INFORMATION RETRIEVAL SYSTEMS (CS3115PE)

(Professional Elective-II)

B.Tech. III Year I Semester								
Course Code	Category	Hours / Week			Credits	Maxumum Marks		
CS3115PE	Elective	L	T	P	C	CIA	SEE	Total
		3	0	0	3	30	70	100
Contact classes: 60	Tutorial Classes : NIL	Practical classes : NIL				Total Classes :60		
Prerequisites: Data Structures								

Course Objectives:

- To learn the important concepts and algorithms in IRS
- To understand the data / file structures that is necessary to design, and implement information retrieval (IR) systems.

Course Outcomes:

- Ability to apply IR principles to locate relevant information large collections of data
- Ability to design different document clustering algorithms
- Implement retrieval systems for web search tasks.
- Design an Information Retrieval System for web search tasks.

COURSE SYLLABUS

UNIT- I

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 DataWarehouses.

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

UNIT- II

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

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

UNIT- III

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.

UNIT- IV

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.

UNIT- V

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

TEXT BOOK:

1.Information Storage and Retrieval Systems-Theory and Implementation, SecondEdition, Gerald J.Kowalski, Mark T.Maybury, Springer

REFERENCE BOOKS:

1. Frakes,W.B., RicardoBaeza-Yates : Information Retrieval Data Structures and Algorithms, Prentice Hall, 1992.
2. Information Storage & Retrieval By Robert Korfhage-JohnWiley& Sons.
3. Modern Information Retrieval By Yates and Neto Pearson Education.

LECTURE NOTES

UNIT-I

Introduction: Definition,Objectives,Functional Overview,Relationship toDBMS, Digitallibraries and Data Warehouses. **Information Retrieval System Capabilities:** Search,Browse,MiscellaneousCapabilities.

WriteaboutInformationSystem?

Thereisapotentialforconfusionintheunderstandingofthedifferencesbetween DatabaseManagement Systems (DBMS) and Information Retrieval Systems. It is easy to confuse the softwarethat optimizes functional support of each type of system with actual information or structured datathat is being stored and manipulated. The importance of the differences lies in the inability of adatabase management system to provide the functions needed to process "information." The opposite,aninformationsystemcontainingstructureddata,alsosuffersmajor functionaldeficiencies.

1. DefinitionofInformationRetrievalSystem

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,videoandothermulti-mediaobjects. Techniquesarebeginningtoemergetosearchtheseothermediatypes(e.g.,EXCALIBUR'sVisualRetrieval Ware,VIRAGEvideoindexer).

The term "item" is used to represent the smallest complete unit that is processed and manipulated bythe system. The definition of item varies by how a specificsource treats information. A completedocument, such as a book, newspaper or magazine could be an item. For example a video newsprogramcouldbeconsidered an item.It is composedof text intheformof closedcaptioning, audiotextprovidedbythespeakers,andthevideo imagesbeingdisplayed.

An Information Retrieval System consists of a software program that facilitateuser in finding theinformation the user needs. The system may use standard computer hardware or specialized hardwareto support the search sub function and to convert non-textual sources to a searchable media (e.g.,transcriptionofaudio to text).

Thussearchcomposition,searchexecution,andreadingnon-relevantitemsareall aspects ofinformationretrievaloverhead.

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 largetextualinformation databasesfortheaverage user.

The introduction and exponential thealgorithmsandtechniquestooptimizetheprocessingandaccessoflargequantitiesoftextualdatawere oncethesoledomainofsegmentsoftheGovernment,afewindustries,andacademics. ImagesacrosstheInternetaresearchablefrommanywebsitessuchasWEBSEEK,DITTO.COM,ALTAVISTA/IMAGES.

Growth of the Internet along with its initial WAIS (Wide AreaInformationServers)capabilityand morerecentlyadvancedsearchservers (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 aremaking historical audio news searchable via the audio transcribed versions of the news. Major videoorganizations such as Disney are using video indexing to assist in finding specific images in theirpreviouslyproduced videostouseinfuturevideosorincorporateinadvertising.

2. ObjectivesofInformationRetrievalSystems.

The general objective of an Information Retrieval System is to minimize the overheadof a userlocating needed information. Overhead can be expressed as the time a user spends in all of the stepsleading to reading an item containing the needed information(e.g., query generation, query execution,scanningresultsofquerytoselectitemstoread,reading non-relevantitems).

In information retrieval the term “relevant” item is used to represent an item containing the neededinformation. Fromauser’sperspective“relevant”and“needed”aresynonymous.

Thetwomajormeasurescommonlyassociatedwithinformationsystemsare

1) Precision
2) Recall

Whenauserdecidestoissuea searchlookingforinformationonatopic,thetotaldatabaseislogicallydividedintofoursegments

Relevant items are those documentsthat contain information that helps the searcher in answering hisquestion. 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’squery.Precisionandrecallaredefinedas:

Precision is directly affected by retrieval of non-relevant items and drops to a number close to zero.Recallis noteffected by retrievalof non-relevantitems and thus remains at 100 percent onceachieved.

Information Retrieval Systems such as RetrievalWare, TOPIC,AltaVista, Infoseek and INQUERYthat the idea of accepting natural language queries is becominga standard system

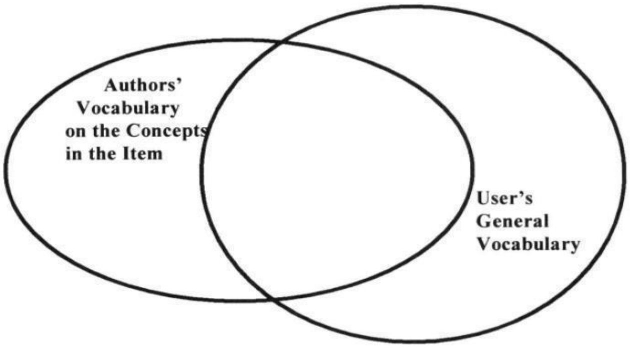


Figure 1.3 Vocabulary Domains

feature. This allowsusers to state in natural language what they areinterested in finding. But the completeness of the userspecificationislimitedbytheuser'swillingnesstoconstructlongnaturallanguagequeries.Mostuser sontheInternetenteroneortwosearchterms.

3. FunctionalOverview

AtotalInformationStorageandRetrievalSystemiscomposedoffourmajorfunctionalprocesses:

1) ItemNormalization
2) SelectiveDisseminationofInformation(i.e.,"Mail")
3) ArchivalDocumentDatabaseSearch,andanIndex
4) DatabaseSearchalongwiththeAutomaticFileBuildprocessthatsupportsIndexFiles

ItemNormalization:

The first step in any integrated system is to normalize the incoming items to a standard format. Itemnormalizationprovideslogicalrestructuringoftheitem.Additionaloperationsduringitemnormalization 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 thetokens. The processing tokens and their characterization are used to define the searchable text from the totalreceived text. Figure 1.5 shows the normalization process. Standardizing the input takes the differentexternal formats of input data and performs the translation to the formats acceptable to the system. Asystemmayhaveasingleformatforallitemsorallowmultipleformats.Oneexampleofstandardizationc ouldbetranslationofforeignlanguagesintoUnicode.Everylanguagehasadifferentinternalbinaryencodingforthecharactersinthelanguage.OnestandardencodingthatcoversEnglish,French,Spanish,etc.isISO-Latin.

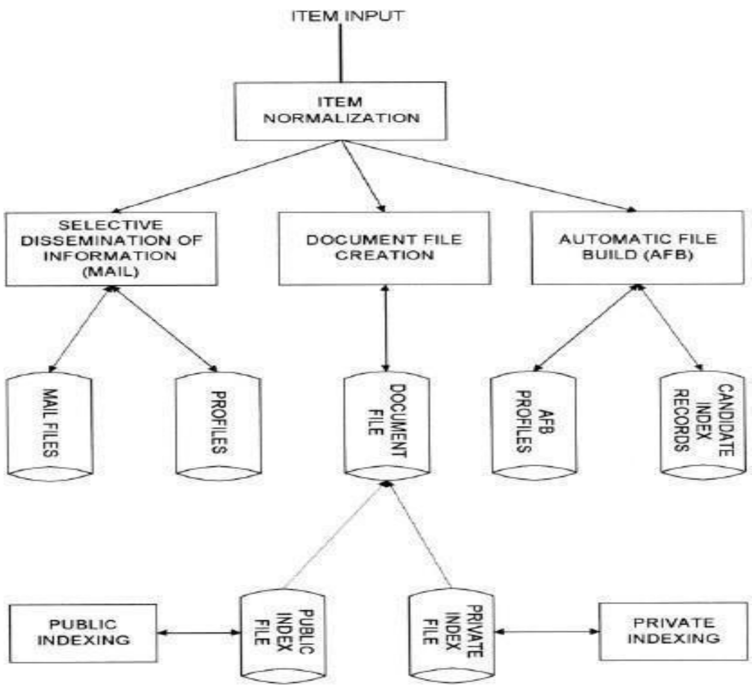


Figure 1.4 Total Information Retrieval System

To assistthe users ingenerating indexes,especiallytheprofessionalindexers, the systemprovidesaprocesscalled *AutomaticFileBuild(AFB)*.

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

The next process is to parse the item into logical sub-divisions that have meaning to the user. Thisprocess, called “Zoning,” is visible to the user and used to increase the precision of a search andoptimize the display. A typical item is sub- divided into zones, which may overlap and can behierarchical, such as Title, Author, Abstract, Main Text, Conclusion, and References. The zoninginformationispassedtotheprocessingtokenidentificationoperationtostoretheinformation,allowing searches to be restricted to a specific zone. For example, if the user is interested in articlesdiscussing “Einstein” then the searchshould notinclude the Bibliography,which could includereferencestoarticleswritten by“Einstein.”

Systemsdeterminewordsbydividinginput symbolsinto3classes:

1) Validwordsymbols
2) Inter-wordsymbols
3) Specialprocessingsymbols.

A word is defined as a contiguous set of word symbols bounded by inter-word symbols. In manysystems inter-word symbols are non-searchable and should be carefully selected. Examples of wordsymbols are alphabetic characters and numbers. Examples of possible inter-word symbols are blanks,periods and semicolons. The exactdefinition of an inter-word symbol is dependent upon the aspectsof the language domain of the items to be processed by the system. For example, an apostrophe maybeoflittleimportance ifonlyused forthepossessivecasein English,butmightbecriticaltorepresentforeignnamesinthedatabase.

Next,a*StopList/Algorithm*isappliedtothelistofpotentialprocessingtokens.TheobjectiveoftheStop function is to save system resources by eliminating from theset of searchable processing tokensthosethathavelittlevaluetothesystem.Giventhesignificantincreaseinavailablecheapmemory,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” (thiswas selected to allowdatesto besearchable)Stopanyprocessingtokenthathasnumbersandcharactersintermixed

2) SelectiveDissemination(Distribution,Spreading)ofInformation

The Selective Dissemination of Information (Mail) Process provides the capability to dynamicallycompare newly received items in the information system against standing statements of interest ofusers 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 usermail 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 ifthesearchstatementintheprofileissatisfied.SelectiveDisseminationofInformationhasnotyetbeenapplied to multimediasources.

