LECTURE NOTES

UNIT-I

Introduction: Definition, objectives, Functional Overview, Relationship to DBMS, Digital libraries and Data Warehouses. **Information Retrieval System Capabilities:** Search, Browse, Miscellaneous Capabilities.

Write about Information System?

There is a potential for confusion in the understanding of the differences between Database Management Systems (DBMS) and Information Retrieval Systems. It is easy to confuse the software that optimizes functional support of each type of system with actual information or structured data that is being stored and manipulated. The importance of the differences lies in the inability of a data base management system to provide the functions needed to process "information." The opposite, an information system containing structured data, also suffers major functional deficiencies.

1. Definition of Information Retrieval System

An Information Retrieval System is a system that is capable of storage, retrieval, and maintenance ofinformation6. Information in this context can be composed of text (including numeric and date data), images, audio, video and other multi-media objects. Techniques are beginning to emerge to search these other media types.

The term "item" is used to represent the smallest complete unit that is processed and manipulated by the system. The definition of item varies by how a specific source treats information. A complete document, such as a book, newspaper or magazine could be an item. For example, a video news program could be considered an item. It is composed of text in the form of closed captioning, audio text provided by the speakers, and the video images being displayed.

An Information Retrieval System consists of a software program that facilitate user in finding the information the user needs. The system may use standard computer hardware or specialized hardware to support the search sub function and to convert non-textual sources to a searchable media Thus search composition, search execution, and reading non-relevant items are all aspects of information retrieval overhead. With the advent of inexpensive powerful personnel computer processing systems and high speed, large capacity secondary storage products, it has become commercially feasible to provide data.

The introduction exponential the algorithms and techniques to optimize the processing and access of large quantities of textual data were once the sole domain of segments of the Government, a few industries, and academics. Images across the Internet are searchable from many websites such as WEBSEEK, DITTO.COM, ALTAVISTA /IMAGES.

Growth of the Internet along with its initial WAIS (Wide Area Information Servers) capability and more recently advanced search servers (e.g., INFOSEEK, EXCITE) has provided a new avenue for access to terabytes of information (over800million indexablepages-Lawrence-99.)
News organizations such as the BBC are processing the audio news they have produced and are making historical audio news searchable via the audio transcribed versions of the news. Major video organizations such as Disney are using video indexing to assist in finding specific images in their previously produced videos to use in future videos or incorporate in advertising.

2. Objectives of Information Retrieval Systems.

The general objective of an Information Retrieval System is to minimize the overhead of a user locating needed information. Overhead can be expressed as the time a user spends in all of the steps leading to reading an item containing the needed information (e.g., query generation, query execution, scanning results of query to select item stored, reading non-relevant items).
In information retrieval the term "relevant" item is used to represent an item containing the needed information. From a user's perspective "relevant" and "needed" are synonymous.
The two major measures common only associated with information systems are

1) Precision
2) Recall

When a user decides to issue a search looking for information on a topic, the total database is logically divided into four segments Relevant items are those documents that contain information that helps the searcher in answering his question. Non-relevant items are those items that do not provide any directly useful information. There are two possibilities with respect to each item: it can be retrieved or not retrieved by the user's query. Precision and recall are defined as Precision is directly affected by retrieval of non-relevant items and drops to a number close to zero. Recall is not affected by retrieval of non-relevant items and thus remains at 100 percent once achieved.
Information Retrieval Systems such as Retrieval Ware, TOPIC, AltaVista, Info seek and INQUERY that the idea of accepting natural language queries is becoming a standard system

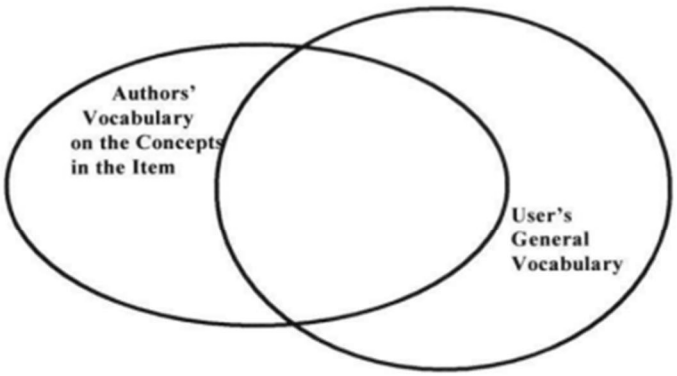

Figure 1.3 Vocabulary Domains

Information Retrieval Systems such as Retrieval Ware, TOPIC, AltaVista, Info seek and INQUERY that the idea of accepting natural language queries is becoming a standard system feature. This allows users to state in natural language what they are interested in finding. But the completeness of the user specification is limited by the user's willingness to construct long natural language queries. Most users on the Internet enter one or two search terms.

