains term X.

When the weights assigned to term Y moves from 0.0 to 1.0, then it means that a change occur on the resultant set. This set changes from items set that contain term X to the set that is actually produced after performing the Boolean operation.

X_1 or Y_1 : It means those items that contain term X and Y are retrieved.

X_1 or Y_0 : It means items containing only term X are extracted.

X_1 or $Y_0 \neq Y$ NOT X i.e., X_1 or Y_0 contain all items except those present in Y NOT X. Therefore, when the weight for term Y moves from 0.0 to 1.0, the items from Y NOT X are added in equal proportion until all the items at value 1.0 are added.

X_1 AND Y_1 : It means items containing both terms X and Y are extracted.

X_1 AND Y_0 : It means that, items containing only terms X are extracted.

X_1 AND $Y_0 \neq X$ NOT Y i.e., X_1 and Y_0 contains all the items except those present in X NOT Y. Therefore, when the weight for term Y moves from 1.0 to 0.0, then item from X NOT Y are added in equal proportion until all the items at the value 0.0 are added.

X_1 NOT Y_1 : The items containing term X but not term Y are extracted.

X_1 NOT Y_0 : All the items containing term X are extracted.

X_1 NOT $Y_0 \neq X$ AND Y. Therefore, as the weight assigned to term Y moves from 0.0 to 1.0 then the item from X AND Y are added in equal proportion until all the items at 1.0 are added.

However, it is very difficult to identify the items that are to be added or removed, while interpreting the weighted values. It is essential to inspect the items in strict case (where X and Y weights is 1.0) as well as the case when the value for both X and Y is 0.0 (i.e., invariant case) when items are being added, then it is necessary to have items that are similar to the item set which are static in both cases. On the other hand, when items are being removed, then it is necessary to have items that are not very much similar to the items that are present in both cases.

Steps for Solving Weighted Searches of Boolean System

The following are the steps that are followed while solving a weighted Boolean query.

1. In the initial step, those items that matches with the strict interpretation of Boolean operations are identified.
2. In the second step, those items that belong to an invariant set are identified.
3. The centroid of invariant set are identified.
4. The number of items to be added or removed are identified. This can be done by calculating the product of term weight and the number of items that lies outside the invariant set. Finally, the resultant number is rounded-off to its nearest integer value.
5. The similarity value that exist between the items exterior to an invariant set and centroid are identified.
6. The items that are to be added or dropped from the resultant set is selected.

Q19. Given the following documents, determine which documents will be returned by the query ($P_{1.0}$ and $Q_{0.5}$).

TERM	P	Q	R	S
DOC1	2	0	2	1
DOC2	0	2	0	3
DOC3	3	2	1	0
DOC4	2	1	2	0
DOC5	1	0	3	3
DOC6	3	0	1	2
DOC7	1	4	0	4
DOC8	4	0	0	3
DOC9	0	4	1	2
DOC10	0	2	0	0
DOC11	2	0	6	2
DOC12	4	0	0	3

Answer :

Q_1 = Query= $P_{1.0}$ or $Q_{0.5}$

$Q_{1\text{strict interpretation}}$ = (DOC1, DOC2, DOC3, DOC4, DOC5, DOC6, DOC7, DOC8, DOC9, DOC10, DOC11, DOC12)

$Q_{1\text{invariant}}$ = (DOC3, DOC4, DOC7)

$Q_{1\text{optional}}$ = (DOC1, DOC2, DOC5, DOC6, DOC8, DOC9, DOC10, DOC11, DOC12)

Thus [0.5 times 9 items] = 5 items

4.1.6 Searching the Internet and Hypertext

Q20. List and explain the six key characteristics of intelligent agents.

Model Paper-II, Q9(b)

Answer :

Characteristics of Intelligent Agents

The following are the six important characteristic features of intelligent agent.

1. **Autonomy (Freedom of Action)**

The search agent should be independent while performing its operation i.e., it should not have any sort of interaction with a human agent. It must have the ability to make its own decision without relying on other and must manage their own internal state. The search agent browses the information sites depending on some standards that were established in advance. The browsing is done so as to gather all the necessary and relevant information.

2. **Communication Capability**

The search agent must have the ability to communicate with the information site while it is browsing them. In order to make this possible, a standard language which is universally accepted and which defines external interfaces must be used by the search agents.

3. **Capacity for Cooperation**

It is essential for an intelligent agent to cooperate with each other in order to carry out tasks that are mutually advantageous.

4. **Capacity for Reasoning**

This concept consist of three kinds of reasoning scenarios,

(i) **Rule-based Scenario**

The user defines a set of rules and the actions that are to be undertaken.

(ii) **Knowledge-based Scenario**

The intelligent agents stores the rules defined in rule-based scenario and carryout the action through which future actions can be deduced.

(iii) **Artificial Evaluation-based Scenario**

Intelligent agents bring out new agents that have more logic capability so as to attain the goals.

5. **Adaptive Behaviour**

This characteristic is closely linked with the first and fourth characteristics. The adaptive behaviour allows an intelligent agent to determine about the current state and decide about the action that should be performed.

6. **Trust Worthiness**

The user must maintain trust on the intelligent agent who is responsible for locating information that is accessible by user on user's behalf. The users must believe that the information retrieved is for the user.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Q21. Write short notes on,

(a) Steps for implementing automatic relevance feedback approach.

(b) Different Internet based search capabilities.

Answer :

(a) Steps for Implementing Automatic Relevance Feedback Approach

The following are the steps for implementing automatic relevance feedback.

1. Whenever an intelligent agent traverses the information sites, it must utilize similarity measure in order to understand the author's language and associate it with the user's search requirement. In this type of relevance feedback approach, it is necessary to perform normalization across different multiple systems. The amount of information extracted enables the user to concentrate on the relevant item which is assigned the highest rank.

2. In this step, hyperlinks are used for searching information over the internet. A hyperlink is defined as an embedded link to a different item that is instantiated by a click on the item reference. A URL which is associated with the textual data being displayed is invisible to the user. The drawback of using hyperlink process is that a link is not capable of defining any attribute. A link is nothing but a pointer that either specifies the address of another item or refer to another item that support existing textual data.

If the conditions of the link available in the item is understood properly, then it is very easy to identify the effectiveness of the immediate related path. Therefore, hyperlinks makes the creation of fixed network possible. This network is created for linked item depending on the item that is being viewed. Just by following the links, it is possible for a user to manually navigate through the network. Searching is defined as the ability of initiating an item and creating the network of related items.

Once the search have been terminated, the output is a network diagram. This diagram identifies all the items that are related to each other. The purpose of network diagram is to,

(i) Guide the user in knowing his location in the network.

(ii) Enable the movement of users from one node to another within the network.

The information retrieval perspective to the hyperlink issue is,

(i) What is the way of following the hyperlinks automatically?

(ii) How the resolution of user's search requirement gets effected when additional information is generated due to link instantiation?

In order to determine the hits associated with the search statements, different function that assign weight to the term in both original as well as linked item can be applied.

(b) Different Internet Based Search Capabilities

There are different search capabilities that are present on the internet. Among those "dissemination system" is the one growing rapidly. These systems extract relevant information which is required by users for either personal or business use. One of the examples of this system include fishwrap newspaper system that enables the user to specify the area in which he is interested. When the user specifies the area, the search is initiated and the resultant items are sent via email to the desired user or archived in mail files for future extraction. It is the responsibility of the system to update the screen, when new items are detected. In addition to search system, there are many search sites that are responsible for gathering relevant information from users after performing necessary action. After collecting the information, relevance feedback algorithms are used for modifying the information that is being delivered.

Firefly is one such site, that communicate with the user along with their preferences. Firefly systems perform comparison between frequently changing interest of one user from the other user. Empirical media is another system that uses relevance feedback strategy across many users. This is used so as to categorize and differentiate the interest of multiple users. User's preference can be learned based on the ranking order of the user. Apart from internal ranking algorithm, empirical system also use collaborative intelligence for providing a final item ranking to each user.

4.2 INFORMATION VISUALIZATION

4.2.1 Introduction to Information Visualization

Q22. Describe the need for information visualization. Under what circumstances is information visualization not useful. Dec.-19(R16), Q8

OR

Define information visualization. Describe the need for information visualization. In what ways it can help its users?

Answer : Model Paper-I, Q9(a)

Information Visualization

Information visualization conception is emerging for specifying the way the brain processes and uses mental images. In order to perform visualization, it is very essential to have advent in the field of technology to enhance information retrieval mechanisms. Doyle is one of the earliest researchers in information visualization that defined the semantic road conception. These maps assisted the user with a view of the entire database. Another purpose of road maps is to display the interrelated items onto a particular theme of semantic. This view provided by road maps can be utilized by user so that he can concentrate on particular part of database semantic.

Later in late 60's, the semantic road conception was enhanced by Sammon in which the emphasis was on self-organization maps. The purpose of self-organization maps is that they maps to the information available in the database. For knowing the inter-relationship between the items, Sammon applied non-linear mapping algorithms. The associations that are revealed out provide all the necessary information which is required for creating road maps or spatial organization maps.

Information visualization mechanisms have the capabilities of improving the ability of users while finding the desired information. The advantage of this technique is that, it allows the users to utilize minimum resources for extracting desired information. Because of the development of many user interfaces, there was a drastic change in the way the users interacted with computers. These interfaces are dependent on Windows, Icons, Menus and Pointing devices (WIMPs).

Donald A.Norman proposed a different approach which states that,

"people are required to adhere (conform) to technology. It is now the time to reverse this trend, time to make technology adhere (conform) to people".

Norman emphasized that in order to efficiently utilize the ability of user in locating the desired information, it is necessary to concentrate on understanding user interface characteristics and information processing. Once the information is processed it is transferred to a computer interface.

If information is textually represented, then it is very difficult for a user to understand it but at the same time textual representation provides the user with a complete and detailed information to the user. In any information retrieval process, search is carried out in order to extract the relevant item. This search process is initiated by performing information filtering through which the irrelevant items are filtered out and only relevant items are presented to the user. However, this process results in more number of relevant items, due to which it becomes difficult for any user to view the desired information from those resultant items. There are many other systems wherein the relevant items are displayed, based on their ranking order. In these systems, the user spends most of the time reading every individual list page to hit list.

The following are the factors that assist the users in visualizing and presenting the information.

1. The time required to understand the search results and the clusters of relevant information should be decreased.
2. Generate information based on the relationship that exists between the items by considering them as independent items.
3. Carry out actions that are very easy to perform and that generates complex and highly developed search functions.

A research was carried out by FOX et. al for identifying the requirements so as to efficiently perform their work associated with the documents. The major goal of the researchers was to find and examine the patterns available in the database of documents. They focused on representing patterns as well as the desired items visually. Earlier, certain tools were available that allowed their users to view and search the desired documents with respect to the available information space. By concentrating on specific area of interest, the user can easily view recent inherent areas of interest in order to perform investigation. They searched an interface using which it is possible to perform trends identification. This interface also allowed to find out the interest of users in existing topics as well as the topics which are newly emerging. Cognitive parallel processing of numerous facts is performed when information is represented visually. This type of processing is also performed for data associations that are also represented in a visual format.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

More number of hit files are generated as a result of the search process, because of the rapid growth in the information. It becomes essential for a user to view some of the status screens so that it is easy to understand the problems associated with the search statement and the extracted items. With the help of intuitive interface provided by information visualization, the search result can be aggregated into a display. This display is responsible for providing detailed summary and enables them to focus on the most relevant items. When a query have been entered, workspace of relevant items is extracted by that query. It is possible for a user to view as well as modify the items extracted. By using aggregate workspace the association that exist between the relevant items can also be viewed. However, these associations cannot be distinguished, recognized if the items are viewed individually. The advantage of aggregate representation of items is that it enables the user to modify the aggregates in order to process the items available in the workspace.

Another usage of aggregating the search result is that, it makes the correlation between the search term and extracted item possible. Examining both relevant and non-relevant items in such a way that the consequence of the search terms that are expanded by adding additional terms provides intuition about the terms that were the main reason for the result. When the search result is displayed visually then it can be seen that the search term is not effected by the item selection process, due to which this search term must be replaced with another term. The drawback of representing the search result in textual format is that it is not possible to view the interrelationship between the extracted items. Human cognition characteristics are used for understanding the technicality of information visualization system.

4.2.2 Cognition and Perception

Q23. Give a brief account on cognition and perception.

Answer : **Model Paper-III, Q8(b)**

Cognition and Perception

The user-machine interface was developed keeping in view the type-writer model. Since computer displays are emerging globally, man-machine interfaces considers display as a paper enhancement that basically concentrated on the accuracy as well as consistency of different operations. The interface complexity increases and it becomes difficult to manipulate several tasks simultaneously, whenever WIMP interfaces and concurrent parallel tasks are developed and are growing in the working environment of a user. The enhancements occurring frequently on the interfaces concentrated on the way of representing the users about the advancements taking place in the computer world. Human computer interface is extended and evolved due to the advancements in the fields of computer technology, information sciences and human information processing. These enhancements improved the flow of information because of which the user's overhead in finding the desired information is reduced. In addition to information visualization factor, there are many other factors that are being focused for developing future interfaces. The audio factor notifies the user about some of the illegal situation taking place in the computer. For example, a beep sound specifies that a wrong input have been provided to the system. The sound effects are present for every situation including from start-up to shut-down.