3) DocumentDatabaseSearch

The Document Database Search Process provides the capability for a query to search against all itemsreceived 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 itemsthat have been received,processedand stored by the system. Typically items in the DocumentDatabasedonotchange(i.e.,arenotedited)oncereceived.

IndexDatabaseSearch

When an item is determined to be of interest, a user may want to save it for future reference. This isineffectfilingit.Inaninformationsystemthisisaccomplishedviathe index process. In this process the user can logically store an item in a file along with additionalindex terms and descriptive text the user wants to associate with the item. The Index Database SearchProcess(seeFigure1.4)providesthecapabilitytocreateindexesandsearchthem.

Thereare2classesofindexfiles:

1) PublicIndexfiles
2) PrivateIndexfiles

Every user can have one or more Private Index files leading to a very large number of files. EachPrivate Index file references only asmall subset of the total number of items in the DocumentDatabase. Public Index files are maintained by professional library services personnel and typicallyindexeveryitemintheDocumentDatabase.ThereisasmallnumberofPublicIndexfiles.Thesefileshave 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* showninFigure1.4 (also calledInformationExtraction).

MultimediaDatabaseSearch

From asystem perspective, the multi-media data is not logically its owndatastructure, but anaugmentation totheexistingstructuresintheInformationRetrievalSystem.

4. RelationshiptoDatabaseManagementSystems

From a practical standpoint, the integration of DBMS’s and Information Retrieval Systems is veryimportant. Commercial database companies have already integrated the two types of systems. One ofthefirstcommercialdatabasestointegratethetwosystemsintoa singleviewistheINQUIREDBMS.This has been available for over fifteen years. A more current example is the ORACLE DBMS thatnow offers an imbedded capability called CONVECTIS, which is an informational retrievalsystemthat uses a comprehensive thesaurus which provides the basis to generate "themes" for a particularitem. The INFORMIX DBMS has the ability to link to RetrievalWare to provide integration ofstructureddataandinformationalong withfunctionsassociatedwithInformationRetrievalSystems.

DigitalLibrariesandDataWarehouses(DataMarts)

As the Internet continued its exponential growth and project fundingbecame available, the topic ofDigital Libraries has grown. By 1995 enough research and pilot efforts had started to support the 1STACMInternationalConferenceonDigitalLibraries(Fox-96).Indexingisoneofthecriticaldisciplines in library science and significant effort has gone into the establishment of indexing andcatalogingstandards. Migrationofmanyof thelibraryproductstoadigitalformatintroducesbothopportunities and challenges. Information Storage and Retrieval technology has addressed a smallsubsetoftheissuesassociatedwithDigitalLibraries.

Data warehouses are similar to information storage and retrieval systems in that they both have aneed for search and retrieval of information. But a data warehouse is more focused on structured dataand decision support technologies. Inaddition to the normal search process, acomplete systemprovidesaflexiblesetofanalyticaltoolsto"mine"thedata.Datamining(originally calledKnowledge Discovery in Databases -KDD)is asearchprocess that automatically analyzes data andextractrelationships anddependenciesthatwerenotpartofthedatabasedesign.

InformationRetrieval SystemCapabilities

Search

CapabilitiesBrowse

CapabilitiesMiscellaneous

CapabilitiesStandards

The search capabilities address both Boolean and Natural Language queries. The algorithms used forsearching are called Boolean, natural language processing and probabilistic. Probabilistic algorithmsuse frequency of occurrence of processing tokens (words) in determining similarities between queriesanditemsandalsoinpredictorsonthepotentialrelevanceofthefounditemtothesearcher.The newer systems such as TOPIC, RetrievalWare, and INQUERY all allow for natural languagequeries.

Browse functions to assist the user in filtering the search results to find relevant information are veryimportant.

SearchCapabilities

The objective of the search capability is to allow for a mapping between a user's specified need andthe items in the information database that will answer that need. It can consist of natural language textin compositionstyle and/orquery terms (referredto as terms in this book)with Boolean logicindicators between them. One concept thathas occasionally been implemented in commercial systems(e.g., RetrievalWare), and holds significant potential for assisting in the location and ranking ofrelevant items, is the "weighting" of search terms. This would allow a user to indicate the importanceofsearch terms in either a Boolean or natural language interface. Given the following natural languagequery statement where the importance of a particular search term is indicated by a value in parenthesisbetween0.0 and1.0with1.0being themostimportant.

The search statement may apply to the complete item or contain additional parameters limiting it to alogical division of the item (i.e., to a zone). Based upon the algorithms used in a system manydifferent functions are associated with the system's understanding the search statement. The functionsdefine 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/Thesaurusexpansion).

BooleanLogic

Boolean logic allows a user to logically relate multiple concepts together to define what informationis needed. Typically the Boolean functions apply to processing tokens identified anywhere within anitem.ThetypicalBooleanoperatorsare**AND,OR,**and**NOT**.

These operations are implemented using set intersection, set union and set difference procedures. Asearch terms in either a Boolean or natural language interface. Given the following natural languagequery statement where the importanceof a particular search term is indicated by a value in parenthesisbetween0.0and1.0with1.0beingthemostimportant.

Thesearchstatementmay apply to the complete item orcontain additional Parame search terms ineither a Boolean or natural language interface. Given the following natural language query statementwhere the importanceof aparticular search term is indicated by a value in parenthesis between 0.0and1.0with1.0beingthemostimportant.

The search statement may apply to the complete item or contain additional parameters limiting it to alogical division of the item (i.e., to a zone). Based upon the algorithms used in a system manydifferent functions are associated with the system's understanding the search statement. The functionsdefine the relationships between the terms in the search statement (e.g., Boolean, Natural Language,Proximity,ContiguousWord Phrases,andFuzzy Searches)andtheinterpretation ofaparticularword (e.g., Term Masking, Numeric and Date Range, Contiguous Word Phrases, and Concept/Thesaurusexpansion).

ters limiting it to a logical division of the item (i.e., to a zone). Based upon the algorithms used in asystem many different functions are associated with the system's understanding the search statement.The functions define the relationships betweenthe terms in the search statement (e.g., Boolean,Natural Language, Proximity, Contiguous Word Phrases, and Fuzzy Searches) and the interpretationof a particular word (e.g., Term Masking, Numeric and Date Range, Contiguous Word Phrases, andConcept/Thesaurusexpansion).fewsystemsintroducedtheconceptof"exclusiveor"butitisequivalent to a slightly more complex query using the other operators and is not generally useful touserssince mostusersdo notunderstandit.

A special type of Boolean search is called "M of N" logic. The user lists a set of possible search termsandidentifies,asacceptable,anyitemthatcontainsasubsetoftheterms.

Forexample,"Findanyitemcontaining any two of the following terms: "AA," "BB," "CC." This can be expanded into a Booleansearch that performs an AND between all combinations of two terms and "OR"s the results together((AAANDBB)or(AAANDCC)or(BBANDCC)).

Proximity

Proximity is used to restrict the distance allowed within an item between two searchterms. Thesemanticconceptis thatthe clossearch terms in either a Boolean ornatural language interface.Giventhe following natural language query statement where the importance of a particular search term isindicatedbyavalueinparenthesisbetween0.0and1.0with1.0beingthemostimportant.

The search statement may apply to the complete item or contain additional parameters limiting it to alogical division of the item (i.e., to a zone). Based upon the algorithms used in a system manydifferent functions are associated with the system's understanding the search statement. The functionsdefine 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/Thesaurusexpansion).

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 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.Thisisbec auseProximity 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).Intheprocessofexpandin gaquerytermfuzzysearchingincludesothertermsthathave 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*”ImbeddedStringSea
rch

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 onelevel 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.

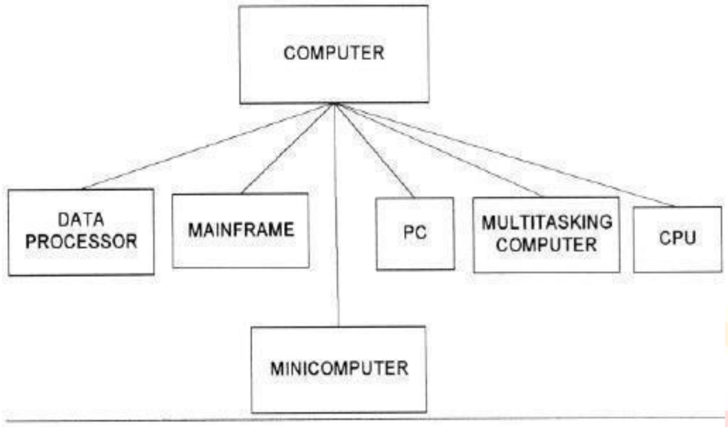


Figure 2.4 Thesaurus for term "computer"

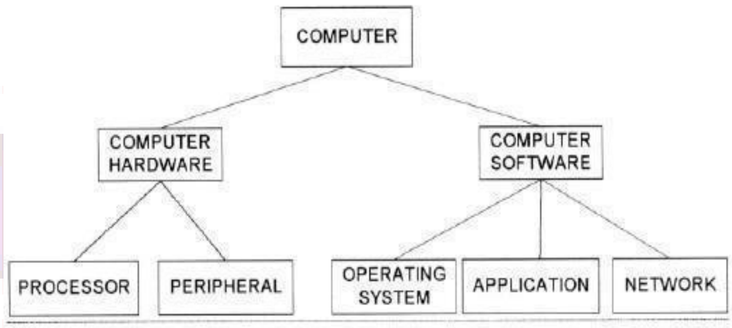


Figure 2.5 Hierarchical Concept Class Structure for "Computer"

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.comindecidingwhatproductstodisplaytousersbasedupontheirqueries.

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 isthecapabilitytodisplayalltheprevioussearchesthatwereexecutedduringthecurrentsession.

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.

<u>Questions</u>:

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

UNIT-II

Cataloging and Indexing: Objectives, Indexing Process, Automatic Indexing, formation Extraction.**Data Structures:** Introduction, Stemming Algorithms, Inverted file structures, N-gram data structure,PATdatastructure,Signaturefilestructure,Hypertextdatastructure.

CATALOGINGAND INDEXING:

Thetransformationfromreceiveditemtosearchabledatastructureiscalledindexing.

- Processcanbemanualorautomatic.
- Creatingadirectsearchindocumentdatabaseorindirectsearchthroughindexfiles.
- Concept based representation: instead of transforming the input into a searchable format somesystemstransformtheinputintodifferentrepresentationthatisconceptbasedSearch?Searchandret urnitem aspertheincomingitems.

Historyof indexing:shows thedependencyof informationprocessingcapabilities onmanual andthen automaticprocessingsystems.

- Indexingoriginallycalledcataloguing:oldesttechniquetoidentitythecontentsofitemsto assistinretrieval.
- Itemsoverlapbetweenfullitemindexing,publicandprivateindexingoffiles

Objectives:

Thepublic fileindexerneedstoconsidertheinformationneedsofallusersoflibrarysystem.Itemsoverlapbetween fullitemindexing,publicand privateindexingoffiles.