3. Functional Overview

A total Information Storage and Retrieval System is composed of four major functional processes:

1) Item Normalization
2) Selective Dissemination of Information (i.e., "Mail")
3) Archival Document Data Base Search, and an Index
4) Database Search along with the Automatic File Build process that supports Index Files

Item Normalization:

The first step in any integrated system is to normalize the incoming items to a standard format. Item normalization provides logical restructuring of the item. Additional operations during item normalization are needed to create a searchable data structure: identification of processing tokens (e.g., words), characterization of the tokens, and stemming (e.g., removing word endings) of the tokens. The processing tokens and their characterization are used to define the searchable text from the total received text. Figure 1.5 shows the normalization process. Standardizing the input takes the different external formats of input data and performs the translation to the formats.

A system may have a single format for all items or allow multiple formats. One example of standardization could be translation of foreign languages into Unicode. Every language has a different internal binary encoding for the characters in the language. One standard encoding that covers English, French, Spanish, etc... is ISO-Latin.

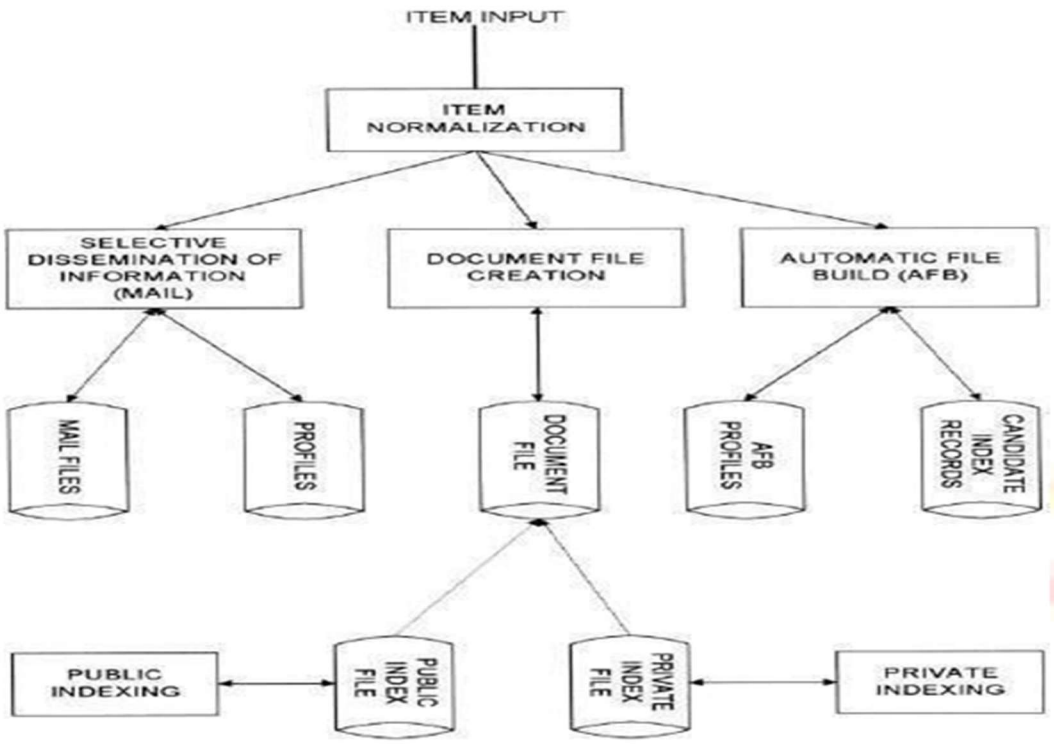

Figure 1.4 Total Information Retrieval System

To assist the users in generating indexes, especially the professional indexers, the system provides a process called Automatic File Build *(*AFB).

Multi-media adds an extra dimension to the normalization process. In addition to normalizing the text input, the multi-media input also needs to be standardized. There are a lot of options to the standards being applied to the normalization. If the input is video the likely digital standards will be either MPEG-2, MPEG-1, AVI or Real Media. MPEG (Motion Picture Expert Group) standards are the most universal standards for higher quality video where Real Media is the most common standard for lower quality video being used on the Internet. Audio standards are typically WAV or Real Media (RealAudio). Images vary from JPEG to BMP.

The next process is to parse the item into logical sub-divisions that have meaning to the user. This process, called "Zoning," is visible to the user and used to increase the precision of a search and optimize the display. A typical item is sub- divided into zones, which may overlap. References. The zoning information is passed to the processing token identification operation to store and can be hierarchical, such as Title, Author, Abstract, Main Text, Conclusion, and the information, allowing searches to be restricted to a specific zone. For example, if the user is interested in Articles discussing "Einstein" then the search should not include the Bibliography. Which Could include references to articles written by "Einstein" Systems determine words by dividing input symbols into 3 classes:

1) Valid word symbols
2) Inter-word symbols
3) Special processing symbols.

A word is defined as a contiguous set of word symbols bounded by inter-word symbols. In many systems inter-word symbols are non-searchable and should be carefully selected. Examples of word symbols are alphabetic characters and numbers. Examples of possible inter-word symbols are blanks, periods and semicolons. The exact definition of an inter-word symbol is dependent upon the aspects of the language domain of the items to be processed by the system. For example, an apostrophe may be of little importance if only used for the possessive case in English, but might be critical to represent foreign names in the database.