Due to the advancement taking place sounds are substituted by speech in input as well as output interfaces. However, the audio is used in research area so as to wrap the information. Apart, from audio factor, touch (tactile) factor is being used on virtual reality arena for providing training about medical procedures where this touch factor plays a vital and crucial role. Information visualization is an important area of interest in information retrieval system.

Until Arnheim postulated his theory cognition and perception, perception was treated as data collection task and thinking was considered as a higher level function. Arnheim modified the definition and specified that,

Perception includes the following,

1. Information understanding process
2. Continuous feedback approach between both perception as well as thinking.

When perception and thinking are viewed as two different functions, then in such situation the mind is treated as serial automata. Visualization is defined as a process of transforming the information into visual format. The advantage of visualization is that it makes the user to easily view, interpret and comprehend the information. It is possible to enhance this visualization conception so as to provide many methods for understanding the information through visual images. This extended visualization concept treated any visual input as understanding process but not as discrete facts. The interfaces that carry information between human and computer enhanced when information processing load is transferred from slower cognitive processes to perceptual system with higher speed.

Q24. Explain the following rules,

(a) Proximity (b) Similarity

(c) Continuity (d) Closure

(e) Connectedness.

Answer :

According to Gestalt postulation, mind always follows certain rules in order to merge input stimuli with the mental representation that varies from certain individual inputs like,

(a) Proximity

In this, figures that are not closely related (i.e., approximate) are combined together to form a single group.

(b) Similarity

In this, figures that are exactly similar to each other are combined to form a group.

(c) Continuity

In this, figures are treated as continuous pattern but not as discrete shape sequences.

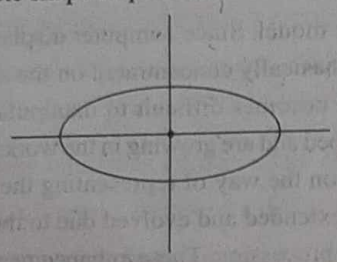


Figure: Ellipse

The above figure can be treated as three continuous shapes an ellipse and two lines but not as four quarter ellipse combined together.

(d) Closure

In this, gaps present inside a figure are filled up, so as to create a complete figure.

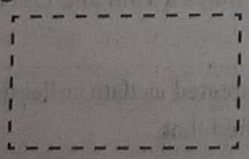


Figure: Rectangle

The above figure can be easily understood that it is a rectangle, though it is represented using dotted lines.

(e) Connectedness

In this, all uniform and interconnected points, lines etc., are treated as a single unit.

Q25. Discuss the limits associated with the use of preattentive processes, configuration aspects and spatial frequency as a basis for information visualization. **Model Paper-I, Q9(b)**

OR

State and explain the principle cognitive engineering. **Oct./Nov.-20(R16), Q7(b)**

Answer :

Various principle cognitive engineering aspects that act as a base for information visualization are as follows,

1. Preattentive Processes

Preattentive process is one of the initial and foremost cognitive processes. It is a process of taking visual information as input from photo receptors and transforming those inputs into primitives. Primitives are generally, a part of preconscious processes that include automatic processing of information at lower level.

Preattentive process is responsible for deleting the boundaries between set of orientation of similar objects. In addition to this process, there is another difficult process that is used to determine the equivalence of object undergone rotation. For example, if a square is rotated, then more effort is required to identify it as a square.

In case of characters, identifying the rotated characters is problematic. Since these characters are normally not available in the alphabetic set. Detecting the symmetry, when the axis is vertical is very easy.

Example

SURE SUER

Optical illusion is another visual factor wherein the dark object on a light background appears to be smaller than the light object placed on the dark background. Because of this factor it is suggested that any object of smaller size can be visually displayed on a background. Colors is another complex area that need to be focused. It is one of the preferred visualization mechanism used for organizing, classifying and enhancing the object's characteristics. The attributes associated with color factor are,

(a) Hue

It specifies the physiological specification of color perception.

(b) Saturation

This attribute specifies the extent to which a hue differs from a normal gray line that have the same level of lightness as the of hue.

(c) Lightness

It specifies the perception about white or black level.

In addition to normal colors, complementary color that are combination of any two colors either form white or gray when merged together. Red, Green, Blue (RGB) are considered as primary colors. Any human being gets naturally attracted towards these colors and the level of image retention associated with these primary colors is very high.

Colors also have negative impact on some people who have color blindness to some or more colors (i.e., colors are of no use for such people). Therefore, it becomes essential for any display to set the other options as well as making use of colors while displaying the objects.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Apart from color, depth is another factor which is more frequently used for visualizing information. The following are various attributes used for representing depth.

Monocular cues, modifications in shading, blurring, perspective, motion, stereoscopic vision, texture.

Lightness is the major factor that effects many monocular cues when compared to contrast factor. It is easier to depict depth as a method for information representation only when colors that increases the level of object's brightness in comparison to background color is selected. The advantage of depth in visualizing information is that, the knowledge of depth is learned in the early stages of life and this knowledge can be used through the life time. Most of the visual processing system prefer depth for processing information and associating (linking) them with the real world models.

2. Configural Aspects

Configural aspect is another higher level processing mechanism that takes place when objects are arranged in such a way that, it enables a user to easily understand the abstract condition present at higher level. The visual process at lower level are replaced by higher level concepts by configural clues. These higher level concepts requires more degree of concentration. The main purpose of these clues is to identify the deviation, distortion taking place in a normal operating environment. These deviations are identified from equal sized objects by visual processing system.

3. Spatial Frequency

A coherent visual image is designed and build-up by human visual and cognitive system that have a tendency to carryout orders. According to multiple spatial channel, whenever inputs are received from external devices (that passes the input as a single index) then only a complex and sophisticated image can be constructed. When all the inputs are received then multiple receptors are used for constructing the final image. These receptors are responsible for detecting changes that take place in spatial frequency, spatial phase orientation,

$$\text{Spatial frequency} = \frac{\text{Number of cycles}}{\text{One degree of visual field}}$$

1 cycle = 1 complete light-dark change

Spatial frequency is defined as an acuity measure concerned with light-dark changes that take place either in the visual field or similar channels. Visual systems are not very much susceptible to spatial frequencies equal to 5 to 6 cycles for one degree of visual field. When a comparison is made between the well defined, distinct image and a blurred image then visual system can easily detect the changes taking place in the well-defined images when compared to the blurred images. It is possible that some of the spatial frequencies retrieve desired patterns when motion is used for combining and displaying information.

Human Sensory System

When a decision is to be made regarding which visual information mechanism should be used as a balance between what should be used for information representation and using these mechanism in the real-world environment human system provide desired output when working with both horizontal and vertical references. The advantage of human system is that it is very easy to detect and process such systems.

When other orientations are being used, then it becomes essential to have extra cognitive processes in order to understand the changes that occur from the expected input. The normal color environment is not considered when bright colored areas of larger sizes are not present. Therefore, bright colors are used to represent items that concentrate on associating themselves with the normal processing.

The hidden risk of information visualization design lies in the user's perspective for understanding the design concepts. The merging of different visual cues for understanding what exactly is been viewed is dependent on two factors,

1. Background, the user use for designing.
2. Information context.

The current information is generally preferred by the users for understanding new information. For example, if user concentrates on items clusters, then they view these clusters in the form of a new presentation that does not exist. This result in modifying the visualization presentation in order to reduce the legacy dispositions. Even the previous knowledge can change the way, a user understands a particular design. Mostly, user understand the figures according to their convenience rather than understanding in the way intended by a designer.

4.2.3 Information Visualization Technologies

Q26. Discuss various information visualization technologies. Nov.-13(R09), Q5(b)

OR

Discuss briefly about information visualization technologies.

Answer : (Model Paper-III, Q9 | May/June-13(R09), Q5(b))

Information Visualization Technologies

Information visualization theories are not only applied to commercial, but also to the experimental systems so as to identify the ways of enhancing the user interface and also to allow information localization. The situations that are being concentrated on information retrieval system are,

1. Examining the best way of displaying the search result
2. Examining the structure data extracted from DBMS
3. Examining the link analysis result associating data.

There are two important classes used for displaying the search result,

(i) Document clustering

(ii) Search statement link analysis.

(i) Document Clustering

Representing the document space visually and presenting it to the user is the major objective of document clustering. This document space is restrained by different searching standard and comprises the document clusters interpreted by the content of the document. These document clusters are visualization tools that specifies the information about the size and topic of the clusters. This information assist the users in navigating through the desired document.

Another objective of document clustering is to help the user in understanding the reason for document extraction so as to provide the information required for query processing. In contrast to conventional Boolean systems, in the current search algorithm and ranking methods it is very difficult to correlate the query with the extracted document and also it is difficult to interpret the effects of adding additional words to the search statement. This problem is solved by visualization mechanism by not only displaying the entire term set but also the additional terms together with the documents retrieved from relevance feedback or thesaurus expansion.

(ii) Search Statement Link Analysis

Link analysis is essential, in addition to structured databases because this analysis maintain information present at aggregate level available in the information system. In this analysis, items are not treated independently, but instead takes into consideration the information which is flowing between the documents whose value is associated with the value of multiple documents. In situation, when query item that is extracted based on particular topic is relevant, then important event dependencies may be viewed by displaying the documents associated by the time factor.

Organizing Information

Information can be organized in a hierarchical tree structure. This method is generally used when the information,

(i) Changes over time

(ii) Represents integral part of larger component

(iii) Combines from higher to lower level.

Whenever the size of the hierarchy increases it becomes very difficult for a user to understand the information representation in two-dimensions. Information visualizer is the earliest research performed on information visualization. This visualizer comprises of many visualization formats like cone tree, perspective wall etc.

(i) Cone Tree

Cone tree is a 3-dimensional hierarchy used for data representation. In this tree, apex level consist of a single node and the base level include all the remaining information subordinate arranged in a circular format. It is also possible that a child node can act as a parent of another cone. Whenever a specific node is selected, it is rotated to the front of the display. In contrast to other hierarchical representation, the cone tree provide maximum available information to the user so that the user can have knowledge about the size of individual subtree.

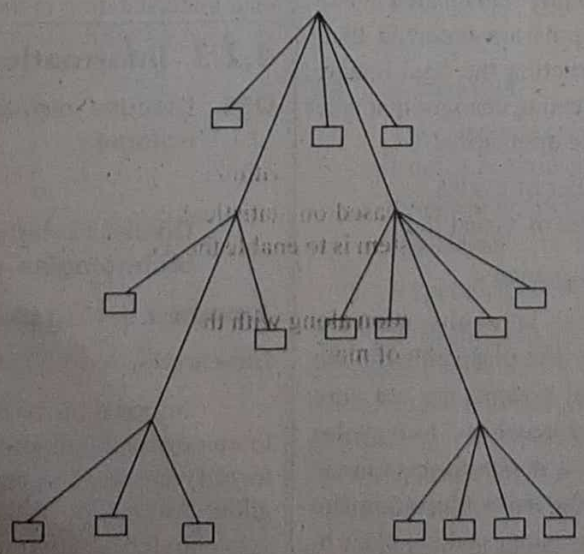


Figure (1): Cone Tree

The leaf-node are the original documents and the higher nodes are treated as centroids that represent the child nodes semantic. The related documents are clustered together whenever the database is of larger size. These clustered document can be extended to form a lower-level of the tree.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

(ii) Perspective Wall

Perspective wall is another kind of visualization format. In this format, the information is divided into three visual areas. The area which is being highly focused is in the front and the remaining two areas will be out of focus present on the either sides of the wall. The advantage of perspective wall is that, the user can maintain the information in accordance to their importance and focuses on specific area at one time.

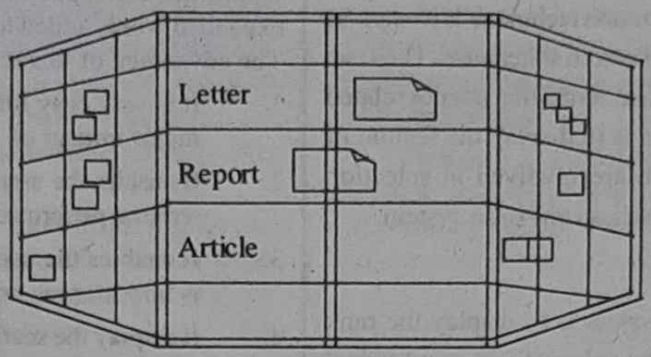


Figure (2): Perspective Wall

Tree Map

Tree maps are another mechanism employed for representing hierarchical information. The information is stored in rectangular boxes that are subdivided continuously depending on the parent-child association that exist between the data. The disadvantage of representing data in rectangular boxes is that, the screen space is used to the maximum extent. The size of these rectangular boxes specifies the total number of items available on a specific topic whereas the location specifies the association between the topics.