- Usersmayusepublicindexfilesaspartofsearchcriteriatoincreaserecall.
- Theycanconstraintheresearchbyprivateindexfiles
- The primary objective of representingtheconceptswithinan item tofacilitate usersfindingrelevantinformation.
- Usersmayusepublicindexfilesaspartofsearchcriteriatoincreaserecall.
- Theycanconstraintheresearchbyprivateindexfiles
- The primary objective of representing theconceptswithinan itemtofacilitate usersfindingrelevantinformation

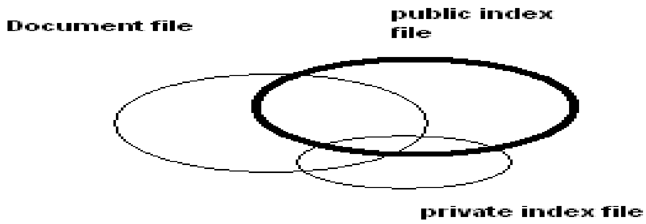
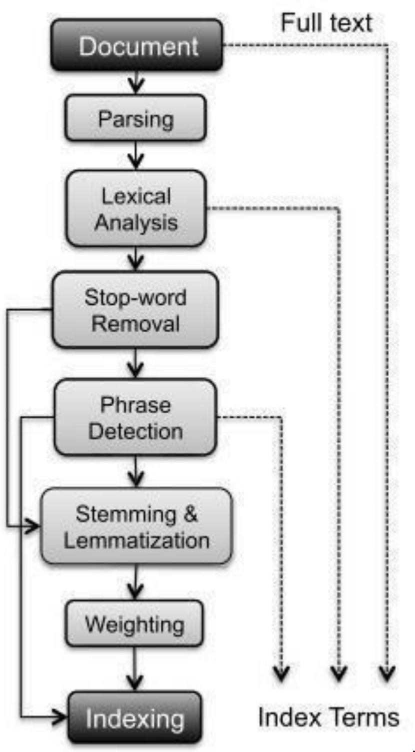


Fig:Indexingprocess

1. Decide thescope of indexing and the levelof detailto be provided. Basedonusagescenario ofusers.
2. Seconddecisionistolinkindextermstogetherinasingleindexforaparticularconcept.

TEXTPROCESSING

Fig. 2.3 Text processing phases in an IR system



1. Document Parsing. Documents come in all sorts of languages, character sets,and formats; often, the same document may contain multiple languages orformats, e.g., a French email with Portuguese PDF attachments. Documentparsing deals with the recognition and "breaking down" of the documentstructure into individual components. In this pre processing phase, unitdocuments arecreated; e.g., emails with attachments are split into one documentrepresentingtheemailandasmanydocumentsastherearattachments.

2. Lexical Analysis. After parsing, lexical analysis tokenizes a document, seen as an input stream, into words. Issues related to lexical analysis include the correct identification of accents, abbreviations, dates, and cases. The difficulty of this operation depends much onthe language at hand: for example, the Englishlanguage has neither diacritics nor cases, French has diacritics but no cases,German has both diacritics and cases. The recognition of abbreviations and, in particular, of time expressions would deserve a separate chapter due to itscomplexityandtheextensiveliteratureinthefieldForcurrentapproaches

3. Stop-Word Removal. A subsequent step optionally applied to the results oflexical analysis is stop-word removal, i.e., the removal of high-frequency words. For example, given the sentence "search engines are the most visible information retrieval applications" and a classic stop words set such as the one adopted bythe Snowball stemmer,1 the effect of stop-word removal would be: "searchenginemostvisibleinformationretrievalapplications".

4. PhraseDetection.Thisstepcapturestextmeaningbeyondwhatispossible withpure bag- of-word approaches, thanks to the identification of noun groups andother phrases. Phrase detection maybe approached in several ways, includingrules(e.g.,retainingterms thatarenotseparatedbypunctuationmarks),morphologicalanalysis,syntacticanalysis,andcombinations thereof.

5. Forexample,scanningourexamplesentence"searchengines arethe mostvisible information retrievalapplications"fornounphraseswouldprobablyresultinidentifying"searchengines"and"informationretrieval".

6. Stemming and Lemmatization. Following phrase extraction, stemmingandlemmatization aim at stripping down word suffixes in order to normalize theword. In particular, stemming is a heuristic process that"chops off" the ends ofwords in the hope of achieving the goal correctly most of the time; a classic rulebased algorithm for this was devised by Porter [280]. According to the Porter stemmer, ourexamplesentence"Searchenginesarethemostvisibleinformationretrieval applications" would result in: "Search engine are the most visibleinformretrievapplic".

7. Lemmatizationisaprocessthattypicallyusesdictionariesandmorphologicalanalysis of words in order to return the base or dictionary form of a word,thereby collapsing its inflectional forms (see, e.g., [278]). For example, oursentence would result in "Search engine are the most visible informationretrieval application" when lemmatized according to a WordNet-basedlemmatizer

8. Weighting. The final phase of text preprocessing deals with termweighting. Aspreviously mentioned, words in a text have different descriptive power; hence,indextermscanbeweighteddifferentlytoaccountfortheirsignificancewithinadocumentand/ora documentcollection.Sucha weightingcanbebinary,e.g.,

assigning0 fortermabsenceand1 forpresence.

SCOPEOFINDEXING

- Whenperform the indexingmanually,problems arisefrom twosourcestheauthorandtheindexertheauthorandtheindexer.
- Vocabularydomainmaybedifferenttheauthorandtheindexer.
- Thisresultsindifferentqualitylevelsofindexing.
- Theindexermustdeterminewhentostoptheindexingprocess.
- Twofactorstodecideonleveltoindextheconceptinaitem.
- Theexhaustivelyandhowspecificindexingisdesired.
- Exhaustivelyofindexistheextenttowhichthedifferentconceptsintheitemareindexe d.
- Forexample,iftwosentencesofa10-pageitemonmicroprocessorsdiscusson-boardcaches,shouldthisconceptbeindexedSpecificrelatestoprecisenessofindextermsus edinindexing.
- For example, whether the term “processor” or the term“microcomputer”ortheterm“Pentium”shouldbeusedintheindexofanitemisbasedupo n thespecificitydecision.

- Indexinganitemonlyonthemostimportantconceptinitandusinggeneralindextermsyieldslo wexhaustivelyandspecificity.
- Another decision on indexing is what portion of an item to beindexedSimplestcaseistolimittheindexingtotitleandabstract(conceptual)zone.
- Generalindexingleadstolossofprecisionandrecall.

PREORDINATIONANDLINKAGES

- Anotherdecisiononlinkages process whetherlinkages areavailablebetweenindextermsforanitem.
- Usedtocorrelateattributesassociatedwithconceptsdiscussedinanitem.‘t hisprocessiscalled preordination.
- When indexterms are notcoordinatedat indextime thecoordination occurs at search time. This is called postcoordination,implementingby“AND”ingindexterms.
- Factorsthatmustbedeterminedinlinkageprocessarethenumberoftermsthatcanberelate d.
- Ex.,anitemdiscusses‘thedrillingofoilwellsinMexicobyCITGOandthei ntroductionofoilrefineriesinPerubytheU.S.’

DATASTRUCTURES

- IntroductiontoDataStructures
- StemmingAlgorithms
- InvertedFileStructure
- N-GramDataStructure
- PATDataStructure
- SignatureFileStructure
- HypertextandXMLDataStructures

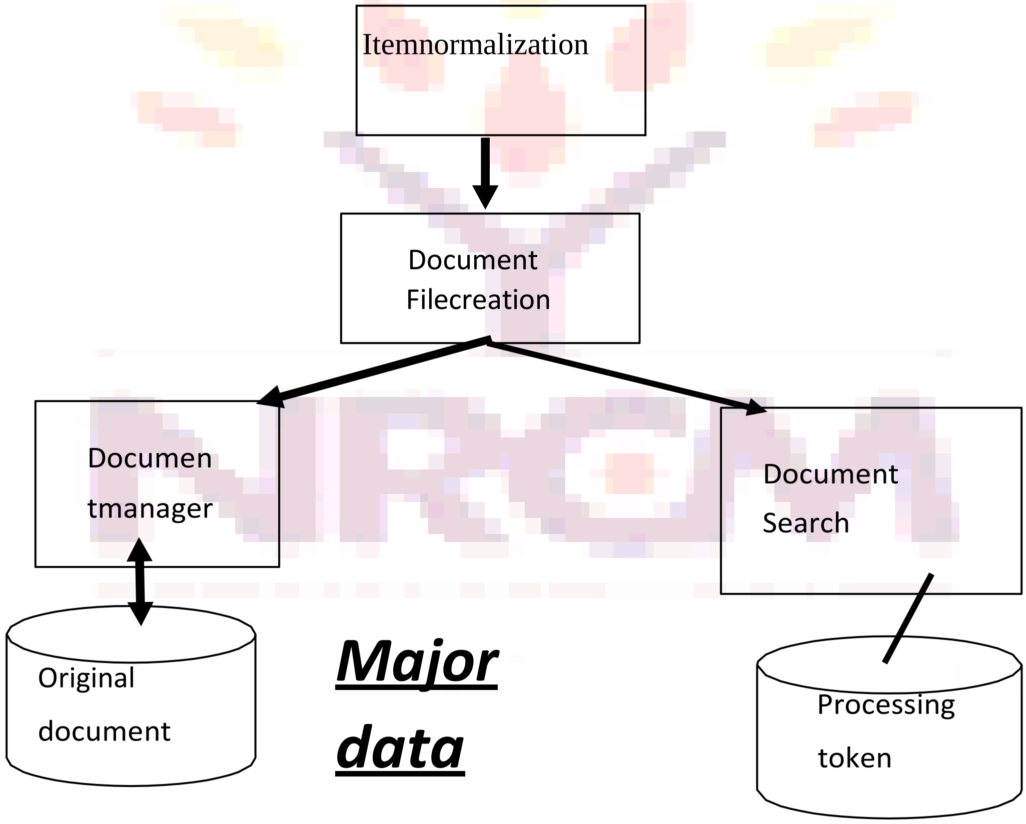
Datastructure:

Theknowledgeofdatastructuregivesaninsightintothecapabilitiesavailabletothesystem.

- Eachdata structurehasasetofassociatedcapabilities.
- Abilitytorepresenttheconceptandtheirr/s.
- Supports locationofthoseconcepts

IntroductionTwomajordata structuresinanyIRS:

1. Onestructurestoresandmanagesreceiveditemsintheirnormalizedformiscalleddocument manger
2. Theotherdatastructurecontainsprocessingtokensandassociateddatatosupportsearch.



Resultsofasearcharereferencestotheitemsthatsatisfythesearchstatementwhicharepassed to thedocumentmanagerforretrieval.

Focus:ondatastructurethatsupportsearchfunction

Stemming:isthetransformationoftenappliedtodatabeforeplacingitinthesearchabledatastru cture.

Stemmingrepresentsconcept(word)toacanonical(authorized;recognized;accepted)morpholo gical(thepatternsofwordformationinaparticularLanguage)representation .Risk with stemming: concept discriminationinformation may belostintheprocess.Causing decreaseinperformance.