Next, a Stop List/Algorithm is applied to the list of potential processing tokens. The objective of the Stop function is to save system resources by eliminating from the set of searchable processing tokens those that have little value to the system. Given the significant increase in available cheap memory, storage and processing power, the need to apply the Stop function to processing tokens is decreasing. Examples of Stop algorithms are: Stop all numbers greater than "999999" (this was selected to allow dates to be searchable) Stop any processing token that has numbers and characters inter mixed.

2) Selective Dissemination (Distribution, Spreading) of Information

The Selective Dissemination of Information (Mail) Process provides the capability to dynamically compare newly received items in the information system against standing statements of interest of users and deliver the item to those users whose statement of interest matches the contents of the item. The Mail process is composed of the search process, user statements of interest (Profiles) and user mail files. As each item is received, it is processed against every user's profile. A profile

contains atypically broad search statement along with a list of user mail files that will receive the

document if the search statement in the profile is satisfied. Selective Dissemination of Information has not yet been applied to multimedia sources.

3) Document Data base Search

The Document Database Search Process provides the capability for a query to search against all items received by the system. The Document Database Search process is composed of the search process, user entered queries (typically ad hoc queries) and the document database which contains all items that have been received, processed and stored by the system. Typically, items in the Document Databased on change (i.e., are noted it) once received.

Index Data base Search

When an item is determined to be of interest, a user may want to save it for future reference. This is in effect filing it. In an information system this is accomplished via the index process. In this process the user can logically store an item in a file along with additional index terms and descriptive text the user wants to associate with the item. The Index Database Search Process (seeFigure1.4) provides the capability to create indexes and search them.

There are 2 classes of index files:

1) Public Index files
2) Private Index files

Every user can have one or more Private Index files leading to a very large number of files. Each Private Index file references only a small subset of the total number of items in the Document Database. Public Index files are maintained by professional library services personnel and typically index every item in the Document Database. There is a small number of Public Index files. These files have access lists (i.e., lists of users and their privileges) that allow anyone to search or retrieve data. Private Index files typically have very limited access lists. To assist the users in generating indexes, especially the professional indexers, the system provides a process called Automatic File Build shown in Figure1.4 (also called Information Extraction).

Multimedia Database Search

From a system perspective, the multi-media data is not logically its own data structure, but an augmentation to the existing structures in the Information Retrieval System.

4. Relationship to Data base Management Systems

From a practical standpoint, the integration of DBMS"s and Information Retrieval Systems is very important. Commercial database companies have already integrated the two types of systems. One of the first commercial databases to integrate the two systems into a single view is INQUIREDBMS. This has been available for over fifteen years. A more current example is the ORACLE DBMS that now offers an imbedded capability called CONVECTIS, which is an informational retrieval system that uses a comprehensive thesaurus which provides the basis to generate "themes" for a particular item. The INFORMIX DBMS has the ability to link to Retrieval Ware to provide integration of structured data and information along with functions associated with Information Retrieval Systems.

Digital Libraries and Data Ware houses (Data Marts)

As the Internet continued its exponential growth and project funding became available, the topic of Digital Libraries has grown. By 1995 enough research and pilot efforts had started to support the 1STACM International Conference on Digital Libraries (Fox96). Indexing is one of the critical disciplines in library science and significant effort has gone into the establishment of indexing and cataloging standards. Migration of many of the library products to a digital format introduces both opportunities and challenges. Information Storage and Retrieval technology has addressed a small subset of the issues associated with Digital Libraries.

Data warehouses are similar to information storage and retrieval systems in that they both have a need for search and retrieval of information. But a data warehouse is more focused on structured data and decision support technologies. In addition to the normal search process, a complete system provides a flexible set of analytical tools to "mine" the data. Data mining (originally called Knowledge Discovery in Databases -KDD) is a search process that automatically analyzes data and extract relationships and dependencies that were not part of the data base design.

Information Retrieval System Capabilities

1. Search Capabilities
2. Browse Capabilities
3. Miscellaneous Capabilities

The search capabilities address both Boolean and Natural Language queries. The algorithms used for searching are called Boolean, natural language processing and probabilistic. Probabilistic algorithms use frequency of occurrence of processing tokens (words) in determining similarities

between Queries and items and also in predictors on the potential relevance of the found item to the searcher.

Queries and items and also in predictors on the potential relevance of the found item to the searcher. The newer systems such as TOPIC, Retrieval Ware, and INQUERY all allow for natural language queries. Browse functions to assist the user in filtering the search results to find relevant information are very important.