When network type association exist between the information, a technique that uses clustering approach can be viewed through the use of semantic scatter plot. The clustering patterns are displayed with the help of 3D scatter plot by both Vineta and Bead systems. All the items are related to each other and multi-dimensional scaling is done so as to plot every individual item as a point over the Euclidean vector space. It is very tedious and difficult to represent the entire axis. This difficulty can be solved by,

(i) Projecting the vector space onto a plane

(ii) Indicating the occurrence frequency by using elevation

(iii) Creating a semantic landscape for specifying the theme importance.

In order to view the detailed association that exist between items and their composite themes, valleys, cliffs displayed on the terrain, map can be used. By creating embedded coordinate space it is possible to overcome the representation of multi-dimensional vector space within a hierarchical environment.

In this approach, the coordinate space of larger size is defined again in which one coordinate lies inside the other coordinates. Therefore, out of 6-D coordinate space, it is possible that three coordinates are defined as a subspace inside the remaining three coordinates. This approach is referred to as "worlds within worlds" approach by Feiners.

The representation issue can be overcome by another approach available in semantic regions. These regions are proposed by Kohenen, linked graphs in Narcissus system. There is also another technique suggested by Lamping and Rao referred to as non-Eulidean landscape for solving the representation issue. Information crystals that are very much similar to VENN diagram, supports the user in detecting the relationship pattern available within the restrained Hit file.

VIBE system is an another clustering system based on statistical information for representing the relationship of terms through spatial positioning. The objective of VIBE system is to enable the user to correlate the terms of query with distinguishable locations available in the visual display.

Lin considered the conception of self-organization along with the Kohenen's algorithm so as to automatically identify the table of contents and to display the result in the form of map.

Goals of Visualization Technique

The goals of visualization technique is to,

1. Display the semantic association that exist between each item. The reason for displaying the association is to support the user in finding the location of the desired item set.
2. Support the users in processing their search statements.

When users are making use of similarity measure for determining the similarity value it is very difficult to identify the main reasons for selecting the item and ranking it in the Hit file. The presence of expanded term in similarity algorithms and its complexity makes it difficult to identify the consequences arising while creating the Hit file because of the availability of different words in the search statement.

The reasons for displaying the association is to support the user in,

(i) Finding the location of the desired item set.

(ii) Refining their search statements.

The main purpose of visualization technique is to support the users in identifying the effects of search statements. They can also assist in identifying the important terms that are not related to the search process. One solution is to display the feature of extracted items graphically which are involved in selection process. This has a significant impact on envision system.

Envision System

The purpose of envision system is to display the rank of relevance item and the relevance value of every individual item identified by a query. Apart from displaying this, it is also capable of displaying other query information. The envision systems are designed graphically using 2D visualization technique which are simple to understand. Basically, envision system consist of three interactive windows for displaying the search results,

(a) Query window

(b) Item summary window

(c) Graphic view window.

(a) Query Window

This window is responsible for maintaining an editable version of the query.

(b) Item Summary Window

The window maintains the bibliographic reference information of the items which is selected in the graphic view window.

(c) Graphic View Window

This window is very much similar to scatter plot graphs.

Every individual item available in the Hit file is represented in the form of an icon. Whenever a particular item is selected in a window, the bibliographic information is displayed on the same window. Single items are represented by circles along with their corresponding relevance weights. Clusters of multiple items are represented by ellipse along with the total number of item value within the ellipse and their relevance weight below the ellipse.

Though some interfaces provides user-friendly environment, at the same time they become sophisticated when large number of items and entries are being displayed. Envision system solves this problem by using a special feature called "ZOOM". This feature allows to view the scatter plot areas of larger size at very less specifications.

Another technique which is similar to envision system is used by system proposed by Veerasamy and Belkin. This system comprises of sequence of vertical columns of bars. This column signifies the documents and the row signifies the index terms. The height of the bar specifies the corresponding term weight within the corresponding item. The system also displays the expanded words added to the system by the relevance feedback. The advantage of this technique is that,

1. It is very easy for a user to identify the term that have a major impact on extracting particular item.
2. It enables the user to identify the procedure of information retrieval performed by many different terms.
3. It enables the user to identify if the term which he feels is important is not contributing to the retrieval process.
4. It display the search terms that contributes to the retrieval of items in such a way that if items are resulting in false hit then they are eliminated from the query or their query weight is reduced.

Document Content Analysis (DCARS) is another system having same characteristics as that of the system used by Veerasamy and Belkin. But DCARS is a business-oriented system developed to Calspan Advanced Technology Center. These systems consist of additional RetrievalWare Search Product. The result of query is displayed in the form of histogram in which the row signifies the item and width of title bar corresponds to the degree of contribution of term to the selection. The advantage of DCARS is that, a user friendly interface is provided, that determines the reason for locating a specific item. The disadvantage of this system is that it is very difficult to modify the search statements in order to enhance the information, which is generally used for determining the way of modifying the search statement.

Cityscape is another way of representing information associated with hierarchical networks. These cityscape utilizes the migratory images in the city.

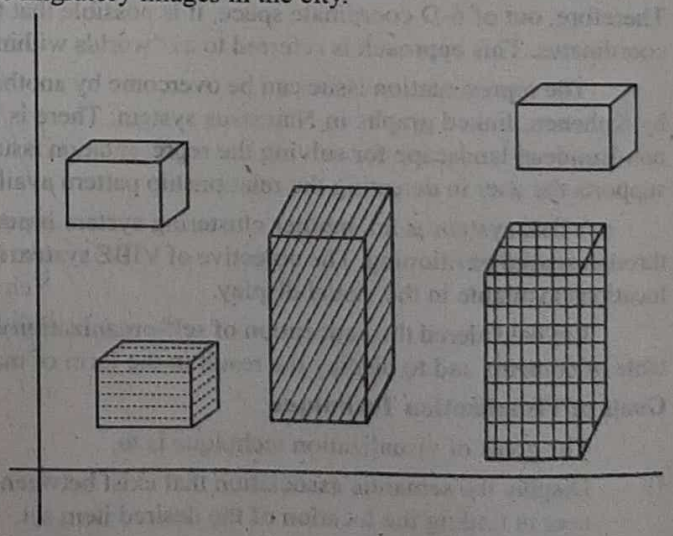


Figure (3): Cityscapes

Cityscape representations is similar to 3D representation of barchart. Performing zooming on a particular domain of information allows the user to view additional design structure that were invisible by the previous presentation.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Whenever entire term relationship approach is used for representing thesaurus and content, then a matrix of larger size is created. In this matrix every individual cell (i.e., combination of row and column) defines the similarity value which is estimated based on the similarity that exist between two terms or items. It is not possible to display the table in a meaningful format. This issue was resolved by MITRE organization that displayed the entire matrix by making use of dot clusters for representing correlations.

When the user performs zooming on a specific correlation domain, then definite set of words becomes visible to the user along with clusters displaying their relationship with the other remaining words.

Thesaurus and content can be represented by another approach which is used by Zizi and Pediotakis. The thesaurus are correlated from the items abstraction that retrieve not only single word expression but also double-word expression. This approach creates two types of views. They are,

1. Presentation view
2. Document view.

The entire display screen space is divided into regions depending on the different classes of the thesaurus. The size of the area is based on the class significance for the collection.

The document view is created only when the presentation view is defined. In the presentation view, ellipse is drawn to specify the documents and the radius of ellipse signifies the document weight.

Information retrieval system is also responsible for representing particular text present in an item visually. There are many situations wherein the items are dynamic in nature i.e., changes periodically through editing. Therefore, it is possible to represent text both dynamically as well as statistically. Dynamic representation of text is very essential when the text being displayed is a software program comprising of hundreds of thousands of code lines. The changes in these line of codes is viewed in SeeSoft system developed by AT&T Bell laboratories by making use of column and color codes. The same technique was used as a fundamental criteria for a similar code visualization tool. These tools represents the line of code within the file in a picture format. The image size is dependent on the number of code line present in the file i.e., the image is scaled in accordance to the code lines.

The purpose of using color codes is to specify the distinct features of the code. The advantage of visualization tool is that it allows the user to view the structure of the code file forming the system together with the individual module's complexity. TileBars tool is a kind of visualization tool generated from Xerox PARC, that enables the users to visually view the distribution of term in a query present in the items, within a Hit file.

Information retrieval system not only concentrates on processing of constructed text but also focuses on the reference data as well as processes structured feature of indexing items. The manipulation of data structure is made possible using structured databases and also using conventional information system. Databases also consist of visualization tools. Query By user interface was the first visualization tool developed by IBM. This interface displays 2-D tables on the display screen which are viewed by users. The user is responsible for defining the desired value to be stored on the table.

Information Visualization and Exploration Environment (IVEE) is a visualization tool that represents the structured database in 3-D format. The representation is restricted depending on the search statement of the user. For example, a larger space can be represented using 3-D box whereas the smaller boxes can be used for representing the realization of particular values. This tool can also represent data in the form of maps and starfields. The user is given the capability of manipulating the search statement because of the presence of sliders and toggles.

Whenever, relevant items are extracted using hyperlink approach, then there is a possibility that the user encounters many orientation issues that are related to the path which the users follows for reaching the current location. This issue can be overcome by allowing the user to view the information space. Pointing devices can be used by a user for specifying the item that the user desires to navigate through the network. A visualization tool was developed by MITRE corporation that allowed the user to view the navigated information space visually in a hierarchical structure format.

Information visualization system also represents pattern and linkage analysis in a visual form. Path finder is one such system that consists of many visualization approaches. This system consist of the following components,

(i) Document Browser

This component is used for signifying the relative items essentiality. It does this by making use of a variety of colors and their density.

(ii) CAMEO

This component performs analytical processing by creating nodes and links in order to represent a problem. Nodes are used for representing a query and the color of the nodes changes depending on "how well the extracted relevant items satisfies the user entered query".

(iii) Counts

This component make use of statistical information associated with the words and phrases and plot them based on the time parameter.

(iv) Cross Field Matrix

This matrix component is responsible for creating 2-D matrix of two attributes within the set of data. The values that are selected are stored in every individual row and column for the two attributes. The lifetime period is represented using different colors.

(v) Oil Stock

It enables to place the data on a geographic mapping tool.

(vi) SPIRE Tool

It is a kind of information scatter graph in which the items are clustered before displaying them in a star-chart format. Similarity value can be estimated by calculating the distance between the two points depending on the concurrency of terms.

UNIT 5

Text Search Algorithms and Multimedia Information Retrieval

Syllabus

Text Search Algorithms: Introduction to Text Search Techniques, Software Text Search Algorithms, Hardware Text Search Systems.

Multimedia Information Retrieval: Spoken Language Audio Retrieval, Non-speech Audio Retrieval; Graph Retrieval, Imagery Retrieval, Video Retrieval.

Learning Objectives

- Typical Architecture of Text Search Techniques
- Various Software Text Search Algorithms
- Importance of Hardware Text Search Systems
- Information Retrieval from Spoken Language
- Retrieval of Non-speech Audio
- Other Multimedia Information Retrievals likes Graph, Image and Video Retrievals.

Introduction

Text search algorithm allow multiple users to enter the queries that contain text. This text is later searched, accessed and compared to query related terms. This generated query is said to be complete if the complete text in it is accessed successfully. The basic advantage of text search algorithms is that, if the searched query is satisfied it result is sent to the user for retrieval. Text search algorithm are of two types type are hardware text search algorithm and software text search algorithm. Hardware text search systems handle multiple search terms concurrently. Whereas software text search algorithm is applied to perform search on same item multiple times simultaneously.

Multimedia information retrieval refers to the criteria of fulfilling information based needs of the users. This criteria involves determination of text, graphics, pictures, documents etc., matching the user queries. The demand for such sophisticated retrieval system is due to the increase in use of world wide web. This platform contains mixture of all types of documents (i.e., text and multimedia).

PART-A SHORT QUESTIONS WITH SOLUTIONS

Q1. Explain the Text Streaming Architecture with a neat diagram.

Answer : Dec.-19(R16), Q1(j)

A text scanning system allows one or more users to enter queries that contain text. This text needs to be searched (scanned), accessed and compared to the query related terms. The query is said to be complete if the complete text in it has been accessed successfully. The basic advantage of this architecture is that when an item satisfies a query, then its results are sent to the user for retrieval.

Q2. Define finite state automata.

Answer : Model Paper-III, Q1(i)

A finite state automata is generally a logical machine that consists of the following five elements.

(i) **X:** It refers to a set of input symbols that are supported by the automata.