Advantage: has a potential to increaserecall.STEMMINGALGORITHMS

- Stemmingalgorithmisusedtoimprovetheefficiencyof IRSandimproverecall.
- Conflation (the process or result of fusing items into one entity; fusion;amalgamation)isatermthatisusedtorefermappingmultiplemorphologicalvariantstosinglerepresentation(stem).
- Stem carries the meaning of the concept associated with the word and theaffixes(ending)introducesubtle(slight)modificationoftheconcept.
- Termswithacommonstemwillusuallyhavesimilarmeanings,fore xample:
- Ex:Termswithacommonstemwillusuallyhavesimilarmeanings,forexa mple:
- CONNECT
- CONNECTED
- CONNECTING
- CONNECTION
- CONNECTIONS
- Frequently, the performance of an IRsystemwill be improved iftermgroups such as this are conflated into asingle term. This may be done byremovalof the various suffixes-ED,-ING,-ION, IONS to leave the singletermCONNECT
- In addition, the suffix stripping process will reduce the total number ofterms in the IR system, and hence reduce the size and complexity of thedatain thesystem,which isalwaysadvantageous.

❖ Majorusageofstemmingistoimproverecall.

➢ Importantforasystemtocategoriesawordpriortomakingthedecisiontostem.

➢ Propernamesandacronyms(AwordformedfromtheinitiallettersofanamesayIARE …)shouldnothavestemmingapplied.

➢ Stemmingcanalsocauseproblems for natural languageprocessing NPLsystemsbycausing lossofinformation.

PORTERSTEMMINGALGORITHM

- Basedonasetconditionofthestem
- AconsonantinawordisaletterotherthanA,E,I,OorU,someimportantstemconditionsare

1. Themeasuremofastemisafunctionofsequenceofvowels(V)followed byasequenceofconsonant(C).
2. C(VC)mV.misnumberVCrepeatsThecasem=0coversthenullword.
3. *<X>-stemendswithaletterX3.*v*-stemcontainsavowel
4. *d-stemendsindoubleconsonant(e.g.-TT,-SS).
5. *o-stemendsinconsonantvowelsequence wherethefinalconsonantisnotw,x,y(e.g.-WIL,-HOP).

Suffixcond.stakestheformcurrent_suffix==patternActionsareintheformold_suffix ->.New_suffix

Rulesaredividedintostepstodefinetheorder

forapplyingtherule.Examplesoftherules

Step	Condition	Suffix	Replacement	Example
1a	Null	Sses	Ss	Stresses->stress
1b	*v*	Ing	Null	Making->mak
1b1	Null	At	Ate	Inflated->inflate
1c	*v*	Y	I	Happy->happi
2	m>0	aliti	al	Formaliti->formal
3	m>0	Icate	Ic	Duplicate->duplie
4	m>1	Able	Null	Adjustable->adjust
5a	m>1	e	Null	Inflate->inflat
5b	m>1and*d	Null	Singleletter	Control-> control

2. Dictionarylookupstemmers

- Useofdictionarylookup.
- Theoriginaltermorstemmedversionofthetermislookedupinadictionaryandreplaced bythestemthatbestrepresentsit.
- ThistechniquehasbeenimplementedinINQUERYandRetrieval waresystems-

INQUERYsystemusesthetechniquecalledKstem.

- Kstemisamorphologicalanalyzerthatconflateswordsvariantstoarootform.
- Itrequiresawordtobeinthedictionary
- Kstemuses6majordatafilestocontrolandlimitthestemmingprocess.

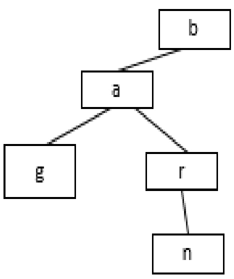
1. Dictionaryofwords(lexicon)
2. Supplementallistofwordsfordictionary
3. Exceptionallistofwordsthatshouldretaina‘e’attheend(e.g.,“suites”to“suite”but“suited”to“suit”).
4. Direct_conflation-wordpairsthatoverridestemmingalgorithm.
5. County_nationality_conflation(BritishmapstoBritain)
6. Propernouns--thatshouldnotbestemmed

- New words that are not special forms (e.g., dates,phone numbers) arelocatedinthedictionarytodeterminesimplerformsbystrippingoffsuffixes andrespellingpluralsasdefinedinthedictionary.

3. Successorstemmers:

- Basedonlengthofprefixes.
- Thesmallestunitofspeechthatdistinguishesonwordfromanother
- Theprocessusessuccessorvarietiesforaword.

Usesinformationtodivideawordintosegmentsandselectsonofthesegmentstostem.



Symbol tree for term bag , barn

Successorvarietyofwordsareusedtosegmentawordbyapplyingoneofthefollowingfourmetho ds.

1. Cutoffmethod:acutofvalueisselectedtodefinethestemlength.
2. Peakandplateau:a segmentbreakismadeafteracharacterwhosesuccessorvarietyexc eedsthatofthecharacter.
3. Completewordmethod:breakonboundariesofcompletewords.
4. Entropymethod:usesthedistributionmethodofsuccessorvarietyletters.

1. Let|Dak|bethenumberofwordsbeginningwithklengthsequenceoflettersa.
2. Let|Dakj|bethenumberofwordsinDakwithsuccessorj.
3. TheprobabilitythatamemberofDakhasthesuccessorjisgivenas|Da kj|/|Dak|Theentropyof|Dak|is26

   Hak=-(|Dakj|/|Dak|)(log(|Dakj|/|Dak|))p=1

Afterawordhasbeensegmentedthesegmenttobeusedasstemmustbeselecte

d.

HaferandWeissselectedthefollowingrule

If(firstsegmentoccursin<=12wordsindatabase)FirstsegmentisstemElse(seco

ndsegmentisstem)

INVERTEDFILESTRUCTURE

Invertedfilestructure

Mostcommondatastructure

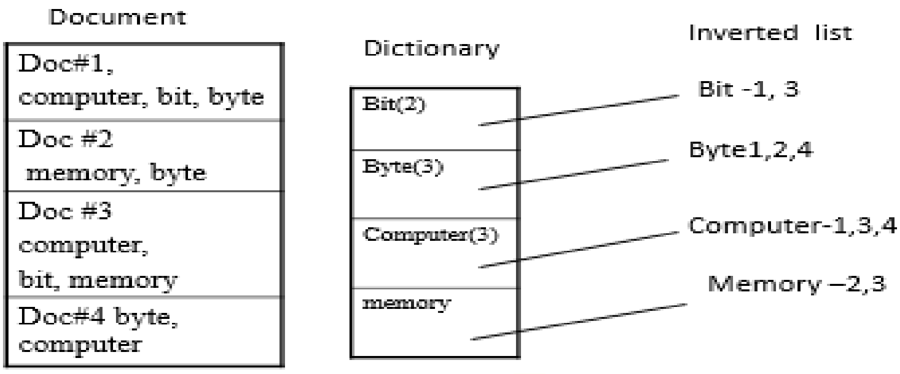
Invertedfilestructures are composedof three

filesThedocumentfile

1. Theinversionlist(PostingList)
2. Dictionary
3. Theinvertedfile:basedonthemethodologyofstoringaninversionofdocuments.
4. Foreachwordalistofdocumentsinwhichthewordisfoundisstored(inversionofdoc ument
5. Eachdocumentisgivenauniquethenumericalidentifierthatisstoredininversionlist.Dict ionaryisusedtolocatetheinversion listforaparticularword.

Thisisasortedlist(processingtokens)inthesystemandapointertothelocationofitsinversionlist.

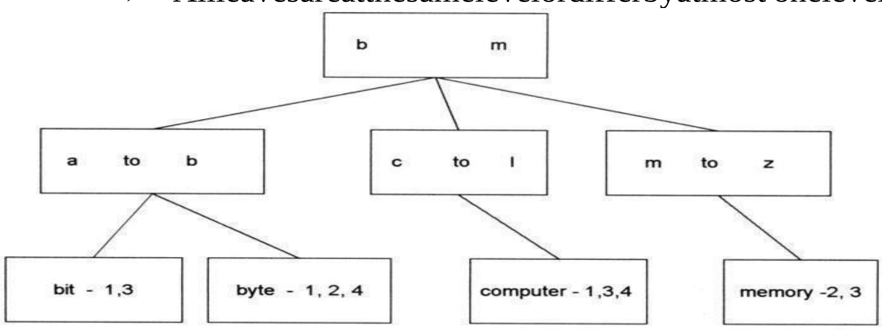
Dictionarycanalsostoreotherinformationusedinqueryoptimizationsuchaslengthofinver sionliststo increasetheprecision.



- Usezoningtoimprove
- PrecisionandRestrictentries.
- Inversionlistconsistsofdocumentidentifierforeachdocumentinwhichthe wordisfound.

Ex:bit1(10),1(12)1(18)isin10,12,18positionofthewordbitinthedocument#1.

- Whenasearchisperformed, theinversionlistsforthetermsinthequeryarelocateandappropriatelogicisapplied betweeninversionlists.
- Weightscanalsobestoredintheinversionlist.
- Inversionlistareusedtostoreconceptandtheirrelationship.
- Wordswithspecialcharacteristicscanbestoredintheirowndictionary.Ex:Date…w hichrequire dateranging andnumbers.
- Systemsthatsupportrankingarere-organizedinrankedorder.
- Btreescanalsobeusedforinversioninsteadofdictionary.
- Theinversionlistsmaybeattheleaflevelorreferencedinhigherlevelpointers.
- AB-treeoforder misdefinedas:
- Arootnodewithbetween2and2mkeys
- Allotherinternalnodeshavebetweenmand2mkeys
- Allkeysarekeptinorder fromsmallertolarger.
- Allleavesareatthesamelevelordifferbyatmost onelevel.



N-GRAMDATASTRUCTURE

- N-Gramscanbeviewedasaspecialtechniqueforconflation(stemming)andasauniquedatastructureininformationsystems.
- N-Gramsareafixedlengthconsecutiveseriesof"n"characters.
- Unlikestemmingthatgenerallytriestodeterminethestemofawordthatrepresentsthesemanticmeaningoftheword,n-gramsdonot careaboutsemantics.
- The searchable datastructure is transformedintooverlappingn-grams,whicharethenusedto create thesearchable database.
- Examplesofbigrams,trigramsandpentagramsforthewordphrase"seacolony."

seeacoolloonnyBigrams(nointerwordsymbols)
seacolololononyTrigrams(nointerwordsymbols)#seseaea##cocolololononyny#Trigrams(with

interwordsymbol#)

#sea##colocolonolonylony#Pentagrams(withi
nterwordsymbol#)

Thesymbol#isusedtorepresenttheinterwordsymbolwhichisanyoneofasetofsymbols(e.g.,blank,period,semicolon,colon,etc.).

- Thesymbol#isusedtorepresenttheinterwordsymbolwhichisanyoneofaset ofsymbols(e.g.,blank, period,semicolon, colon,etc.).
- Eachofthen-gramscreatedbecomesaseparateprocessingtokensandaresearchable.
- Itispossiblethatthesamen-gramcanbecreatedmultipletimesfromasingleword.