Search Capabilities

The objective of the search capability is to allow for a mapping between a user's specified need and the items in the information database that will answer that need. It can consist of natural language text in composition style and/or query terms (referred to as terms in this book) with Boolean logic indicators between them. One concept that has occasionally been implemented in commercial systems (e.g., Retrieval Ware), and holds significant potential for assisting in the location and ranking of relevant items, is the "weighting" of search terms. This would allow a user to indicate the importance of search terms in either a Boolean or natural language interface. Given the following natural language query statement where the importance of a particular search term is indicated by a value in parenthesis between 0.0 and 1.0 with 1.0 being the most important.
The search statement may apply to the complete item or contain additional parameters limiting it to a logical division of the item (i.e., to a zone). Based upon the algorithms used in a system many different functions are associated with the system's understanding the search statement. The functions define the relationships between the terms in the search statement (e.g., Boolean, Natural Language, Proximity, Contiguous Word Phrases, and Fuzzy Searches) and the interpretation of a particular word (e.g., Term Masking, Numeric and Date Range, Contiguous Word Phrases, and Concept/Thesaurus expansion).

Boolean Logic

Boolean logic allows a user to logically relate multiple concepts together to define what information is needed. Typically, the Boolean functions apply to processing tokens identified anywhere within an item. The typical Boolean operators are **AND, OR,** and **NOT**.

These operations are implemented using set intersection, set union and set difference procedures. A search terms in either a Boolean or natural language interface. Given the following natural language query statement where the importance of a particular search term is indicated by a value in parenthesis between 0.0 and1.0 with 1.0 being the most important.

The search statement may apply to the complete item or contain additional Parame search terms in either a Boolean or natural language interface. Given the following natural language query Statement where the importance of a particular search term is indicated by a value in parenthesis between 0.0 and 1.0 with1.0 being them important.

The search statement may apply to the complete item or contain additional parameters limiting it to a logical division of the item (i.e., to a zone). Based upon the algorithms used in a system many different functions are associated with the system's understanding the search statement. The functions define the relationships between the terms in the search statement (e.g., Boolean, Natural Language, Proximity, Contiguous Word Phrases, and Fuzzy Searches) and the interpretation of a particular word (e.g., Term Masking, Numeric and Date Range, Contiguous Word Phrases, and Concept/Thesaurus expansion). limiting it to a logical division of the item (i.e., to a zone). Based upon the algorithms used in a system many different functions are associated with the system's understanding the search statement. The functions define the relationships between the terms in the search statement (e.g., Boolean, Natural Language, Proximity, Contiguous Word Phrases, and Fuzzy Searches) and the inter pretation of a particular word (e.g., Term Masking, Numeric and Date Range, Contiguous Word Phrases, And Concept / Thesaurus expansion).few systems introduced the concept of exclusive or but it is equivalent to a slightly more complex query using the other operators and is not generally useful trousers since most users do not understand it.

A special type of Boolean search is called "M of N" logic. The user lists a set of possible searches terms and identifies, as acceptable, any item that contains a subset of the terms.For example, "Find any item containing any two of the following terms: "AA," "BB," "CC." This can be expanded into a Boolean search that performs an AND between all combinations of two terms and "OR" s the results together ((AA AND BB) or (AA AND CC) or (BB AND CC)).

Proximity

Proximity is used to restrict the distance allowed within an item between two search terms. The semantic concept is that the close search terms in either a Boolean or natural language interface. Given the following natural language query statement where the importance of a particular search term is indicated by a value in parenthesis between 0.0 and 1.0 with 1.0 being the most important. The search statement may apply to the complete item or contain additional parameters limiting it to a logical division of the item (i.e., to a zone). Based upon the algorithms used in a system many different functions are associated with the system's understanding the search statement.

The functions define the relationships between the terms in the search statement (e.g., Boolean, Natural Language, Proximity, Contiguous Word Phrases, and Fuzzy Searches) and the interpretation of a particular word (e.g., Term Masking, Numeric and Date Range, Contiguous Word Phrases, and Concept/Thesaurus expansion). Two terms are found in a text the more likely they are related in the description of a particular concept. Proximity is used to increase the precision of a search. If the terms COMPUTER and DESIGN are found within a few words of each other then the item is more likely to be discussing the design of computers than if the words are paragraphs a part.

The typical format for proximity is TERM1 with in "m" "units" of TERM The distance operator "m" is an integer number and units are in Characters, Words, Sentences, or Paragraphs.

SEARCH STATEMENT	SYSTEM OPERATION
COMPUTER OR PROCESSOR NOT MAINFRAME	Select all items discussing Computers and/or Processors that do not discuss Mainframes
COMPUTER OR (PROCESSOR NOT MAINFRAME)	Select all items discussing Computers and/or items that discuss Processors and do not discuss Mainframes
COMPUTER AND NOT PROCESSOR OR MAINFRAME	Select all items that discuss computers and not processors or mainframes in the item

Figure 2.1 Use of Boolean Operators

A special case of the Proximity operator is the Adjacent (ADJ) operator that normally has a distance operator of one and a forward only direction (i.e., in WAIS). Another special case is where the distance is set to zero meaning with in the same semantic unit.