(ii) **A:** It refers to a set of possible states.

(iii) **P:** It refers to a set of productions that describes the next possible state which depends upon the current state and the input symbol.

(iv) A_0: It refers to a special state which is known as the initial state.

(v) A_F: It refers to a set of one or more final states from the set A.

Q3. Write short notes on hardware text search algorithm.

Answer : Model Paper-I, Q1(i)

Hardware text search systems are developed to overcome the limitations of software text search systems that failed to handle multiple search terms concurrently against similar textual data. The major drawback of software text search systems is that they perform the search process with a limited I/O speed. One method to reduce the resource searching overhead is to utilize a specialized hardware machine. This machine is used to carry out the search process and then transfer the search result to actual computer that supports both user interfaces as well as hits retrieval. In hardware search machine, many search machines are executed parallely against the similar textual data so that more queries can be handled at one time.

Q4. Write about fast data finder.

Answer :

Fast Data Finder (FDF) is another type of hardware text search system which is developed to perform textual search as well as english and foreign language search. The first FDF machine comprised of array of programming cells which are connected sequentially to form a pipeline hardware search processor. VLSI chips are employed for implementing these text processing cells. Individual chip consist of 24 processor cell and a usual integrated system generally consists of 3600 cells where each individual cell acts as a comparator. All cells are connected to each other by using data path and control path of 8-bits and 20-bits respectively. The searching in FDF is done in a pipeline format i.e., the text passes from one cell to another in the form of a pipe until the entire database is searched. Data analysis is performed at every cell due to which, the 20-bit control paths are modified depending on the existing state as well as the result received from the comparator.

Q5. What is brute force approach?

Answer : Model Paper-I, Q1(j)

In brute force approach, the string to be searched is matched against the input text. If a mismatch occurs during comparison, the input text is shifted to one position, thereby starting the comparison process again. The expected number of comparisons during the search of an input string of 'm' characters for a text pattern of 'n' characters is,

$$E_s = \frac{s}{s-1}\left(1-\frac{1}{s^n}\right)(m-n+1)+0(1)$$

Where,

E_s refers to the expected number of comparisons and

's' refers to the size of the alphabet for the text pattern.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Q6. Write a brief note on Spoken Language Audio Retrieval.

Answer : (Model Paper-II, Q1(i) | Dec.-19(R16), Q1(i))

Spoken language audio retrieval refers to the process of retrieving information from spoken data. It is used for its vast range of applications like retrieving data from radio broadcasts, speeches and conversations. In the process of its implementation, higher error rate is produced which can be minimized with redundancy, inclusion of more dictionary words and various language recognition processing.

Q7. Write a short note on graph retrieval.

Answer : Model Paper-III, Q1(j)

Graph retrieval refers to the process of retrieving data carrying tables, charts etc., which are drawn using points, lines and labels. An example interface of such a retrieval is Sagebook developed at Carnegie Mellon University. It is capable of allowing its users to search and customize the graphic data. It offers features like indexing, searching, representing and others.

The queries in this interface can be generated using SageBrush where it displays a set of options like objects, colours, spaces etc. The results can be filtered based on exact matches and similar matches. The result (object) generated by the query can be customized accordingly.

Q8. Explain in brief about TDT.

Answer : Model Paper-II, Q1(j)

TDT (Topic Detection and Tracking) for newswire sources and broadcast news is used for introducing new algorithms including,

(i) Story Segmentation : It identifies story boundaries.

(ii) Topic Tracking : It identifies stories that discuss the topics related to target.

(iii) Topic Detection : It identifies stories that discuss an arbitrary topic.

PART-B ESSAY QUESTIONS WITH SOLUTIONS

5.1 TEXT SEARCH ALGORITHMS

5.1.1 Introduction to Text Search Techniques

Q9. Draw and explain the text streaming architecture.

Answer : **Model Paper-I, Q10(a)**

A text scanning system allows one or more users to enter queries that contain text. This text needs to be searched (scanned), accessed and compared to the query related terms. The query is said to be complete if the complete text in it has been accessed successfully. The basic advantage of this architecture is that when an item satisfies a query, then its results are sent to the user for retrieval. The figure below shows a text streaming search system.

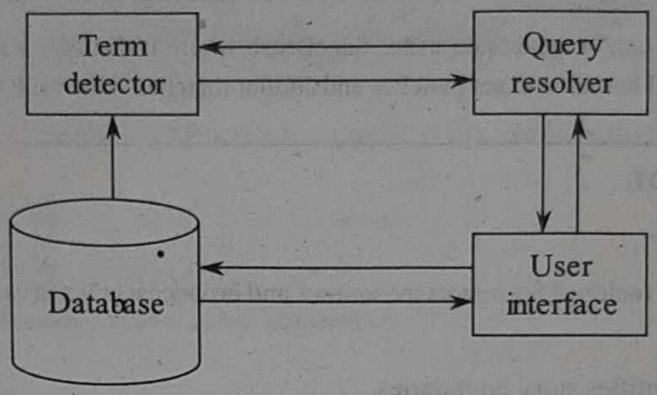


Figure: Text Streaming Architecture

The above architecture consists of database, term detector, query resolver and user interface components.

(i) Database

The database holds all the text related to items.

(ii) Term Detector

The term 'detector' refers to a hardware or software that consists of all the search terms and in certain systems the logic between the items is also stored within it. It inputs the text and detects the presence of the search terms. The detected terms are given to the query resolver so that final logical processing of a query is done.

(iii) Query Resolver

The 'query resolver' accepts the search statements given by the users. It extracts the logic and the search terms and outputs the search terms to the term detector. The results of the searched terms are sent by the detector to the query resolver to identify which queries satisfy the item and its associated weight.

(iv) User Interface

The information generated from the results of searched terms is passed by the query resolver to the 'user interface', which continually updates the search status and provides it to the user and on request the items that are satisfied by the user search statements are retrieved.

The main objective of text streaming process is to find at least one or all the occurrences of text pattern in a text stream. The worst case for searching "n" sequence of characters in a string of "m" characters is at least $(m - n + 1)$ or it requires a search time of $O(m)$. The brute force method requires $O(m*n)$ and certain improved methods require a time of $O(m + n)$.

Some hardware search systems like multiple parallel search systems (term detectors) may have the ability to work against the same or different data streams, thereby reducing the accessing time of the complete database. But, in case of software systems, various term detectors may execute simultaneously.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

The two approaches that are applied to the data stream are,

1. The complete database is sent to the term detectors for performing a search operation on it.
2. The randomly retrieved items are sent to the term detectors.

The second approach deals with an index search of the database. It allows the text streamer to perform a different search logic, which doesn't satisfy the index search. The following are the examples of index search limits.

(i) Searching for stop words.

(ii) Searching for accurate matches on performing streaming operation.

(iii) Searching for the terms that consists of both leading and trailing "don't cares".

(iv) Searching for the symbols that are available on the interword symbol list (i.e., " , ;).

The major disadvantage of text streaming search is that the search is dependent on the I/O module, which is the slowest module in the computer. Inversion lists or indexes minimizes the data that is to be retrieved, thereby increasing their processing speed. They provide an optimum ratio of the total number of items that are to be delivered to the user and the total number of items that are to be retrieved in response to a query. The inversion systems require storage overheads (i.e., 50% to 300%) of the original databases but, the full text search function doesn't need any of the additional storage overhead. An advantage of text streaming search is that as soon as hits are detected, they are returned to the user. Typically, in case of an index system, before determining or detecting any hits, the complete query must be processed successfully. Text streaming systems maintain the current search status and the completion time of the query. Inversions or indexes also face problems in imbedded query search terms and in fuzzy searches. The possible index values are difficult to locate, due to which, the complete search of dictionary of possible terms is also difficult. The imbedded query terms are located by the streaming algorithm and fuzzy searches are performed by the hardware search units and by some other algorithms. The use of hardware text search units protect a searchable environment where the problems associated with the performance bottlenecks can be overcome by providing additional search units so that they can work in parallel with the data that is being streamed.

Q10. Define finite state automata. Draw an FSA that can identify the character string CPU in an input stream.

Answer :

A finite state automata is generally a logical machine that consists of the following five elements.

(i) **X:** It refers to a set of input symbols that are supported by the automata.

(ii) **A:** It refers to a set of possible states.

(iii) **P:** It refers to a set of productions that describes the next possible state which depends upon the current state and the input symbol.

(iv) A_o: It refers to a special state which is known as the initial state.

(v) A_F: It refers to a set of one or more final states from the set A.

A finite state automata is used by most hardware and software text searchers. It is designated by a directed graph which consists of a sequence of nodes and edges that exists between the nodes. The edges denote transitions defined by a set of productions. The symbols present at each edge defines the input that transit from one state A_i to another state A_j. Following figure represents a finite state automata which determines the character string i.e., CPU.

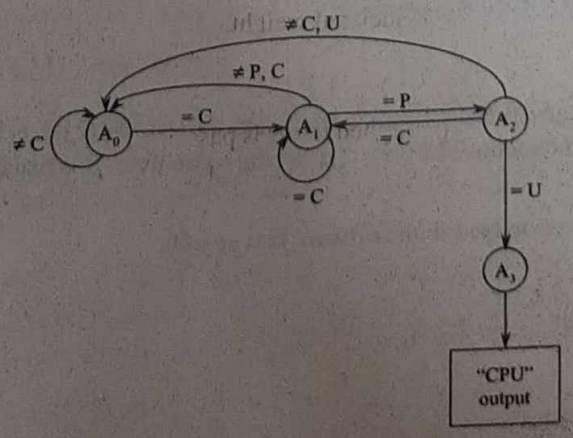


Figure: Representation of Finite State Automata

The automata definition is described as in the following and is used to define the automata.

$X = \{\text{alphabetic characters}\}$

$A = \{A_0, A_1, A_2, A_3\}$

$P = \{$if $X = C$ then $A_0 \rightarrow A_1$

if $X \neq C$ then $A_0 \rightarrow A_0$

if $X = P$ then $A_1 \rightarrow A_2$

if $X \neq \{P, C\}$ then $A_1 \rightarrow A_0$

if $X = C$ then $A_1 \rightarrow A_1$

if $X = U$ then $A_2 \rightarrow A_3$

if $X = C$ then $A_2 \rightarrow A_1$

if $X \neq \{C, U\}$ then $A_2 \rightarrow A_0$

and $A_0 = \{A_0\}$

$A_F = \{A_3\}$

The automata will remain in the initial state i.e., in A_0 unless it contains an input string character "C" which is allowed to transit from A_0 to A_1 state. It remains in A_1 state until all "C's" are received as input and if a "P" is received by A_1 then it transits to A_2. If the automata receives any other string character, then it goes back to the initial state A. If "U" is received by the automata in state A_2 then it will go to the final state A_3, if "C" is received then it will go to the state A_1 or if any other character is received then it goes back to the initial state A_0.

The productions can be represented in the form of tables with states as rows and input symbols as columns. The states and values (input symbols) in the table designates the current and next state, respectively.

5.1.2 Software Text Search Algorithms

Q11. What are software text search algorithms? Explain them clearly. **(Nov.-13(R09), Q6 | May/June-13(R09), Q6)**

OR

Explain in detail about software text search algorithms. **(Model Paper-I, Q11 | Dec.-19(R16), Q10)**

OR

Discuss any one software text search algorithm. **Nov./Dec.-12(R09), Q6(a)**

OR

Explain Boyer-Moore algorithm with relevant example. **Dec.-14(R09), Q6(a)**

(Refer Only Topic: Boyer-Moore Algorithm)

OR

Write Boyer Moore algorithm for string searching.

(Refer Only Topic: Boyer-Moore Algorithm)

Answer : **Oct./Nov.-20(R16), Q8(a)**

Software Text Search Algorithms

In software text streaming techniques, the algorithm is applied to the item after it has been read in the memory. As software streaming can be applied to multiple simultaneous searches against the same item, it is usually used to facilitate the search for an item.

The following are the algorithms associated with software text search,

1. The Brute Force approach
2. Knuth-Morris-Pratt algorithm
3. Boyer-Moore algorithm
4. Karp-Rabin algorithm
5. Shift-OR (or) Shift-AND algorithm.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

1. The Brute Force Approach

In this approach the string to be searched is matched against the input text. If a mismatch occurs during comparison, the input text is shifted to one position, thereby starting the comparison process again. The expected number of comparisons during the search of an input string of 'm' characters for a text pattern of 'n' characters is,

$$E_s = \frac{s}{s-1}\left(1-\frac{1}{s^n}\right)(m-n+1)+O(1)$$

Where,

E_s refers to the expected number of comparisons and

's' refers to the size of the alphabet for the text pattern.

Example

Position	:	1	2	3	4	5	6
Input stream	:	p	q	p	p	p	q
Search pattern	:	p	p	q			
			↑				

Figure (i)

In figure (i), a mismatch occurs in position 2 of the search pattern. Thus, as per Brute-Force approach one position is shifted to the right of the search pattern.