Uses:

- WidelyusedascryptographyinworldwarIISpellingerrorsdetectionandcorrection
- Usebigramsforconflatingterms.
- N-gramsaspotentialerroneouswords.
- Damerauspecified4categoriesoferrors:

ErrorCategory	Example
singlecharinsertion	compuuter
singlechardeletion	compter
singlecharsubstitution	compiter

Transpositionof2adjacent comp*tu*er chars

- Zamorashowedtrigramanalysisprovidedaviabledatastructurefori dentifying misspellingsandtransposedcharacters.
- Thisimpactsinformationsystemsasapossiblebasisforidentifyingpo tentialinputerrorsforcorrectionasaprocedurewithinthenormalization process.
- Frequencyofoccurrenceofn-grampatternscanalsobeusedforidentifyingthelanguageofanite m.
- Trigramshavebeenusedfortextcompressionandtomanipulatethelengthofindexterms.
- ToencodeprofilesfortheSelectiveDisseminationofInformation.
- Tostorethesearchabledocumentfileforretrospectivesearchdatabases.

Advantage:

TheyplaceafinitelimitonthenumberofsearchabletokenMaxSeg

n=(□ □

)nmaximumnumberofuniquengramsthatcanbegenerated.“n”ist

helength ofn-grams

□ number of process able symbols Disadvantage:

longerthengramthesizeofinversionlistincrease.Performanceha

s85 %precision.

<u>PATdatastructure</u>(practicalalgorithmtoretrieveinformationcodedinalphanumeric)

- PATstructureorPATtreeorPATarray:continuoustextinputdat astructures(stringlikeN-Gramdatastructure).
- The inputstream is transformed into asearchabledatastructureconsistingofsubstrings,allsubstringsareunique.
- Eachpositioninainputstringisaanchorpointforasubstring.
- IncreationofPATtreeseachpositionintheinputstringistheanchorpointfor a sub-string that starts at that point and includesall new text up to theendoftheinput.
- Binarytree,mostcommonclassforprefixsearch,ButPattreesaresortedlo gicallywhichfacilitaterangesearch,andmoreaccuratetheninversionfile.
- PATtreesprovidealternatestructureifsupportingstringssearch.

Text EconomicsforWarsawiscomplex.

----

sistring1Economics

forWarsawiscomplex.

sistring2

conomicsforWarsawiscomplex.

sistring5omicsforWarsawiscomplex.s

istring 10 for Warsaw is

complex.sistring20w is

complex.sist

ring30ex.

Examplesofsistrings

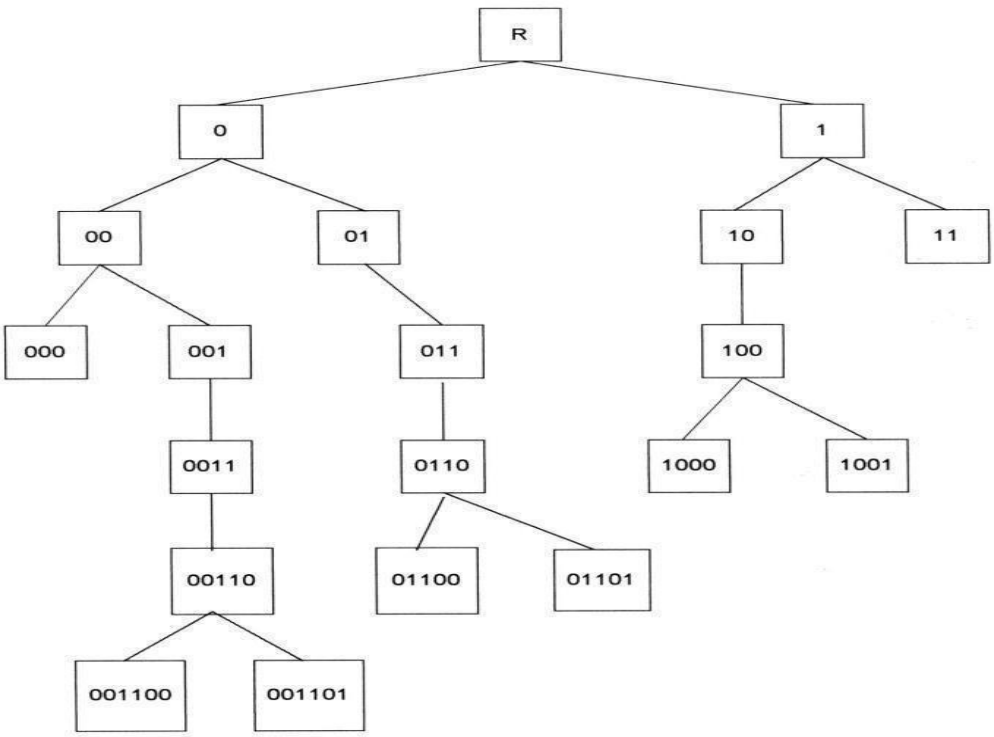
- Thekeyvaluesarestoredattheleafnodes(bottomnodes)inthePATTree.
- Foratextinputofsize“n”thereare“n”leafnodesand“n-1”atmosthigherlevelnodes.
- Itispossibletoplaceadditionalconstraintsonsistringsfortheleafnodes.
- Ifthebinaryrepresentationsof“h”is(100),“o”is(110),“m”is

(01) and“e”is(101)thentheword“home”producestheinput

100110001101.. Usingthesistrings.

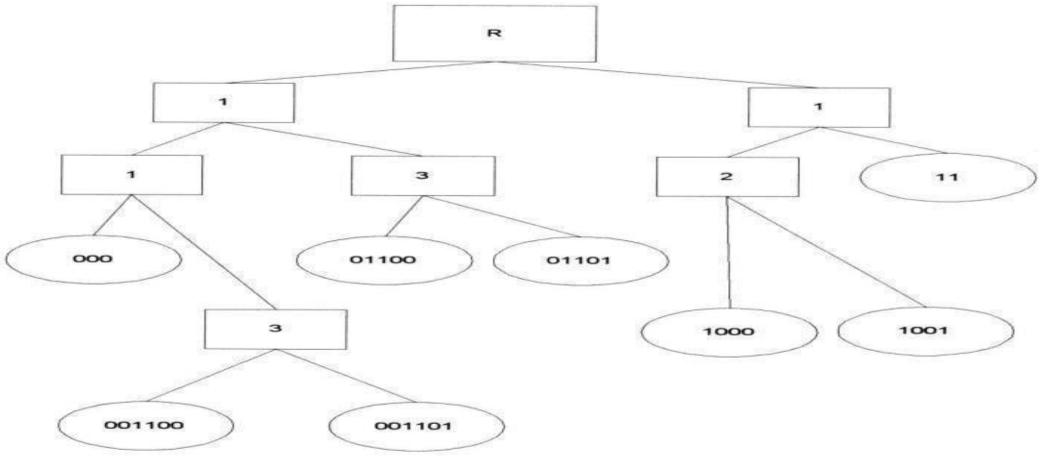
INPUT	100110001101
sistring1	1001....
sistring2	001100...
sistring3	01100....
sistring4	11.......
sistring5	1000...
sistring6	000.....
sistring7	001101
sistring8	01101

ThefullPATbinarytreeis



Thevalueintheintermediatenodes(indicatedbyrectangles)isthenumberofbitstoskipuntilthen
ext

bittocomparethatcausesdifferencesbetweensimilarterms.



SkippedfinalversionofPATtree

Signaturefilestructure

- Thecodingisbaseduponwordsinthecode.
- Thewordsaremappedintowordsignatures.
- Awordsignatureisfixedlengthcodewithafixednumberofbitssetto1.
- Thebitpositionsthataresettoonearedeterminedviaahashfunctionoftheword.
- ThewordsignaturesareOredtogethertocreatesignatureofanitem..
- Partitioningof wordsisdoneinblocksize,Whichisnothingbutsetofwords,Code lengthis16 bits.
- Searchisaccomplishedbytemplatematchingonthebitposition.
- provide apracticalsolution applied inparallel processing, distributedenvironmentetc.

- Toavoidsignaturesbeingtoodensewith“1”s,amaximumnumberof words is specified and an item is partitioned intoblocks of thatsize.
- Theblocksizeissetatfivewords,thecodelengthis16bitsandthenumb erofbitsthatareallowedtobe“1”foreachwordisfive.
- TEXT:ComputerSciencegraduatestudentsstudy(assumeblocksizeisfivewords)

WORD	Signature
computer	00010110 0000 0110
Science	1001000011100000
graduate	1000010101000010
students	0000011110000100
study	00000110 0110 0100
BlockSignature	1001011111100110

SuperimposedCoding

Application(s)/Advantage(s)

- Signaturefiles provide apracticalsolutionfor storing and locatinginformationinanumberofdifferentsituations.
- Signaturefileshavebeenappliedasmediumsizedatabases,databasesw ith low frequency of terms, WORM devices, parallelprocessingmachines,and distributedenvironments

HYPERTEXTANDXMLDATASTRUCTURES

- ❖ The adventof the Internet andits exponential growthandwide acceptanceas anew global information network has introduced new mechanisms forrepresentinginformation.
- ❖ Thisstructureiscalledhypertextanddiffersfromtraditionali nformationstoragedatastructuresinformat anduse.
- ❖ ThehypertextisHypertextisstoredinHTMLformatandXML.
- ❖ Botoftheselanguagesprovidedetaileddescriptionsforsubsetsoftextsimilartothezoning.
- ❖ Hypertextallowsoneitemtoreferenceanotheritemviaaembeddedpointer.
- ❖ HTMLdefinesinternalstructureforinformationexchangeoverWWWontheinternet.
- ❖ XML:definedbyDTD,DOM,XSL,etc.

Documentandtermclustering

Twotypesofclustering:

1) clusteringindextermstocreateastatisticalthesaurusand
2) Clusteringitemstocreatedocumentclusters.Inthefirstcaseclusteringisusedtoincreaserecallby expanding searches with related terms. In document clustering the search can retrieve itemssimilar to an item of interest, even if the query would not have retrieved the item. The clusteringprocess is not precise and care must be taken on use of clustering techniques to minimize thenegativeimpactmisusecanhave.

Questions:

1. ExplainaboutCatalogingandIndexing?
2. Writeaboutdatastructures?ExplainaboutStemmingAlgorithms?
3. Writeabouta)InvertedFileStructureb)N-GramDataStructurec)PATDataStructure?
4. ExplainaboutHypertextandXMLDataStructures?
5. a)ExplainaboutProbabilisticWeighting?
   b)WhatisVectorSpaceRetrievalModel withanexample?
6. ExplainaboutHiddenMarkovModels?

UNIT-III

Automatic Indexing: Classes of automatic indexing, Statistical indexing, Natural language, Conceptindexing, Hypertext linkages **Document and Term Clustering:** Introduction, Thesaurus generation,Itemclustering,Hierarchyofclusters.