A special case of the Proximity operator is the Adjacent (ADJ) operator that normally has a distance operator of one and a forward only direction (i.e., in WAIS). Another special case is where the distance is set to zero meaning with in the same semantic unit.

Contiguous Word Phrases

A Contiguous Word Phrase (CWP) is both a way of specifying a query term and a special search operator. A Contiguous Word Phrase is two or more words that are treated as a single semantic unit. An example of a CWP is "United States of America." It is four words that specify a search term representing a single specific semantic concept (a country) that can be used with any of the Operators discussed above. Thus a query could specify "manufacturing" AND "United States of America" which returns any item that contains the word "manufacturing" and the contiguous words "United States of America."

A contiguous word phrase also acts like a special search operator that is similar to the proximity (Adjacency) operator but allows for additional specificity. If two terms are specified, the contiguous word phrase and the proximity operator using direction alone word parameters or the Adjacent operator are identical. For contiguous word phrases of more than two terms the only way of creating an equivalent search statement using proximity and Boolean operators is via nested Adjacencies which are not found in most commercial systems. This is because Proximity and Boolean operators are binary operators but contiguous word phrases are an "N" operator where"N" is the number of words in the CWP.

Contiguous Word Phrases are called Literal Strings in WAIS and Exact Phrases in Retrieval Ware. In WAIS multiple Adjacency (ADJ) operators are used to define a Literal String (e.g., "United" ADJ "States" ADJ "of" ADJ "America").

SEARCH STATEMENT	SYSTEM OPERATION
"Venetian" ADJ "Blind"	would find items that mention a Venetian Blind on a window but not items discussing a Blind Venetian
"United" within five words of "American"	would hit on "United States and American interests," "United Airlines and American Airlines" not on "United States of America and the American dream"
"Nuclear" within zero paragraphs of "clean-up"	would find items that have "nuclear" and "clean-up" in the same paragraph.

Figure 2.2 Use of Proximity

Fuzzy Searches

Fuzzy Searches provide the capability to locate spellings of words that are similar to the entered search term. This function is primarily used to compensate for errors in spelling of words. Fuzzy searching increases recall at the expense of decreasing precision (i.e., it can erroneously identify terms as the search term). In the process of expanding a query term fuzzy searching includes other

terms that have similar spellings, giving more weight (in systems that rank output) to words in the database that have similar word lengths and position of the characters as the entered term. A Fuzzy Search on the term "computer" would automatically include the following Words from the information database: "computer", "compiter", "computer", "computter", "compute".

Term Masking

Term masking is the ability to expand a query term by masking a portion of the term and accepting as valid any processing token that maps to the unmasked portion of the term. The value of term masking is much higher in systems that do not perform stemming or only provide a very simple stemming algorithm. There are two types of search term masking: fixed length and variable length. Sometimes they are called fixed and variable length "don't care" functions. Fixed length masking is a single position mask. It masks out any symbol in a particular position or the lack of that position in a word. Variable length "don't cares" allows masking of any number of characters within a processing token. The masking may be in the front, at the end, at both front and end, or imbedded. The first three of these cases are called suffix search, prefix search and imbedded character string search, respectively. The use of an imbedded variable length don't care is seldom used .Figure 2.3 provides examples of the use of variable length term masking. If "*" represents a variable length don't care then the following are examples of its use:

"*COMPUTER" Suffix Search

"COMPUTER*" Prefix Search

"*COMPUTER*" Imbedded String Search

SEARCH STATEMENT	SYSTEM OPERATION
multi$national	Matches"multi-national," "multinational," "multinational" but does not match "multi national" since it is two processing tokens.
computer	Matches,"minicomputer" "microcomputer" or "computer"
comput*	Matches "computers," "computing," "computes"
comput	Matches "microcomputers" , "minicomputing," "compute"

Figure 2.3 Term Masking

DESIGN arefound within a few words of each other then the item is more likely to be discussing the design ofcomputersthaniftheWordsareparagraphsapart.Thetypicalformatforproximityis:

TERM1within"m""units"ofTERM2

The distance operator "m" is an integer number and units are in Characters, Words, Sentences, orParagraphs.

SEARCH STATEMENT	SYSTEM OPERATION
COMPUTER OR PROCESSOR NOT MAINFRAME	Select all items discussing Computers and/or Processors that do not discuss Mainframes
COMPUTER OR (PROCESSOR NOT MAINFRAME)	Select all items discussing Computers and/or items that discuss Processors and do not discuss Mainframes
COMPUTER AND NOT PROCESSOR OR MAINFRAME	Select all items that discuss computers and not processors or mainframes in the item

Figure 2.1 Use of Boolean Operators

A special case of the Proximity operator is the Adjacent (ADJ) operator that normally has a distanceoperator of one and a forward only direction (i.e., in WAIS). Anotherspecial case is where thedistanceissettozeromeaningwithinthesamesemanticunit.