Position	:	1	2	3	4	5	6
Input stream	:	p	q	p	p	p	q
Search pattern	:		p	p	q		
			↑				

Figure (ii)

In figure (ii), a mismatch occurs again in position 2. Thus, to align the input character to a character in the search patterns, one position of the search pattern is shifted until the two patterns match. After aligning the characters of two patterns the result will appear in the following way in figure (iii).

Position	:	1	2	3	4	5	6
Input stream	:	p	q	p	p	p	q
Search pattern	:				p	p	q

Figure (iii)

The time complexity for the comparison process is O(mn) which is optional and the no preprocessing phase exists disadvantage is that it has worst case quadratic (O(nm)) time complexity which is not good.

2. The Knuth-Morris-Pratt (KMP) Algorithm

The KMP algorithm provides an improvement in the previous algorithms i.e., even in the worst case, this algorithm is independent of the length of the text pattern to be searched. In this algorithm, if a mismatch is detected, the previously matched characters determine the number of positions that can be skipped in the input pattern instead of starting the comparison process again. Consider an example where,

Position	:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Input stream	:	q	p	q	r	q	p	q	r	p	q	r	p	p	q	r	p
Search pattern	:	p	q	r	p	q	r	p	r	p	q						
		↑															

Figure (iv)

In the above input stream, a mismatch occurs in position 1 with a "*q*" in the stream. Using Brute-Force approach the input text may be shifted by exactly one position to the right and start the comparison again.

Position	:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Input stream	:	q	p	q	r	q	p	q	r	p	q	r	p	p	q	r	p
Search pattern	:		p	q	r	p	q	r	p	r	p	q					
						↑											

Figure (v)

In the above input stream, a mismatch occurs in position 5 with a "*q*" in the stream. As three positions in search pattern i.e., "*pqr*" matches, they are allowed to jump to at least three positions to the right.

Position	:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Input stream	:	q	p	q	r	q	p	q	r	p	q	r	p	p	q	r	p
Search pattern	:					p	q	r	p	q	r	p	r	p	q		
						↑											

Figure (vi)

Again a mismatch occurs after skipping three positions. Now, shift one position to the right.

Position	:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Input stream	:	q	p	q	r	q	p	q	r	p	q	r	p	p	q	r	p
Search pattern	:						p	q	r	p	q	r	p	r	p	q	
														↑			

Figure (vii)

In the above pattern, a mismatch occurs in position 13 with a "*p*" in the input stream. The longest repeating pattern in the search text is "*pqrp*". Thus, three positions are to be skipped in the search pattern. The result is as follows,

Position	:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Input stream	:	q	p	q	r	q	p	q	r	p	q	r	p	p	q	r	p
Search pattern	:									p	q	r	p	q	r	p	q

Figure (viii)

3. Boyer-Moore Algorithm

Boyer-Moore recognized that the algorithm could be more efficiently done if the comparison process starts from right to left of the search pattern. The advantage of this kind of comparison is that large jumps are possible whenever large number of characters match in the input stream. The following are the two possible sources of determining the number of characters that are to be jumped.

(i) Algorithm 1

Whenever a mismatch occurs, the characters in the input stream are compared to the search pattern to determine how many input characters are to be jumped to align them to a characters in the search pattern. If a mismatch occur i.e., a character does not exist in the search pattern then the length of the matched characters of the search pattern are shifted.

(ii) Algorithm 2

Whenever a mismatch occurs the matching process jumps to the previously matched input sequence, thereby aligning it with the text in the search string.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

On a mismatch, maximum of (Algorithm 1, Algorithm 2) is used to skip the characters in the pattern. Figure (ix) gives an example of the comparison process where the alphabets used are (p, q, r, s, t, u)

Position	:	1	2	3	4	5	6	7	8	9	10	11	12	13
Input stream	:	u	p	q	u	p	p	q	q	s	p	q	p	q
Search pattern	:		p	q	s	p	p	q						
					↑									

Figure (ix)

The comparison begins from right end of the search pattern and move towards the left of the search pattern. In figure (ix), the characters in positions 7, 6 and 5 match but the characters in position 4 does not match , hence Algorithm 1 wants to align the mismatched character (i.e., u) of the input stream by skipping three positions to the right. Algorithm 2 determines that the mismatch occurs in position 4 after matching on three position in the search pattern. Thus, four places are moved to align the characters.

Algorithm 1 = 3

Algorithm 2 = 4

Max(Algorithm 1, Algorithm 2)

→ Max(3, 4)

= 4

Position	:	1	2	3	4	5	6	7	8	9	10	11	12	13
Input stream	:	u	p	q	u	p	p	q	q	s	p	q	p	q
Search pattern	:						p	q	s	p	p	q		
									↑					

Figure (x)

In figure (x) a mismatch occurs in position 9. Algorithm 1 determines that the mismatch character "*s*" can be aligned in the input text by moving only one position to the right of the pattern. Algorithm 2 recognizes that again after a match of three characters the fourth character mismatches. Thus, four positions of the search pattern are moved.

Algorithm 1 = 1

Algorithm 2 = 4

Max(Algorithm 1, Algorithm 2)

⇒ Max(1, 4)

= 4

Position	:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Input stream	:	u	p	q	u	p	p	q	q	s	p	q	s	p	p	q
Search pattern	:										p	q	s	p	p	q

Figure (xi)

Similarly upon moving four places of the search pattern, the two patterns match. The advantages is that the time complexity for the comparison process is $O(m + n)$ which is efficient among all the algorithms for patterns of any length. The disadvantage of this algorithm is that more preprocessing time is required to setup the tables.

4. Karp-Rabin Algorithm

An approach which has a similar functionality as that of n-grams and signature files is used to partition the input text string into 'n'-characters and then calculates a hash function i.e., the signature value for each of the individual strings. The calculated hash value for the search pattern is compared to the input text. Karp and Rabin discovered a significantly enhanced signature function so that the hash values can be calculated.

$$h(l) = l \bmod p$$

Where, 'p' refers to a large prime number.

Each location in the input text is associated with a signature value that depends upon the calculated values for the previous locations. Hashing functions do not provide any unique characteristics, hence Karp-Rabin algorithm is used to find only those positions in the input text whose values are similar to the hash values of the search pattern. Then the actual text must be compared with the search pattern to ensure that a match exists.

5. Shift-OR (or) Shift-AND Algorithm

Baeza-Yates and Gonnet enhanced the software text search conception so that it is possible to handle "don't care" as well as the compliment symbols. In contrast to the other searching algorithm, Baeza-Yates and Gonnet search algorithm which is based on vector approach handles cases upto 'n' mismatches. The vector consists of 'v' distinguishable states, where 'v' specifies the length of the search pattern. The state i gives the information about the search state which lies between the pattern's positions value 1, 2, 3,.....,k. It also specifies the text positions $(l-k+1)$ ..... l, where 'l' denotes the current pattern position in the text. The search process is expanded so that it can be viewed as if 'v' comparators are being executed simultaneously.

Let the set of states be denoted by $s_{k,l}$ after the character at l^{th} position is read [where k ranges between 1 and v]. This set specifies a distinguishable character with respect to their position between sp_1,.....,sp_k (search pattern) and st_{l-k+1}.....st_l (searching text). When the value of $s_{k,l}$ reaches zero it means that the search pattern exactly matches the text that is being searched. If the value of $s_{k,l}$ is not equal to zero then fuzzy search techniques need to be employed. In this technique the length of both search term as well as found term are similar and the value is used to specify the number of mismatches.

Let us consider a search pattern as '*pqpqr*' (where $r = 5$) and input text as '*rqqpqpqpqrpqp*'. The value for $s_{k,l-1}$ is calculated in the following way.

	Value of Vector	Position of Vector
p q p q r	1	1
p q p q r	0	2
p q p q r	3	3
p q p q r	0	4
p q p q r	5	5
.....r q q p q p q p q r p q p.....		

Table (i): Vector for Position l – 1

The value of vector for vector position 2 is 0 since two sp characters (i.e., pq) mismatch with the two st characters (i.e., qp). A new vector is created by moving one position ahead, which is shown in the table (ii) below.

	Value of Vector	Position of Vector
p q p q r	0	1
p q p q r	2	2
p q p q r	0	3
p q p q r	4	4
p q p q r	1	5
.....rq q p q p q p q r p q p....		

Table (ii): Vector for Position l

Let the above table be denoted as tab (t). The value of $tab_k(t) = 0$ if $t = sp_k$ which implies that whenever the search pattern matches the text, the position of vector becomes zero because of which tab(t) = '0' at that position and other position will have tab (t) = 1. The value of $tab_k(t)$ at $l-1$ is (10101) and the value of $tab_k(t)$ at l is (01011).

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Thus, it is possible to estimate the value of $s(k,l)$ using the formula,

$$s(k, l) = s(k, l-1) + \text{tab}_k (\text{text}_l).$$

If the value of $s(k,l) = 0$ then the following new values are generated using the above formula by moving from old position (0) to new position (n).

$$s(1, n) = s(0, 0) + T_1 (n) = 0 + 0 = 0$$

$$s(2, n) = s(1, 0) + T_2 (n) = 1 + 1 = 2$$

$$s(3, n) = s(2, 0) + T_3 (n) = 0 + 0 = 0$$

$$s(4, n) = s(3, 0) + T_4 (n) = 3 + 1 = 4$$

$$s(5, n) = s(4, 0) + T_5 (n) = 0 + 1 = 1$$

Baeza-Yates and Gonnet algorithm is referred to as Shift-ADD algorithm because of the above operations. This algorithm handles "don't care" and compliment of a character by considering the following three possibilities.

(a) A character from the alphabet

(b) A "don't care" character (*)

(c) A compliment of character (c').

Let 'v' denotes the total number of items in each class (i.e., the pattern size). The 'don't care' character (*) is assigned weight 1 and the compliment character is not considered.

Pattern : $[k_k]\ l\ [\overline{aeiou}]^*\ \bar{l}\ [k,.....os.....u]$

Value : 2 1 5 1 1 5 + 3

Therefore, $v = 6$ and 'v' = 18 [since, 2 + 1 + 5 + 1 + 1 + 5 + 3 = 18].

In order to process every individual character for each position, the Shift-ADD algorithm is modified.

Advantage

It can be easily implemented in hardware text search system.

Disadvantages

(i) It does not handle insertion and deletion of items.

(ii) It fails to identify positional mismatches.

Q12. Explain the significance of GOTO, failure and OUTPUT functions in Aho-Corasick algorithm.

Answer :

Aho-Corasick algorithm is an approach that uses a finite state machine to process various query related items. The machine contains a number of states and processes the text by successively reading it. Based upon the current state if a match occurs then the state transitions are done. This algorithm is based on Knuth-Moris-Pratt algorithm.

The following functions are used for processing the machine.

(a) GOTO (State transition)

(b) Failure function

(c) Output function.

(a) GOTO Function (State Transition)

GOTO function refers to a directed graph where the lines connected between two states are used to specify the transitions for the given input.

Example

The GOTO function, is applied for the following set of words tool, tin, test and stone.

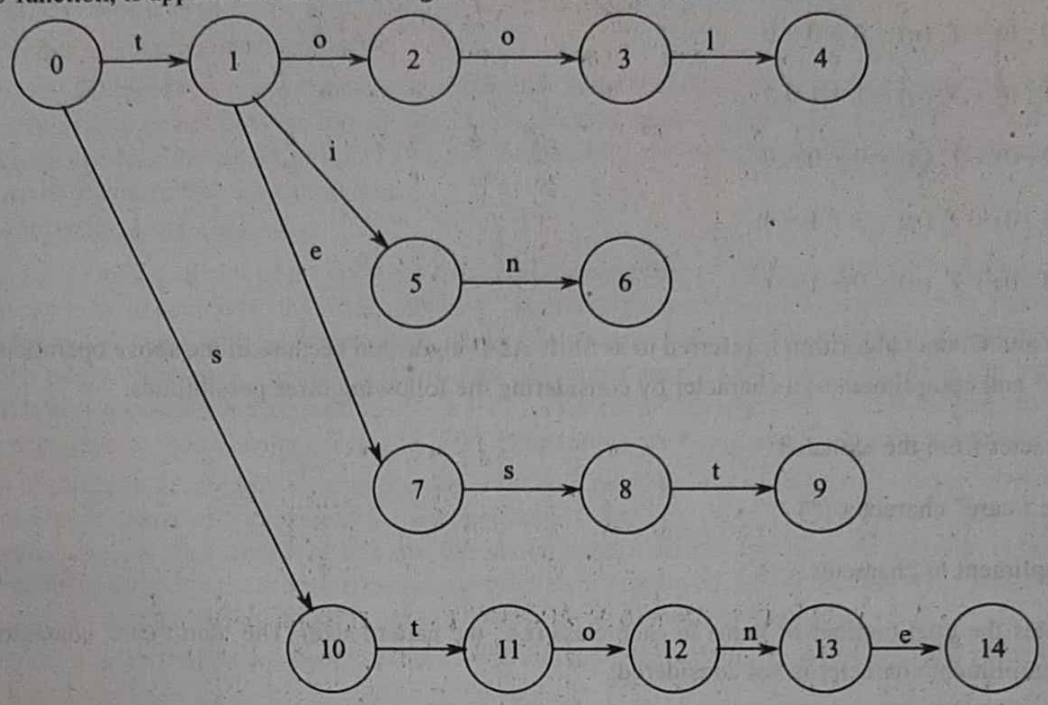


Figure (i): GOTO Function

In figure (i), state '0' represents the initial state. The words begin with either character *t* or *s*. Initially if "*t*" has been received then the system moves to state 1. If the next input character is "o" then the system goes to state 2, if '*i*' and '*e*' are received then it goes to states 5 and 7, respectively. Similarly, if "*s*" has been received, then the system goes to state 10. This process continues until all the input characters of the given set of words match. Thus, a state transition diagram is drawn by showing all the input text symbols on each transition line.