AUTOMATIC INDEXING

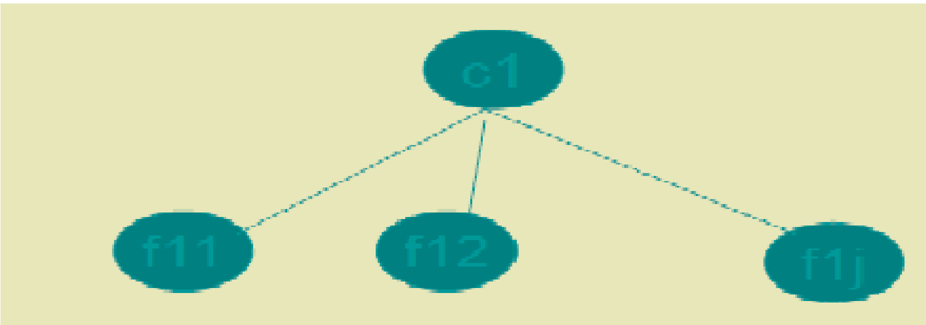
- Case:Totaldocumentindexing.
- Automatic indexing requiresfewsecondsbased on theprocessorandcomplexityofalgorithmstogenerateindexes.'
- Adv.isconsistencyinindextermselectionprocess.
- Indexresultingfromautomatedindexingfallintotwoclasses,weightedandunweighted.
- Unweighted indexingsystem: theexistence of an index termin adocumentandsometimesitswordlocationarekeptaspartofsearchabledatastruc ture.
- Weightedindexingsystem:aattemptismadetoplaceavalueontheindexter m associated with concept in the document. Based on the frequency ofoccurrence ofthetermintheitem.
- Valuesarenormalizedbetween0and1.
- Theresultsarepresentedtotheuserinorderofrankvaluefromhighestnu mbertolowestnumber.
- IndexingByterm
- Terms(vocabulary)oftheoriginalitemareusedasbasisofindexprocess.
- Therearetwomajortechniquesforcreationofindexstatisticalandnaturallanguage.
- Statisticalcanbebaseduponvectormodelsandprobabilisticmodelswitha special case being Bayesian model (accounting foruncertainty inherentinthe modelselectionprocess).
- Calledstatistical because theircalculationof weights useinformationsuchasfrequencyofoccurrenceofwords.
- Naturallanguagealsousessomestatisticalinformation,butperform

morecomplexparsingtodefinethefinal setofindexconcept.

- Otherweightedsystemsdiscussedasvectoris usedinformationsystem.
- Thesystememphasizesweightsasafoundationforinformationdetectionandstorestheseweightsi navectorform.
- Eachvectorrepresentsadocument.Andeachpositioninavectorrepresentau niqueword(processingtoken)inadatabase..
- Thevalueassignedtoeachpositionistheweightofthatterminthedocument.
- 0indicatesthatthewordwasnotinthedocument.
- Searchisaccomplishedbycalculatingthedistancebetweenthequeryvect orand documentvector.
- Bayesianapproach:basedonevidencereasoning(drawingconclusionfromevidence)
- Couldbeappliedaspartofindextermweighing.Butusually appliedaspartof retrieval process by calculating the relationship between an item andspecificquery.
- Graphic representation eachnode represents a randomvariable archbetweenthenodesrepresentaprobabilisticdependenciesbetweent henodeanditsparents.

TwolevelBayesiannetwork

- “c”“representsconceptinaquery
- “f”representingconceptsinanitem



- Anotherapproachisnaturallanguageprocessing.

- DR-LINK(documentretrievalthroughlinguisticsknowledge)
- Indexingbyconcept
- Concept indexing determines a canonical set of concept basedupon atestset of terms and uses them as base for indexing all items.Calledlatentsemanticsindexing.
- Ex:matchplussystemdevelopedbyHNCinc
- UsesneuralNWstrengthofthesystemwordrelationship (synonyms)andusestheinformationingeneratingcontextvectors.
- Twoneuralnetworks are used one togenerated stemcontextvectorsandanotheroneto performquery.
- Interpretationissameastheweights.
- Multimediaindexing:
- Indexingvideoorimagescanbeaccomplishedatrawdatalevel.
- Positionalandtemporal(time)searchcanbedone.

INFORMATIONEXTRACTION

Therearetwoprocessesassociatedwithinformationextraction:

1. determinationoffactstogointostructuredfieldsinadatabaseand
2. Extractionoftextthatcanbeusedtosummarizeanitem.

TheprocessofextractingfactstogointoindexesiscalledAutomaticFileBuild.

Inestablishingmetricstocompareinformationextraction,precisionandrecallareappliedwithslightmodifications.

- Recallreferstohowmuchinformationwasextractedfromanitemversusho wmuch should havebeenextractedfromtheitem.
- Itshowstheamountofcorrectandrelevantdata extractedversusthecorrectandrelevantdatain theitem.
- Precision refers to how much information was extractedaccuratelyversusthetotalinformationextracted.
- Additionalmetricsusedareovergenerationandfallout.
- Overgenerationmeasurestheamountofirrelevantinformationthatisextracted.
- Thiscouldbecausedbytemplatesfilledontopicsthatarenotintendedtobeextra ctedorslotsthat get filledwithnon-relevantdata.
- Falloutmeasureshowmuchasystemassignsincorrectslotfillersasthenumberof
- Thesemeasuresareapplicabletobothhumanandautomatedextractionprocesses.

- Anotherrelatedinformationtechnologyisdocumentsummarization.
- Ratherthantryingtodeterminespecificfacts,thegoalof documentsummarization is to extract a summary of an item maintaining the mostimportantideaswhilesignificantlyreducingthesize.
- Examples of summaries that are often part of any item are titles, table ofcontents,andabstractswiththeabstractbeingtheclosest.

- The abstract can be used to represent the item for search purposes or as away for a user to determine the utility of an item without having to read thecompleteitem.

IntroductiontoClustering

The goal of the clustering was to assist in the location of information. Clustering ofwords originatedwith the generation of thesauri. Thesaurus, coming from the Latin word meaning “treasure,” is similarto a dictionary in that it stores words. Instead of definitions, it provides the synonyms and antonymsforthewords.Itsprimarypurposeistoassistauthorsinselectionofvocabulary.Thegoalofclusteringis to provide a grouping of similar objects (e.g., terms or items) into a “class” under a more generaltitle. Clustering also allows linkages between clusters to be specified. The term class is frequentlyused asasynonymforthetermcluster.

Theprocessofclusteringfollowsthefollowingsteps:

- Definethedomainfor theclusteringeffort.Definingthedomainfortheclusteringidentifiesthoseobjectstobeusedintheclusteringprocess. Ex:Medicine, Education,Financeetc.
- Oncethedomainisdetermined,determinetheattributesoftheobjectstobeclustered.(Ex:Title, Place,jobetczones)
- Determinethestrengthoftherelationshipsbetweentheattributeswhoseco-occurrenceinobjectssuggestthoseobjectsshouldbeinthesameclass.
- Applysomealgorithmtodeterminetheclass(s)towhicheachitemwillbeassigned.

Classrules:

- ➢ Awell-definedsemanticdefinitionshouldexistforeachclass.
- ➢ Thesizeoftheclassesshouldbeless.
- ➢ Within a class, one object should not dominate the class. For example, assume a thesaurusclass called “computer” exists and it contains the objects (words/word phrases) “microprocessor,”“286-processor,” “386-processor” and “pentium.” If the term

“microprocessor” is found 85 per centof the time and the other terms are used 5 per cent each, there is a strong possibility that using“microprocessor”asasynonymfor“286-processor”willintroducetoomanyerrors.Itmaybebetter toplace“microprocessor”into itsownclass.

- Whether an object can be assigned to multiple classes or just one mustbe decided atcreationtime.

There are additional important decisions associated with the generation of thesauri that are not part ofitemclustering.Theyare

1) Wordcoordinationapproach:specifiesifphrasesaswellasindividualtermsaretobeclustered
2) Word relationships: Aitchison and Gilchrist specified three types of relationships:equivalence, hierarchical and nonhierarchical. Equivalence relationships are the mostcommonandrepresentsynonyms.Hierarchicalrelationshipswheretheclassnameisageneral term and the entries are specific examples of the general term. Theprevious example of“computer” class name and “microprocessor,” “pentium,” etc Non-hierarchical relationships coverothertypesofrelationships suchas“object”-“attribute” thatwouldcontain“employee”and“jobtitle.”
3) Homograph resolution: a homograph is a word that has multiple, completely differentmeanings.Forexample,theterm“field”couldmeanaelectronicfield,afieldofgrass,etc.
4) Vocabularyconstraints:thisincludesguidelinesonthenormalizationandspecificityofthevocabulary.Normalizationmayconstrainthethesaurusto stemsversuscompletewords.

ThesaurusGeneration

There are three basic methods for generation of a thesaurus; hand crafted, co- occurrence, and header-modifier based. In header-modifier based thesauri term relationships are found based upon linguisticrelationships.Words appearinginsimilargrammaticalcontexts areassumedtobesimilar. Thelinguistic parsing of the document discovers the following syntacticalstructures: Subject-Verb, Verb-Object, Adjective-Noun, and Noun-Noun. Each noun has a set of verbs, adjectives and nounsthat itco-occurswith,anda mutualinformationvalueiscalculatedforeachusingtypicallyalogfunction.

ManualClustering

The art of manual thesaurus construction resides in the selection of the set of words to be included. .Care is taken to not include words that are unrelated to the domain of the thesaurus. If a concordanceis used, other tools such as KWOC, KWIC or KWAC may help in determining useful words. A KeyWord Out of Context (KWOC) is another name for a concordance. Key Word In Context (KWIC)displays a possible term in its phrase context. It is structured to identify easily the location of the termunderconsiderationinthesentence.KeyWordAndContext(KWAC)displaysthekeywordsfollowedbytheircontext.

KWOC

TERM	FREQ	ITEM Ids
chips	2	doc2, doc4
computer	3	doc1, doc4, doc10
design	1	doc4
memory	3	doc3, doc4, doc8, doc12

KWIC

chips/	computer design contains memory
computer	design contains memory chips/
design	contains memory chips/ computer
memory	chips/ computer design contains

KWAC

chips	computer design contains memory chips
computer	computer design contains memory chips
design	computer design contains memory chips
memory	computer design contains memory chips

Figure 6.1 Example of KWOC, KWIC and KWAC

IntheFigure6.1thecharacter“/”isusedinKWICtoindicatetheendofthephrase.TheKWIC and KWACareusefulindeterminingthemeaningofhomographs.

Once the terms are selected they are clustered based upon the word relationship guidelines and theinterpretation of the strength of the relationship. This is also part of the art of manual creation of thethesaurus,usingthejudgmentofthehumananalyst.

AutomaticTermClustering

There are many techniques for the automatic generation of term clusters to createstatistical thesauri.Whenthenumberofclusterscreatedisverylarge,theinitialclustersmaybeusedasastartingpoint togenerate more abstract clusters creating a hierarchy. The basis for automatic generation of a thesaurusis a set of items that represents the vocabulary to be included in the thesaurus. Selection of this set ofitems is the first step of determining the domain for the thesaurus. The processing tokens (words) inthesetofitemsaretheattributestobeusedtocreatetheclusters.

Implementation of the other steps differs based upon the algorithms being applied. The automatedmethod of clustering documents is based upon the polythetic clustering where each cluster is definedby a set of words and phrases. Inclusion of an item in a cluster is based upon the similarity of theitem'swordsandphrasesto thoseofotheritemsinthecluster.