ContiguousWordPhrases

A Contiguous Word Phrase (CWP) is both a way of specifying a query term and a special searchoperator. A Contiguous Word Phrase is two or more words that are treated as a single semantic unit.An example of a CWP is "United States of America." It is four words that specify a search termrepresenting a single specific semantic concept (a country) that can be used with any of the operatorsdiscussedabove.Thusaquerycouldspecify"manufacturing"AND"UnitedStatesofAmerica" whichreturns any item that contains the word "manufacturing" and the contiguous words "United States ofAmerica."

Acontiguouswordphrase also acts likeaspecialsearchoperatorthatissimilar to theproximity(Adjacency)operatorbutallowsforadditionalspecificity.

If two terms are specified, the contiguous word phrase and the proximity operator using directionalone word parameters or the Adjacent operator are identical.For contiguous word phrases of morethan two terms the only way of creating an equivalent search statement using proximity and

BooleanoperatorsisvianestedAdjacencieswhicharenotfoundinmostcommercialsystems.Thisisbecause Proximity and Boolean operators are binary operators but contiguous word phrases are an "N"aryoperatorwhere"N"isthenumber ofwordsintheCWP.

Contiguous Word Phrases are called Literal Strings in WAIS and Exact Phrases in RetrievalWare. InWAIS multiple Adjacency (ADJ) operators are used to definea Literal String (e.g., "United" ADJ"States"ADJ"of"ADJ"America").

SEARCH STATEMENT	SYSTEM OPERATION
"Venetian" ADJ "Blind"	would find items that mention a Venetian Blind on a window but not items discussing a Blind Venetian
"United" within five words of "American"	would hit on "United States and American interests," "United Airlines and American Airlines" not on "United States of America and the American dream"
"Nuclear" within zero paragraphs of "clean-up"	would find items that have "nuclear" and "clean-up" in the same paragraph.

Figure 2.2 Use of Proximity

FuzzySearches

Fuzzy Searches provide the capability to locate spellings of words that are similar to the enteredsearch term. This function is primarily used to compensate for errors in spelling of words. Fuzzysearchingincreasesrecallatthe expense ofdecreasingprecision(i.e.,itcanerroneouslyidentifytermsasthesearchterm).Intheprocessofexpandingaquerytermfuzzysearchingincludesothertermsthathave similar spellings, giving more weight (in systems that rank output) to words in the database thathave similar word lengths and position of the characters as the entered term. A Fuzzy Search on theterm"computer" wouldautomaticallyincludethe following

Wordsfromtheinformationdatabase:"computer,""compiter,""conputer,""computter,""compute."

TermMasking

Term masking is the ability to expand a query term by masking a portion of the term and accepting asvalid any processing token that maps to the unmasked portion of the term. The value of term maskingismuchhigherinsystemsthatdonotperform stemmingoronlyprovideaverysimplestemmingalgorithm. There are two types of search term masking: fixed length and variable length. Sometimestheyarecalledfixedand variablelength "don"tcare"functions.

Fixedlengthmaskingisa singlepositionmask.Itmasksoutanysymbolinaparticularpositionorthelack of that position in a word. Variable length "don"t cares" allows masking of any number ofcharacters within a processing token. The masking may be in the front, at the end, at both front andend, or imbedded. The first three of these cases are called suffix search, prefix search and imbeddedcharacter string search, respectively. The use of an imbedded variable length don"t care is seldomused.Figure 2.3 provides examples of the use of variable length term masking. If "*" represents avariablelengthdon"tcarethenthefollowingareexamplesofitsuse:

"*COMPUTER" Suffix Search"COMPUTER*" Prefix Search"*COMPUTER*"ImbeddedStringSearch

SEARCH STATEMENT	SYSTEM OPERATION
multi$national	Matches"multi-national," "multinational," "multinational" but does not match "multi national" since it is two processing tokens.
computer	Matches,"minicomputer" "microcomputer" or "computer"
comput*	Matches "computers," "computing," "computes"
comput	Matches "microcomputers" , "minicomputing," "compute"

Figure 2.3 Term Masking

NumericandDateRanges

Term masking is useful when applied to words, but does not workfor finding ranges of numbers ornumeric dates. To find numbers larger than "125," using a term "125*" will not find any numberexceptthose thatbeginwiththe digits"125."

Concept/ThesaurusExpansion

Associated with both Boolean and Natural Language Queries is the ability to expand the search termsvia Thesaurus or Concept Class database reference tool. A Thesaurus is typically a one-

level or two-level expansion of a term to other terms that are similar in meaning. A Concept Class is a treestructure that expands each meaning of a word into potential concepts that are related to the initialterm (e.g., in the TOPIC system). Concept classes are sometimes implemented as a network structurethat links word stems (e.g., in the RetrievalWare system). An example of Thesaurus and ConceptClassstructuresareshownin Figure2.4 (Thesaurus-93)and Figure2.5.