(b) Failure Function

A failure function is used to determine that there exists no directed line and an input character associated on that line. Whenever a failure occurs at a particular state, then its function maps a state into the another state or that state itself to continue the search process.

Example

Step x	1	2	3	4	5	6	7	8	9	10	11	12	13	14
States f(x)	0	0	1	7	0	10	11	12	0	0	0	1	2	3

Figure (ii): Failure Function

In the table, shown in figure (ii) failure function i.e., $f(x)$ determines that if a difficult character is received or an input character doesn't match, then this function is returned by indicating that the system should be moved back to state 0 until the input characters match. Now consider an example where the input string is *x, u, t, e, a, s, t, o, d, g, m, t, o, o*. From figure (ii), the first step has state '0', which indicates that the initial character i.e., "*x*" of the given input string does not match with the characters in the given set of words (to, tool, test, tin, stone). Hence, the system remains in state '0'. Next, the second character i.e., "*u*" is matched with the characters of the words. Since, it doesn't match, the system remains in state '0'. The third character i.e., '*t*' is matched. Since, '*t*' exists in the given set of words, step 3 results in state 1. Similar procedure is followed for the other characters in the input string.

Thus, in $f(x)$ the state "0" is the failure function and the states 1, 7, 10, 11, 12, 1, 2, 3 specify the input characters that has been matched at these states.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

(c) **Output Function**

Output function is used to determine that the query terms have been matched.

State	2	4	6	9	14	12
Output	to	tool	tin	test	stone	to

Figure (iii): Output Function

The given table determines the output states because upon reaching at these states the input query terms match during the search process of the Aho-Corasick algorithm.

Both the Aho-Corasick and the KMP algorithms have similar number of characters to be compared. In case of a new algorithm, the state transitions that are used to process a string do not depend upon the search terms and the search operation to be performed is linear in accordance with the characters in the input string. The number of characters that are compared in the input stream has an order of magnitude of $kO(T)$ where k refers to a constant which is greater than 1 and T refers to the characters that are available in the input string. This improves the performance of both the Knuth-Morris-Pratt and Boyer-Moore algorithms.

5.1.3 Hardware Text Search Systems

Q13. Write short notes on "Hardware text search systems".

Answer : **(Model Paper-I, Q10(b) | Dec.-14(R09), Q6(b)**

Hardware Text Search Algorithms

Hardware text search systems are developed to overcome the limitations of software text search systems that failed to handle multiple search terms concurrently against similar textual data. The major drawback of software text search systems is that they perform the search process with a limited I/O speed. One method to reduce the resource searching overhead is to utilize a specialized hardware machine. This machine is used to carry out the search process and then transfer the search result to actual computer that supports both user interfaces as well as hits retrieval. In hardware search machine, many search machines are executed parallely against the similar textual data so that more queries can be handled at one time. It also handles different textual data streams in order to reduce accessing time. The scalability issue is present in hardware text search systems which in turn increases the cost. The scalability of system can be attained by increasing the number of hardware devices. In such a system, the I/O speed is limited because of time factor, which specifies the amount of time taken for text to be transferred from disk drives to the searchers. The time taken to search a database is equivalent to the time taken to search a single disk when there is one search machine for individual disk drives. In addition to resource overhead reduction, another benefit of hardware text search system is that it eliminates the index that represents an item of a database.

Apart from these, the other advantages of hardware text search systems are,

1. The speed for searching newly received items is very fast and deterministic because the hardware system bypasses the waiting time required for creation of index.
2. It is easy for a user to predict the search time by estimating the time taken to perform data flow.
3. This system does not wait until the entire search process is completed i.e., it returns the hits as soon as they are identified.

Q14. Explain the importance of hardware text search unit in text streaming architecture.

Model Paper-III, Q10

OR

Explain fast data finder architecture.

(Refer Only Topic: Fast Data Finder (FDF))

Answer : **Oct./Nov.-20(R16), Q8(b)**

Text streaming architecture consists of both hardware and software text search systems. These systems algorithmetically concentrates on term detector. The following are the ways of implementing term detector.

(i) Parallel comparators (Associative memory)

(ii) Cellular structure

(iii) Universal FSA (Finite State Automata).

When the term comparator is implemented along with parallel comparator, every individual term present in the query is assigned to independent comparison component. Once all the input data is received, it is synchronously streamed into the term detector where the comparison is performed. When a term queried matches with any of the terms in the database, the term which is comparator is responsible to pass this matching information to external query resolver. It does this by setting the values of status flag.

The following are some of the hardware text search system units.

1. Parasel Searcher

It was developed in early 70's as the most recent product for searching disk drives. These searchers comprises of specialized hardware that interfaces with computers. The main purpose for using this specialized hardware is because of limited computer resources. The speed search in hardware search system is dependent on the I/O speed.

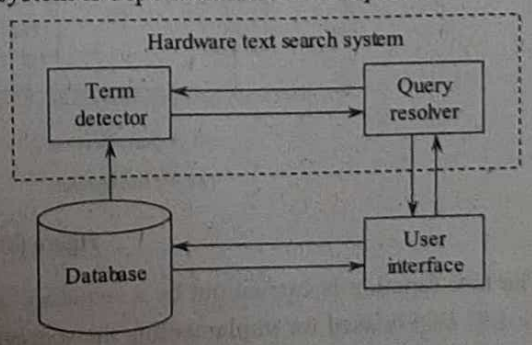


Figure (a): Hardware Text Search System

2. Rapid Search Machine

It is the earliest hardware text search system which was developed by General Electric. This machine is constructed by special purpose search unit. The unit is responsible for transferring a single query to the magnetic tapes that contains the document.

3. Associate File Processor (AFP)

It is another type of hardware search unit which has the ability to perform a search process by searching multiple queries concurrently.

4. High Speed Text Search (HSTS)

This machine is another hardware text search machine that employs an algorithm which is similar to the Aho-Corasick software finite state machine algorithm. But the difference is that the former algorithm executes three state machines concurrently. Out of these, the first machine is solely responsible for creating sequential word phrases, the second machine is responsible for matching imbedded term and the last machine is responsible for finding the exact word match.

5. GESCAN Text Array Processor (TAP)

GE modified the rapid search machine and renamed it as GESCAN unit. This system utilizes Text Array Processor (TAP). This processor is responsible for concurrently performing the match between the terms and conditions in accordance to the input data stream. Once the user enters the query, it is passed to TAP, which accesses the entered textual data from secondary disk drives. TAP processors comprises of large cache memory and an array of query processors that ranges from 4 to 128. The query processor itself is responsible for performing the search every time a text is stored in the cache. The entire query is handled by independent query processor which can be loaded any time. There are two operations that are performed by query processors simultaneously.

(i) To match every query term against the input text

(ii) To perform boolean logic resolution.

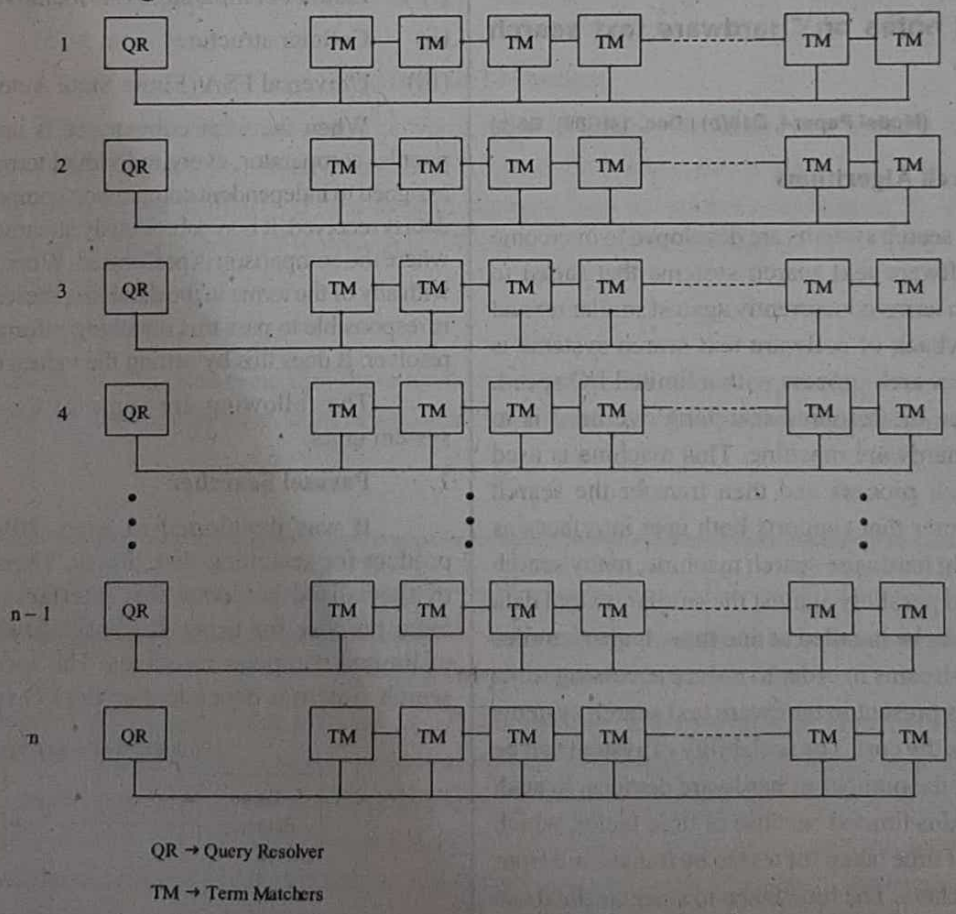


Figure (b): GESCAN Query Processor

The first function is carried out by a sequence of character cells. Every individual cell consists of a single character of the query. LSI chip is used for implementing the sequence of character cells. This chip is connected so as to form a sequence of longer size strings.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Whenever a match is found, a signal is transferred to the logic resolution subprocess available on the LSI chip. This resolution resolves the boolean logic between terms and proximity requirements. When the match is completely found, the result is sent back to the user's computer. The TAP processor make use of these chips by arranging them in matrix format, where each row specifies a query processor. The first chip in this processor is responsible for resolving query and the other remaining chips are responsible for performing the match between the query terms. The length of a row is limited by the maximum number of characters and the number of queries to be processed parallely is limited by the number of rows.

6. **CASSM (Content Addressable Segment Sequential Memory)**

It is another hardware system developed by university of Florida, which utilizes associative search elements simultaneously in order to extract data structure from the database. It can also be used to carry out the search associated with the string across the database.

7. **Fast Data Finder (FDF)**

It is another type of hardware text search system which is developed to perform textual search as well as english and foreign language search. The first FDF machine comprised of array of programming cells which are connected sequentially to form a pipeline hardware search processor. VLSI chips are employed for implementing these text processing cells. Individual chip consist of 24 processor cell and a usual integrated system generally consists of 3600 cells where each individual cell acts as a comparator. All cells are connected to each other by using data path and control path of 8-bits and 20-bits respectively. The searching in FDF is done in a pipeline format i.e., the text passes from one cell to another in the form of a pipe until the entire database is searched. Data analysis is performed at every cell due to which, the 20-bit control paths are modified depending on the existing state as well as the result received from the comparator.

Architecture of FDF System

A single cell is made up of register cell and a comparator. Once input is received from the database, it is controlled and stored by microprocessor and passed through the comparator. The characters which are searched are archived in the interconnected registers. This connection specifies the control paths which are passing the state information.

The query terms are identified by a set of cells together with a boolean logic. The identification is done by appropriately programming the control paths. The micro processor receives a hit, when a match is found. After receiving, the microprocessor is responsible for transferring it back to the host processor so that a user can easily access the matched item.

The following are the functions which are performed by FDF.