CompleteTermRelationMethod

Inthecompletetermrelationmethod,thesimilaritybetweeneverytermpairiscalculatedasabasisfordeter mining the clusters. The easiest way to understand this approach is to consider the vector model.The vector model is represented by a matrix where the rows are individual items and the columns aretheuniquewords(processingtokens)intheitems.Thevaluesinthematrixrepresenthowstronglythatp articularwordrepresentsconceptsinthe item.

Figure 6.2 provides an example of a database with 5 items and 8 terms. To determine the relationshipbetween terms, a similarity measure is required. The measure calculates the similarity between twoterms. In Chapter 7 a number of similarity measures are presented. The similarity measure is notcritical

	Term1	Term2	Term3	Term4	Term5	Term6	Term7	Term8
Item 1	0	4	0	0	0	2	1	3
Item 2	3	1	4	3	1	2	0	1
Item 3	3	0	0	0	3	0	3	0
Item 4	0	1	0	3	0	0	2	0
Item 5	2	2	2	3	1	4	0	2

Figure 6.2 Vector Example

in understanding the methodology so the following simple measure is used:

$$\mathbf{SIM(Term_i, Term_j) = \Sigma\ (Term_{k,i})\ (Term_{k,j})}$$

where “k” is summed across the set of all items. In effect the formula takes the two columns of thetwotermsbeinganalyzed,multiplyingandaccumulatingthevaluesineachrow.Theresultscanbe paced in a resultant “m” by “m” matrix, called a Term-Term Matrix (Salton-83), where “m” is thenumber of columns (terms) in the original matrix. This simple formula is reflexive sothat the matrixthatisgeneratedissymmetric.Othersimilarityformulas couldproduceanon-symmetricmatrix.

Using the data in Figure 6.2, the Term-Term matrix produced is shown in Figure 6.3. There are novalues on the diagonal since that represents the auto correlation of a word to itself. The next step is toselect athreshold that determines if two terms are considered similarenough to each other to be in thesame class. In this example, the threshold value of 10is used.Thus two terms areconsideredsimilar ifthe similarity value between them is 10 or greater. This produces a new binarymatrix called the TermRelationshipmatrix (Figure6.4)thatdefineswhichtermsaresimilar.
A one in the matrix indicates that the terms specified by the column and the row are similar enough
tobeinthesameclass.Term7demonstratesthatatermmayexistonitsownwithnoothersimilartermsidenti fied. In any of the clustering processes described below this term will always migrate to a classby itself.

The final step in creating clusters is to determine when two objects (words) are in the same cluster.There are many different algorithms available. The following algorithms are the mostcommon:cliques,singlelink,starsandconnectedcomponents.

	Term1	Term2	Term3	Term4	Term5	Term6	Term7	Term8
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	

Figure 6.3 Term-Term Matrix

	Term1	Term2	Term3	Term4	Term5	Term6	Term7	Term8
Term 1		0	1	1	1	1	0	0
Term 2	0		0	1	0	1	0	1
Term 3	1	0		1	0	1	0	0
Term 4	1	1	1		0	1	0	0
Term 5	1	0	0	0		0	0	0
Term 6	1	1	1	1	0		0	1
Term 7	0	0	0	0	0	0		0
Term 8	0	1	0	0	0	1	0	

Figure 6.4 Term Relationship Matrix

ApplyingthealgorithmtoFigure6.4,thefollowingclassesarecreated:Class1(Term1,Term3, Term4,Term6)

Class2(Term1,Term5)

Class3(Term2,Term4,Term6)

Class4(Term2,Term6,Term8)

Class5(Term7)

NoticethatTerm1andTerm6areinmorethanoneclass.Acharacteristicofthisapproachisthatterms canbefoundinmultipleclasses.Insinglelinkclusteringthestrongconstraintthateverytermina class is similar to everyother term is relaxed. The rule to generate single link clusters is that anyterm that is similar to any term in the cluster can be added to the cluster. It is impossible for a term tobeintwodifferentclusters.Thisineffectpartitionsthesetoftermsintotheclusters.Thealgorithmis:

1. Selectatermthatisnotinaclassandplaceitinanewclass
2. Placeinthatclassallothertermsthatarerelatedtoit
3. Foreachtermenteredintotheclass,performstep2
4. Whennonewtermscanbeidentifiedinstep2,gotostep1.

ApplyingthealgorithmforcreatingclustersusingsinglelinktotheTermRelationshipMatrix,Figure6.4,thefo llowingclassesarecreated:

Class1(Term 1,Term3,Term4,Term5,Term6,Term2,Term8)

Class2(Term7)

Therearemanyotherconditionsthatcanbeplacedontheselectionoftermstobeclustered.

Questions:

1. WriteaboutClassesofAutomaticIndexing?
2. Writeabouta)Statisticalindexingb)NaturalLanguagec)ConceptIndexingd)HypertextLinkages?
3. WriteaboutDocumentandTermClustering?
4. ExplainaboutThesaurusGeneration?
5. ExplainaboutItemClustering?HierarchyofClusters?

UNIT-4

User Search Techniques: Search statements and binding, Similarity measures and ranking, Relevancefeedback, Selective disseminationof informationsearch,weightedsearches of Booleansystems,SearchingtheInternetandhypertext.**InformationVisualization:**Introduction,Cognitionandperception,Informationvisualizationtechnologies.

SearchStatementsandBinding

Searchstatementsarethestatementsofaninformationneedgeneratedbyuserstospecifytheconceptstheyaretryingtolocatein items.
In generation of the search statement, the user may have the ability to weight (assign an importance) todifferentconceptsinthestatement.Atthispointthebindingistothevocabularyandpastexperiencesoftheuser. Binding in this sense is when a more abstract form is redefined into a more specific form. The searchstatement is the user's attempt to specify the conditions needed to subset logically the total item space tothatclusterofitemsthatcontainstheinformationneededbytheuser.
The next level of binding comes when the search statement is parsed for use by a specific search system.Thefinallevelofbindingcomesasthesearchisappliedtoaspecificdatabase.Thisbindingisbasedupont he statistics of the processing tokens in the database and the semantics used in the database. This isespeciallytrueinstatisticalandconceptindexingsystems.
Figure7.1illustratesthethreepotentialdifferentlevels

INPUT	Binding
"Find me information on the impact of the oil spills in Alaska on the price of oil"	User search statement using vocabulary of user
impact, oil (petroleum), spills (accidents), Alaska, price (cost, value)	Statistical system binding extracts processing tokens
impact (.308), oil (.606), petroleum (.65), spills (.12), accidents (.23), Alaska (.45), price (.16), cost (.25), value (.10)	Weights assigned to search terms based upon inverse document frequency algorithm and database

Figure 7.1 Examples of Query Binding

SimilarityMeasuresandRanking

Avarietyofdifferentsimilaritymeasurescanbeusedtocalculatethe similaritybetweentheitemandthesearchstatement.A characteristic of asimilarityformulais that the results of the formulaincrease as theitems become more similar. The value is zero if the items are totally dissimilar. An example of a simple"sum of the products" similarity measure from the examples in Chapter 6 to determine the similaritybetweendocumentsforclusteringpurposesis:

$$SIM(Item_i, Item_j) = \Sigma\, (Term_{i,k})\, (Term_{j,k})$$

Thisformulausesthesummationoftheproductofthevarioustermsoftwoitemswhentreatingtheindexasa vector. If is replaced with then the same formula generates the similarity between every Item and Theproblemwiththissimplemeasureisinthenormalizationneededtoaccountfor variancesinthelengthofitems.Additionalnormalizationisalsousedtohavethefinalresultscomebetweenzero and+1(someformulasusetherange-1 to+1)

This assumptionof the availabilityof relevance information in theweightingprocesswaslater relaxedbyCroftandHarper(Croft-79). Croftexpandedthisoriginalconcept,takingintoaccountthefrequencyofoccurrenceoftermswithinanitemproducingthefollowingsimilarity formula(Croft-83):

$$SIM(DOC_i, QUERY_j) = \sum_{i=1}^{Q} (C + IDF_i) * f_{i,j})$$

where C is a constant used in tuning, **IDF**$_i$ is the inverse document frequency for term "i" in the collection and

$$f_{i,j} = K + (K - 1)\, TF_{i,j}/maxfreq_j$$

whereKisatuningconstant,isthefrequencyof"i"andisthemaximumfrequencyofanyterminitem"j."The best values for K seemed to range between 0.3 and 0.5. Another early similarity formula was used bySaltonin theSMARTsystem(Salton-83).

Todeterminethe"weight"anitemhas withrespecttothesearchstatement,theCosineformulaisusedtocalculatethedistancebetweenthevectorfortheitemandthevectorforthequery:

$$SIM(DOC_i, QUERY_j) = \frac{\sum_{k=1}^{n} (DOC_{i,k} * QTERM_{j,k})}{\sqrt{\sum_{k=1}^{n} (DOC_{i,k})^2 * \sum_{k=1}^{n} (QTERM_{j,k})^2}}$$

where is the kth term in the weighted vector for Item "i" and is the kth term in query "j." The Cosineformula calculatestheCosineoftheanglebetweenthetwovectors.AstheCosineapproaches"1,"thetwovectors become coincident (i.e., the term and the query represent the same concept). If the two are totallyunrelated,thentheywillbeorthogonalandthevalueoftheCosineis"0."Whatisnottakenintoaccountisthelength ofthevectors

Forexample,ifthefollowingvectorsareinathreedimensional(threeterm)system:Item =(4,8,0)

Query1=(1,2,0)

Query2=(3,6,0)

$$QTERM_{i,k} = (0.5 + (0.5\ TF_{i,k}/maxfreq_k)) * IDF_i$$

$$\mathrm{SIM}(\mathrm{DOC}_i, \mathrm{QUERY}_j) = \frac{\sum_{k=1}^{n} (DOC_{i,k} * QTERM_{j,k})}{\sum_{k=1}^{n} DOC_{i,k} + \sum_{k=1}^{n} QTERM_{j,k} - \sum_{k=1}^{n} (DOC_{i,k} * QTERM_{j,k})}$$

The Dice measure simplifies the denominator from the Jaccard measure and introduces a factor of 2 in the numerator. The normalization in the Dice formula is also invariant to the number of terms in common.

$$\mathrm{SIM}(\mathrm{DOC}_i, \mathrm{QUERY}_j) = \frac{2 * \sum_{k=1}^{n} (DOC_{i,k} * QTERM_{j,k})}{\sum_{k=1}^{n} DOC_{i,k} + \sum_{k=1}^{n} QTERM_{j,k}}$$

QUERY = (2, 2, 0, 0, 4)
DOC1 = (0, 2, 6, 4, 0)
DOC2 = (2, 6, 0, 0, 4)

	Cosine	Jaccard	Dice
DOC1	36.66	16	20
DOC2	36.66	−12	20

Figure 7.2 Normalizing Factors for Similarity Measures

similarityformulaisusedtocalculatesimilaritybetweenthequeryandeachdocument.Ifnothresholdisspecified, all three documents are considered hits. If a threshold of 4 is selected, then only DOC1 isreturned.