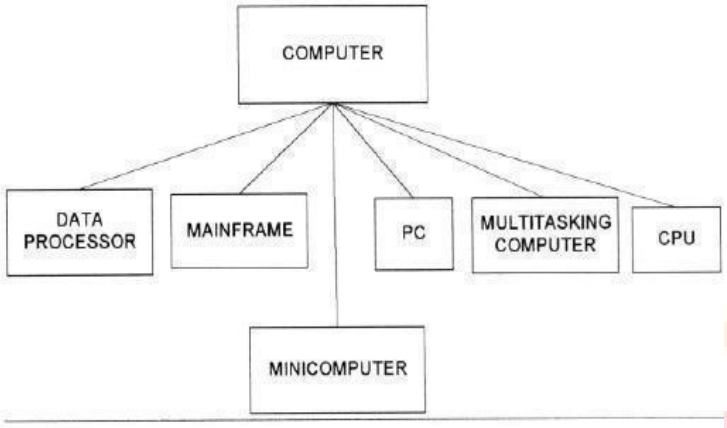

Figure 2.4 Thesaurus for term "computer"

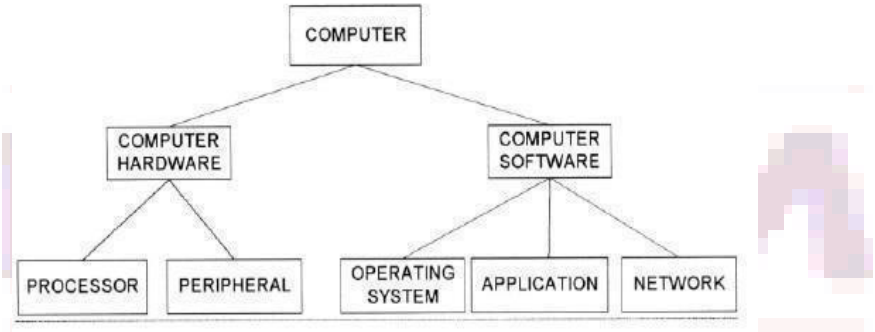

Figure 2.5 Hierarchical Concept Class Structure for "Computer"

Thesauri are either semantic or based upon statistics. A semantic thesaurus is alisting of words andthen otherwordsthataresemanticallysimilar.

The problem with thesauri is that they are generic to a language and can introduce many search termsthat are not found in the document database. An alternative uses the database or a representativesample of it to create statistically related terms. It is conceptually a thesaurus in that words that arestatistically related to other words by their frequently occurring together in the same items. This typeof thesaurus is very dependent upon the database being searched and may not be portable to otherdatabases.

NaturalLanguageQueries

Natural language interfaces improve the recall of systems with a decrease in precision when negationisrequired.

BrowseCapabilities

Once the search is complete, Browse capabilities provide the user withthe capability to determinewhich items are of interest and select those to be displayed. There are two ways of displaying asummary of the items that are associated with a query: line item status and data visualization. Fromthesesummarydisplays,theusercanselectthespecificitemsandzones

Withintheitemsfordisplay.

Ranking

Typically relevance scores are normalized to a value between 0.0 and 1.0. The highest value of 1.0 isinterpreted that the system is sure that the item is relevant to the search statement. In addition

torankingbaseduponthecharacteristicsoftheitemandthedatabase,inmanycircumstancescollaborative filtering isprovidinganoptionforselectingand orderingoutput.

Collaborative filtering has been very successful in sites such as AMAZON.COM MovieFinder.com,andCDNow.comindecidingwhatproductstodisplaytousersbaseduponheirquerie s.

Rather than limiting the number of items that can be assessed by the number of lines on a screen,other graphical visualization techniques showing the relevance relationships of the hit items can beused.

For example, a two or three dimensional graph can be displayed where points on the graph representitems and the location of the points represent their relative relationship between each other and theuser‟s query. In some cases color is also used in this representation. This technique allows a user toseetheclusteringofitemsbytopicsandbrowsethroughacluster or movetoanothertopicalcluster.

Zoning

Relatedtozoningforuseinminimizingwhatanenduserneedstoreview

fromahititemistheideaoflocalityandpassagebasedsearchandretrieval.

Highlighting

Most systems allow the display of an item to begin with the first highlight within the item and allowsubsequent jumping to the next highlight. The DCARS system that acts as a user frontend to theRetrieval Ware search system allows the user to browse an item in the order of the

paragraphs orindividual words that contributed most to the rank value associated with the item. The highlightingmay vary by introducing colors and intensities to indicate the relative importance of a particular wordintheiteminthedecisiontoretrievetheitem.

MiscellaneousCapabilities:

VocabularyBrowse

Vocabulary Browse provides the capability to display in alphabetical sorted order words from thedocument database. Logically, all unique words (processing tokens) in the database are kept in sortedorder along with a count of the number of unique items in which the word is found. The user canenter a word or word fragment and thesystem will begin to display the dictionary around the enteredtext.

It helps the user determine the impact of using a fixed or variable length mask on a search term andpotentialmis-spellings. The usercandeterminethatentering the searchterm "compul*"ineffect issearching for "compulsion" or "compulsive" or "compulsory." It also shows that someone probablyenteredtheword"computen"whenthey really meant"computer."