(i) Variable length "don't care" symbol
(ii) Proximity in an arbitrary pattern
(iii) Term counting and threshold
(iv) Fuzzy matching
(v) Term weight
(vi) Numeric ranges
(vii) Boolean logic resolution that include negation.

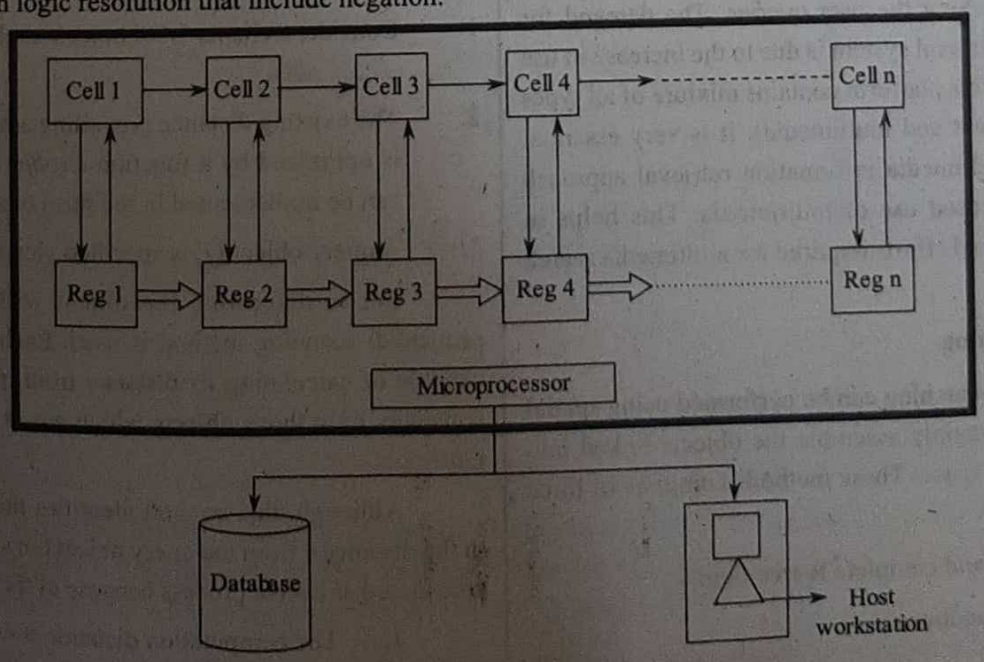


Figure (c): Architecture of Fast Data Finder

Parasel merged the searching capability of FDF with BTK (Biology Tool Kit). Using this combination search can be made between DNA's, proteins or between DNA and protein.

Smith-Waterman algorithm and sequence-profile algorithm can be implemented by configuring the FDF. The Smith-Waterman algorithm is a dynamic programming algorithm which is responsible for locating similarities between the local sequences FDF is implemented with a sequence which is reported back to the database only if the local similarity value is above the threshold value. The analysis process is completed by BTK software.

5.2 MULTIMEDIA INFORMATION RETRIEVAL

Q15. What is meant by multimedia information retrieval? Discuss about indexing and searching.

May/June-13(R09), Q7

OR

Discuss the significance of multimedia information retrieval. Discuss about indexing and searching. **Nov.-13(R09), Q7**

OR

Describe efficient indexing and searching in multimedia information retrieval.

Answer : **(Model Paper-II, Q10 | Dec.-14(R09), Q7)**

Multimedia Information Retrieval

Multimedia information retrieval refers to the criteria of fulfilling information based needs of the users. This criteria involves determination of text, graphics, pictures, documents etc., matching the user queries. The demand for such sophisticated retrieval system is due to the increase in use of world wide web. This platform contains mixture of all types of documents (i.e., text and multimedia). It is very essential to adopt efficient multimedia information retrieval approach because of the increased use of multimedia. This helps in minimizing the time and efforts required for multimedia search and retrieval.

Indexing and Searching

Indexing and searching can be performed using spatial access methods that mainly assemble the objects linked into points within the f-D space. These methods comprise of three classes. They are,

1. R* trees and complete R-tree family
2. Linear quadtrees
3. Grid-files.

1. R* Trees and Complete R-Tree Family

These trees are characteristic representation of the spatial access methods and describe a spatial object by using Minimum Body Rectangle (MBR). A tree hierarchy is formed by data rectangles which are clustered to form parent nodes and eventually these nodes are clustered again to form grand-parent nodes. However, the minimum bound rectangle of a parent node can be determined easily as it contains MBRs of its children entitled to intersect. The tree containing the nodes co-relates the disk-pages which are referred to as disk blocks. These blocks are the repeated byte positions which are typically present on the surface of the disk and can easily be obtained with a single disk access. Besides this, in order to achieve better clustering, the trees can adopt insertion, deletion and splitting operations with less parent nodes.

When the R-tree nodes fanout is less than 2, the methods can easily be implemented on higher-dimensions.

2. Linear Quadtrees

The time taken for searching via this method is directly proportional to the hyper surface of the query region. This hyper surface increases exponentially with respect to dimensionality.

3. Grid Files

This method uses directory which increases exponentially with respect to dimensionality.

Basic Idea Behind Generic Multimedia Indexing Approach

The basic idea behind the generic multimedia indexing approach can be explained by determining the objects which are with in the distance 'ε' from the query object. This can be done by considering "whole match" queries. These queries can be solved as follows,

1. Consider a cluster '*s*' of objects such as *obj*1, *obj*2, *obj*3 ..... *objs*.
2. The existing distance prevailing among the two objects is optimized by a function $D(obji, objj)$. This function can be implemented in the form of a software program.
3. A query object Q_0 is specified along with tolerance ε.

So, to determine the object with in the cluster, a sequential scanning method is used. Each *obji* ($1 \leq i \leq s$) is capable of calculating its distance from query object Q_0 and considers only those objects which are at distance of $D(Q_0, Obji) \leq \varepsilon$.

Although, this method identifies the objects which are in the distance ε from the query object but at the same time it is considered as a slow process because of the following reasons,

1. The computation distance is expensive
2. The size of the database is very large.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

5.2.1 Spoken Language Audio Retrieval

Q16. Discuss in brief about spoken language audio retrieval.

Model Paper-III, Q11(a)

Answer :

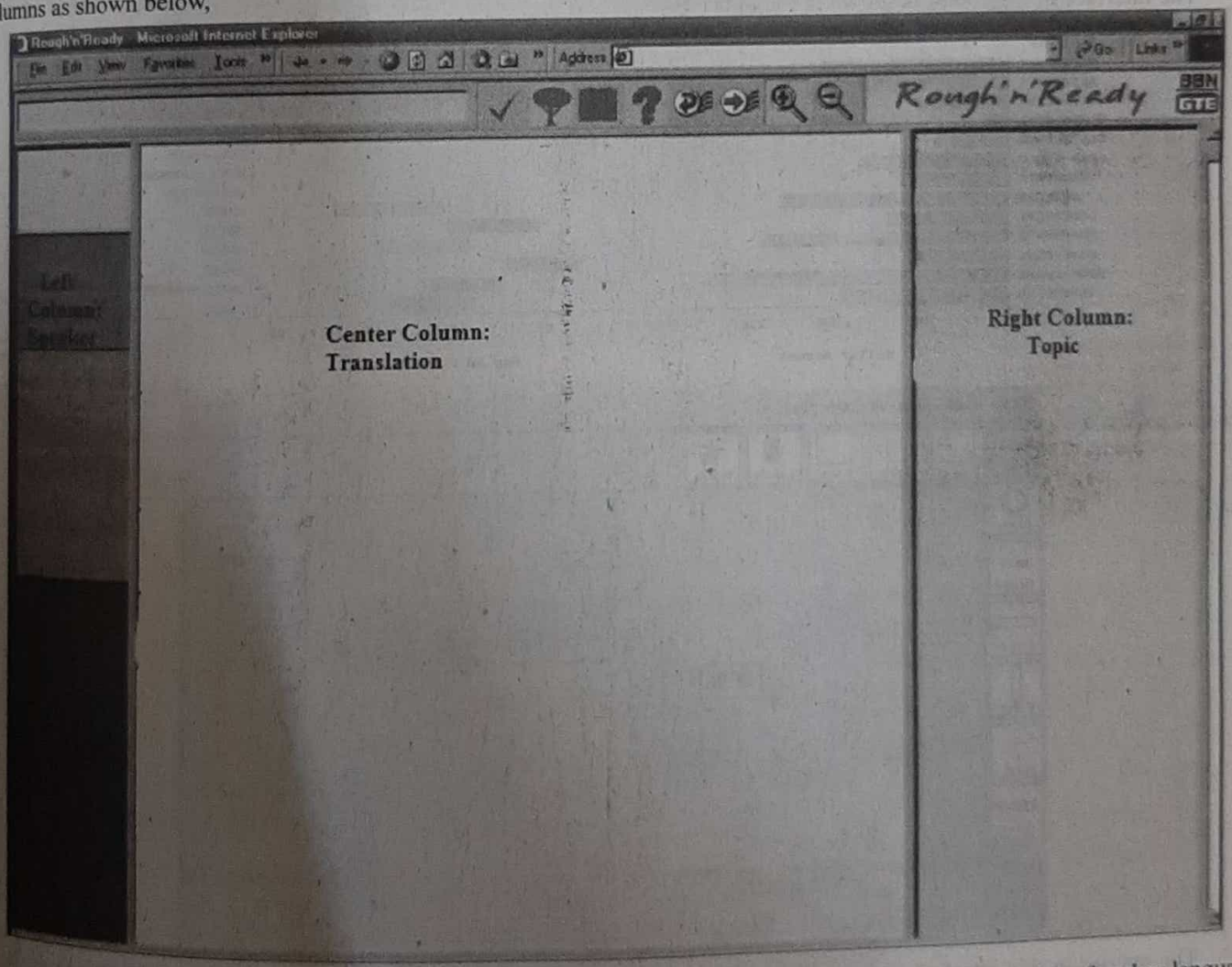
Spoken Language Audio Retrieval

Spoken language audio retrieval refers to the process of retrieving information from spoken data. One effective way of doing so is to create a system that recognizes the spoken words and convert the data into searchable text format. The sources of audio can be radio broadcasts, speeches, conversations and other programs. Certain applications are available that are capable of recognizing speech that can be used over speaker, transcription, command and control. However, the error rate for audio retrieval is approximately double the text retrieval. One example of such a recognition application is VMR (Video Mail Retrieval). The error of such applications can be minimized by employing redundancy in the source.

The challenges associated with recognition systems range from speaker dependency to time and effort involved in their development.

Rough 'n' Ready Interface

The Rough 'n' Ready interface is one of the application of spoken language recognition from audio and video sources. The output of this system is a summarization which can be searched or browsed instantly. This interface arranges the data in three columns as shown below,



As the figure shows that the left column carries speaker details, the middle column carries the translation of spoken language and the right column carries the topic details. This interface makes use of a dictionary whose range of words (Vocabulary) is around 60,000. Its error rate for a broadcast news task is recorded as 18.8%.

An added feature in spoken language recognition is the use of multiple languages for accesses information. An example of such a system is Gavvain whose error rate is 13.6 and works well with German and French languages.

5.2.2 Non-speech Audio Retrieval, Graph Retrieval

Q17. Discuss in brief about non-speech audio retrieval and graph retrieval.

Answer : Model Paper-II, Q11(a)

Non-speech Audio Retrieval

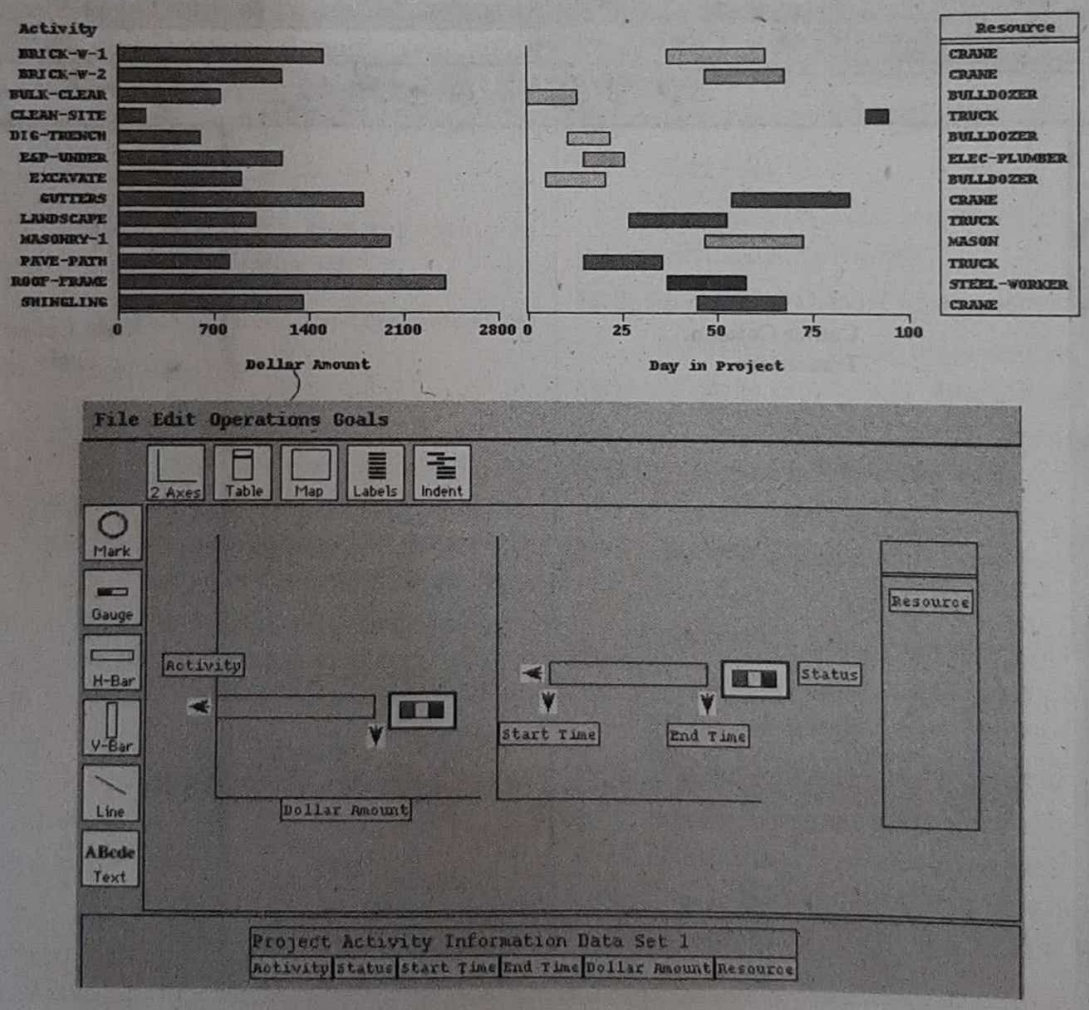
Non-speech audio retrieval refers to the process of retrieving sound other than speech like music, noise etc. Such retrieval plays a major role in fields like movie production or music industry. An example interface supporting non-audio retrieval is SoundFisher developed by Thom Blum et al. (1997). This interface was capable of retrieving signal processing, music and multimedia data apart from speech recognition. It utilizes the features of audio like loudness, pitch, duration etc., for obtaining the index of sound files. Through this, users can search for required sound by forwarding a query like "Search for files encoded with AIFF having vocal sounds similar to barking sounds". Moreover, system can also be molded to consider the perceptual properties also.

The performance of such systems can be evaluated by considering sound files carrying sounds extracted from different sources like animals, instruments, nature, speeches etc.

Graph Retrieval

Graph retrieval refers to the process of retrieving data carrying tables, charts etc., which are drawn using points, lines and labels. An example interface of such a retrieval is Sagebook developed at Carnegie Mellon University. It is capable of allowing its users to search and customize the graphic data. It offers features like indexing, searching, representing and others.

The interface of Sagebook is shown in the figure below,



The queries in this interface can be generated using SageBrush where it displays a set of options like objects, colours, spaces etc. The results can be filtered based on exact matches and similar matches. The result (object) generated by the query can be customized accordingly.

SageBook follows certain syntax and semantics like relationship existing among objects, attributes etc., for effective retrieval. The criteria follows multiple alternatives to identify matches of generated query. Moreover, irrelevant matches can be filtered out automatically.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

5.2.3 Imagery Retrieval, Video Retrieval

Q18. Write short notes on Imagery Retrieval.

Answer :

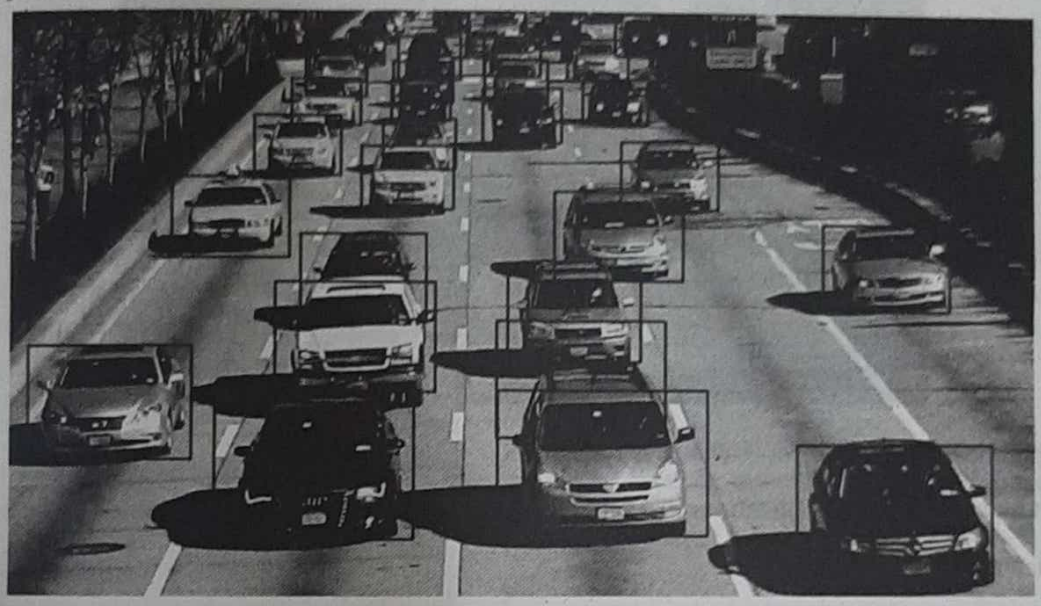
Content based image retrieval is the process of detecting images based on colour, shape and texture (automatic indexing) rather than annotations and captions (manual indexing).

QBIC

QBIC (Query By Image Content) uses multiple features like color parameters, texture selection and representation of shapes for retrieving an image. QBIC facilitates large data sets for better results.

Example

Retrieving specific coloured images using QBIC.

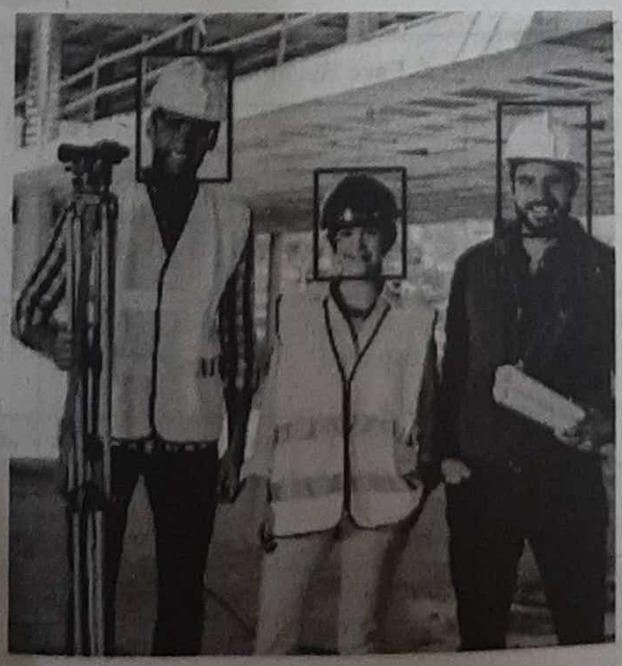


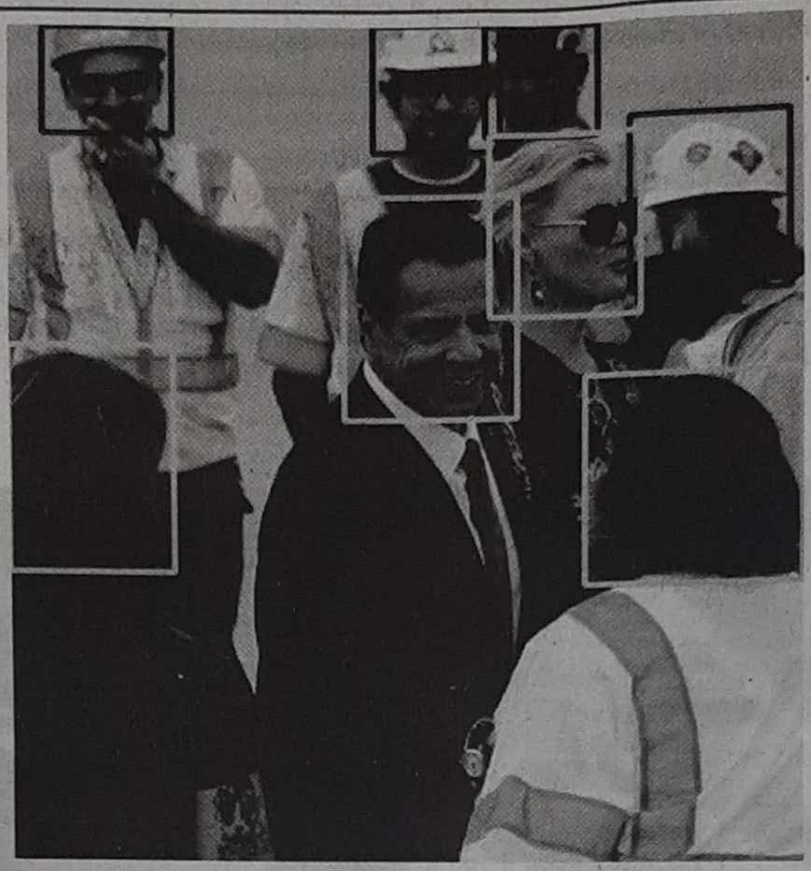
The above image explains, how images can be retrieved based on colour.

Various parameters can be added to QBIC for getting desired output. Though object identification is a difficult process, tools for scanning the background and foreground are developed.

Example

Helmet detection.





Above images show, how an object (Helmet) is recognized.

With respect to wide range of applications, the problems of content based image retrieval led to the concepts of,

(i) Face detection: Identifying a face

(ii) Face recognition : Authenticating a detected face.

(iii) Face retrieval : Matching the detected face

(iv) Emotion recognition : Identifying facial expressions.

Q19. Distinguish between face detection, face recognition and a face retrieval.

Answer :

(Model Paper-II, Q11(b) | Dec.-19(R16), Q11(b))

	Face Detection		Face Recognition		Face Retrieval
1.	Face detection is the process of finding a face in any scenario.	1.	Face recognition is the process of finding the database of a face that is detected.	1.	Face retrieval is the process of finding the match for a face that is recognized.
2.	It is done using cameras or surveillance.	2.	It is done using radio frequency tags.	2.	It is done using the database system.
3.	It does not considers the features of face to be detected.	3.	It considers the features of face to be detected.	3.	It takes the features as input and generates the output based on given features.
4.	It counts the number of persons in an image without their identification.	4.	It counts the number of persons in an image and identifies them.	4.	It does not count the number of persons in the image but goes through the physical features of all the faces.
5.	Example: Phone camera used for clicking pictures.	5.	Example: Face recognition based lock system of phone.	5.	Example: Image Searching by name on phones Gallery.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Q20. Explain in detail about video retrieval. Model Paper-III, Q11(b)

OR

List three applications for content based video retrieval. Dec.-19(R16), Q11(a)

Answer :

Content based video retrieval is the process of searching, analyzing and retrieving information from video mails, surveillance videos, broadcast television etc. Some applications of content based video retrieval are as follows,

1. **BNN**

BNN (Broadcast News Navigator) provides multiple features of capturing, integrating (of text, image and speech) analyzing and visualizing the video content for retrieval of data. BNN is used for broadcast news video.

BNN query page allows the user to select and search news based on data range, closed captions, transcriptions, keywords and named entities. Using the data provided, BNN creates a custom query page offering parameters from extracted content. This simplifies the work of query formulation for a user.

BNN can also be used for simple story browsing with a time constraint. Data mining is also possible for named entities in the analyzed stories.

The results from the evaluation test of BNN claim that BNN is fast and effective when searched using named entities and keywords. Automated segmentation plays a vital role in BNN's performance. The significance of keyframes selection and quality source is explained by the evaluation test. However, BNN struggles with the issues of the time-consumption and manual video annotations.

2. **TDT**

TDT (Topic Detection and Tracking) for newswire sources and broadcast news is used for introducing new algorithm including,

(i) Story Segmentation : It identifies story boundaries.

(ii) Topic Tracking : It identifies stories that discuss the topics related to target.

(iii) Topic Detection : It identifies stories that discuss an arbitrary topic.

3. **GeoNODE**

GeoNODE(Geospatial News On Demand Environment) is used for topic detection and tracking. It produces output in temporal and geospatial context and is used for extracting information, visualizing components and mining the data. This gives the data that matches with desired information from different geographical locations. GeoNODE is capable of extracting information from various sources like audio, video and text.