TERM	OCCURRENCES
compromise	53
comptroller	18
compulsion	5
compulsive	22
compulsory	4

IterativeSearchandSearchHistoryLog

Frequently a search returns a Hit file containing many more items than the user wants to review.Rather than typing in a complete new query, the results of the previous search can be used as aconstraining list tocreate a new query that is applied against it. This has the same effect as taking theoriginal query and adding additional search statement against it in an AND condition. This process ofrefining the results of a previous search to focus on relevant items is called iterative search. This alsoapplies when a user uses relevance feedback to enhance a previous search. The search history log isthecapabilitytodisplayallthepreviussearchesthatwereexecutedduringthecurrentsession.

CannedQuery

The capability to name a query and store it to be retrieved and executed during a later user

session iscalled canned or stored queries. A canned query allows a user to create and refine a search thatfocuses on the user‟s general area of interest one timeand then retrieve it to add additional searchcriteria to retrieve data that is currently needed. Canned query features also allow for variables to beinsertedintothequeryandboundto specific valuesatexecutiontime.

Z39.50andWAISStandards

The Z39.50 standard does not specify an implementation, but the capabilities withinan application(ApplicationService)and theprotocolusedtocommunicate betweenapplications (InformationRetrieval Application Protocol). It is a computerto computer communications standard for databasesearchingandrecordretrieval.Itsobjectiveistoovercomedifferentsystemincompatibilitiesassociatedwith multipledatabase searching.

The first version of Z39.50 was approved in 1992. An international version of Z39.50, called theSearchandRetrieveStandard(SR),wasapprovedbytheInternationalOrganizationforStandardization(ISO)in1991.Z39.50-1995,thelatestversionofZ39.50,replacesSRastheinternationalinformationretrievalstandard.Thestandarddescribeseightoperationtypes:Init(initialization),Search,Present,Delete,Scan,Sort,Resource-report,andExtendedServices.Therearefivetypesofqueries(Types0,1,2,100,101,and102).

Theclientisidentifiedas the “Origin”andperformsthecommunicationsfunctions relatingtoinitiating a search, translation of the query into a standardized format, sending a query, and requestingreturn records. The server is identified as the “Target” and interfaces to the database at the remoteresponding to requests from the Origin (e.g., pass query to database, return records in a standardizedformat and status). The end user does not have to be aware of the details of the standard since theOriginfunctionperformsthemappingfromtheuser‟squeryinterfaceintoZ39.50format.

This makes the dissimilarities of different database systems transparent to the user and facilitatesissuing one query against multiple databases at different sites returning to the user a single integratedHitfile.WideAreaInformationService(WAIS)isthedefactostandardformanysearchenvironmentsonthe INTERNET.WAISwasdevelopedbyaprojectstartedin1989by threecommercial companies (Apple, Thinking Machines, and Dow Jones). The original idea was to create aprogramthatwouldactasapersonallibrarian.

A free version of WAIS is still available via the Clearinghouse for Networked Information Discoveryand Retrieval (CINDIR) called "FreeWAIS." The originaldevelopment of WAIS started with the 1988Z39.50protocolasabasefollowingtheclient/serverarchitectureconcept.Thedevelopersincorporated the information retrieval concepts that allow for ranking, relevance feedback and naturallanguageprocessingfunctionsthatapplytofulltextsearchabledatabases.

Center for National Research Initiatives (CNRI) that is working with the Department of Defense andalso the American Association of Publishers (AAP), focusing on an Internet implementation thatallows for control of electronic published and copyright material. In addition to the Handle Serverarchitecture, CNRI is also advocating a communications protocol to retrieve items from existingsystems. This protocol call Repository Archive Protocol (RAP) defines the mechanisms for clients touse the handles to retrieve items. It also includes other administrative functions such as privilegevalidation. The Handlesystem is designed to meet the Internet Engineering TaskForce (IETF)requirements for naming Internet objects via Uniform Resource Names to replace URLs as defined intheInternet"sRFC-1737 (IETF-96).

WAIS(WideAreaInformationServers)

WAIS (Wide Area Information Servers) is an Internet system in which specialized subject databasesare createdatmultipleserverlocations, kepttrackof by a*directoryof servers* at one location, andmade accessible for searching by users with WAIS client programs. The user of WAIS is providedwith or obtains a list of distributed databases. The user enters a search argument for a selecteddatabase and the client then accesses all the servers onwhich the database is distributed. The resultsprovide a description of each text that meets the search requirements. The user can then retrieve thefull text. **RetrievalWare** is an enterprise search engine emphasizing natural language processing andsemanticnetworks.

Questions:

1. Explainthefunctionaloverviewofinformationstorageandretrievalsystem?
2. ListtheObjectivesofIRSandExplainaboutPrecisionandRecall?
3. Doesaprivateindexfiledifferfromastandarddatabasemanagementsystem(DBMS)?
4. ExplainbrieflyaboutFunctionalOverview"sinIRS?
5. Explainbrowsecapabilities?