Onespecialareaofconcernarisesfromsearchofclustersoftermsthatarestoredina hierarchicalscheme

Vector:	American, geography, lake, Mexico, painter, oil, reserve, subject
DOC1	geography *of* Mexico *suggests* oil reserves *are available* vector (0, 1, 0, 2, 0, 3, 1, 0)
DOC2	American geography *has* lakes *available everywhere* vector (1, 3, 2, 0, 0, 0, 0, 0)
DOC3	painters *suggest* Mexico lakes *as* subjects vector (0, 0, 1, 3, 3, 0, 0, 2)
QUERY	oil reserves *in* Mexico vector (0, 0, 0, 1, 0, 1, 1, 0)

SIM(Q, DOC1) = 6, SIM (Q, DOC2) = 0, SIM(Q, DOC3) = 3

Figure 7.3 Query Threshold Process

The items are stored in clusters that are represented by the centroid for each cluster.

showsaclusterrepresentationofanitemspace.Eachletterattheleaf(bottomnodes)representanitem(i.e.,K, L, M, N, D, E, F, G, H, P, Q, R, J). The letters at the higher nodes (A, C, B, I) represent the centroid oftheir immediate children nodes. The hierarchy is used in search by performing a top-down process. Thequery is compared to the centroids "A" and "B." If the results of the similarity measure are above thethreshold, the query is then applied to the nodes' children. If not, then that part of the tree is pruned and notsearched

The problem comes from the nature of a centroid which is an average of a collection of items (in Physics,the center of gravity). The risk is that the average may not be similar enough to the query for continuedsearch,butspecificitemsusedtocalculatethecentroidmaybecloseenoughtosatisfythesearch.Therisksof missing items and thus reducing recall increases as the standard deviation increases. Use of centroidsreduces the similarity computations but could cause a decrease in recall. It should have no effect onprecisionsincethat isbaseduponthesimilaritycalculationsat theleaf(item)level.

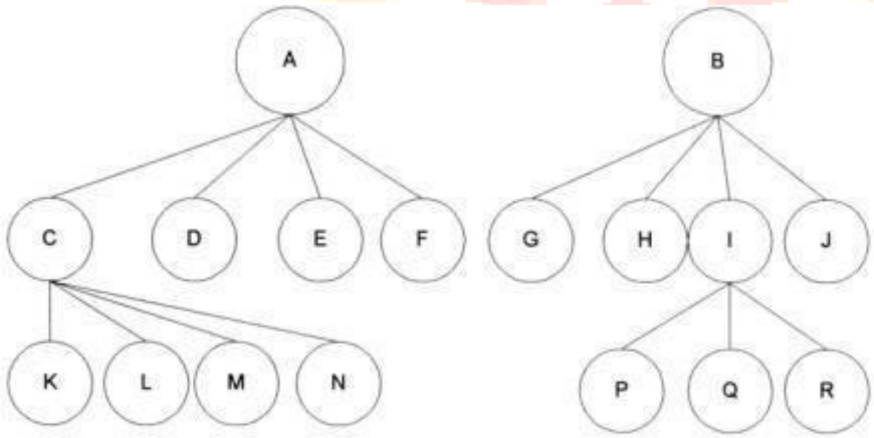


Figure 7.4 Item Cluster Hierarchy

HiddenMarkovModelsTechniques

Use of HiddenMarkovModels forsearching textualcorporahas introduceda new paradigm forsearch. Inmostof thepreviussearch techniques, thequeryis thoughtof as another"document"and thesystem triestofindotherdocumentssimilartoit.InHMMsthedocumentsareconsideredunknownstatisticalprocesses that can generate output that is equivalent to the set of queries that would consider the document relevant.Another way to look at it is by taking the general definition that a HMM is defined by output that isproducedbypassingsomeunknownkeyvia statetransitionsthrougha noisychannel.Theobservedoutputisthequery,andtheunknownkeysaretherelevantdocuments.Thenoisychannelisthe mismatchbetweenthe author's way of expressing ideas and the user's ability to specify his query. Leek, Miller and Schwartz(Leek-99)computedforeachdocumenttheprobabilitythatDwastherelevantdocumentintheusersmindgiventhatQ

$$P(D \text{ is } R/Q) = P(Q/D \text{ is } R) * P(D \text{ is } R) / P(Q)$$

wasthequeryproduced,i.e.,P(DisR/Q).ThedevelopmentforaHMMapproachbeginswithapplyingBayesruletotheconditionalprobability

Thebiggestprobleminusingthisapproachistoestimatethetransitionprobabilitymatrixandtheoutput(queriest hatcouldcausehits)foreverydocumentinthecorpus.

RankingAlgorithms

Aby-productofuseofsimilaritymeasuresforselectingHititemsisavaluethatcanbeusedinrankingtheoutput.Rankingthe outputimpliesorderingtheoutputfrommostlikelyitemsthatsatisfythequerytoleastlikely items. This reduces the user overhead by allowing the user to display the most likely relevant itemsfirst. The original Boolean systems returned items ordered by date of entry into the system versus bylikelihood of relevance to the user's search statement. With the inclusion of statistical similarity techniquesintocommercialsystemsandthelargenumberofhitsthatoriginatefromsearchingdiversecorpora,su chasthe Internet, ranking has become a common feature of modern systems. A summary of ranking algorithmsfromtheresearchcommunityisfoundinanarticlewrittenbyBelkinandCroft(Belkin-87)

<u>RelevanceFeedback</u>

The first major work on relevance feedback was published in 1965 by Rocchio (republished in 1971:Rocchio-71). Rocchio was documenting experiments on reweighting query terms and query expansionbased upon a vector representation of queries and items. The concepts are also found in the probabilisticmodelpresentedbyRobertsonandSparckJones(Robertson-76).Therelevancefeedbackconceptwasthatthenewqueryshouldbebasedontheoldquerymodifiedtoincrease theweightof

$$Q_n = Q_o + \frac{1}{r}\sum_{i=1}^{r} DR_i - \frac{1}{nr}\sum_{j=1}^{nr} DNR_j$$

where

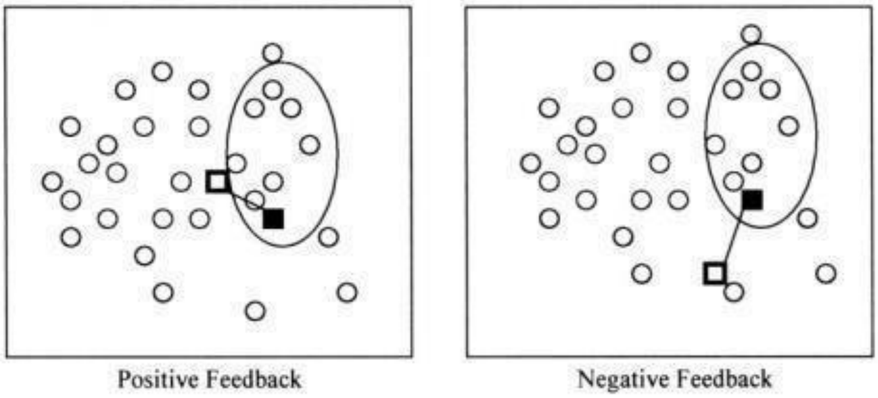
Q_n	= the revised vector for the new query
Q_o	= the original query
r	= number of relevant items
DR_i	= the vectors for the relevant items
nr	= number of non-relevant items
DNR_j	= the vectors for the non-relevant items.

The factors r and nr were later modified to be constants that account for the number of items along with the importance of that particular factor in the equation. Additionally a constant was added to Q_o to allow adjustments to the importance of the weight assigned to the original query. This led to the revised version of the formula:

$$Q_n = \alpha Q_o + \beta\sum_{i=1}^{r} DR_i - \gamma\sum_{j=1}^{nr} DNR_j$$

Termsinrelevantitemsanddecreasetheweightoftermsthatareinnon-relevantitems.Thistechniquenotonly modified the terms in the original query but also allowed expansion of new terms from the relevantitems.Theformulaused is:

Whereandaretheconstantsassociatedwitheachfactor(usually1/nor1/nrtimesaconstant).The factorisreferred to as positive feedback because it is using the user judgments on relevant items to increase thevaluesoftermsforthenextiterationofsearching.



Thefactorisreferredtoasnegative

Figure 7.6 Impact of Relevance Feedback

Relevancefeedback,inparticularpositivefeedback,hasbeenproventobeofsignificantvalueinproducingbetter queries. Some of the early experiments on the SMART system (Ide-69, Ide-71, Salton-83) indicatedthe possible improvements that would be gained by the process. But the small collection sizes andevaluationtechniquesputintoquestiontheactual gainsbyusingrelevancefeedback.

Oneoftheearlyproblemsaddressedinrelevancefeedbackishowtotreatquerytermsthatarenotfoundinany retrieved relevant items. Just applying the algorithm would have the effect of reducing the relativeweightofthosetermswithrespecttootherqueryterms.Fromtheuser'sperspective,thismaynotbedesiredbecause the term may still have significant value to the user if found in the future iterations of the searchprocess.

Harper and van Rijisbergen addressed this issue in their proposed EMIM weighting scheme (Harper-78,Harper-80). Relevance feedback has become a common feature in most information systems. When theoriginalqueryismodifiedbaseduponrelevancefeedback,thesystemsensurethattheoriginalquerytermsare in the modified query, even if negative feedback would have eliminated them. In some systems themodified query is presented to the user to allow the user to readjust the weights and review the new termsadded.

SelectiveDisseminationofInformationSearch

Selective Dissemination of Information, frequently called dissemination systems, are becoming moreprevalent with the growth of the Internet. A dissemination system is sometimes labeled a "push" systemwhile a search system is called a "pull" system. The differences are that in a search system the userproactivelymakesadecisionthatheneedsinformationanddirectsthequerytotheinformationsystemtosearch. In a dissemination system, the user defines a profile (similar to a stored query) and as newinformationisaddedtothesystemitisautomaticallycomparedtotheuser'sprofile.

WeightedSearchesofBooleanSystems

ThetwomajorapproachestogeneratingqueriesareBooleanandnaturallanguage.Naturallanguagequeriesare easily represented within statistical models and are usable by the similarity measures discussed. Issuesarise whenBooleanqueries are associatedwithweightedindexsystems.Some of the issues are associatedwith how the logic (AND, OR, NOT) operators function with weighted values and how weights areassociatedwiththequeryterms.

Iftheoperatorsareinterpretedintheirnormalinterpretation,theyacttoorestrictiveortoogeneral(i.e.,ANDandORoperatorsrespectively).Salton,FoxandWushowedthatusingthestrictdefinitionoftheoperators will sub optimize the retrieval expected by the user (Salton-83a). Closely related to the strictdefinitionproblemisthelackofrankingthatismissingfromapureBooleanprocess.

Some of the early work addressing this problem recognized the fuzziness associated with mixing Booleanand weighted systems (Brookstein-78, Brookstein-80) To integrate the Boolean and weighted systemsmodel,FoxandSharatproposedafuzzysetapproach(Fox-86).Fuzzysetsintroducetheconceptofdegreeofmembershiptoaset(Zadeh-65).ThedegreeofmembershipforANDandORoperationsaredefinedas:

TheMMMtechniquewasexpandedbyPaice(Paice-84)consideringallitemweightsversusthemaximum/minimumapproach.Thesimilaritymeasureiscalculatedas:

$$DEG_{A\cap B} = \min(DEG_A, DEG_B)$$

$$DEG_{A\cup B} = \max(DEG_A, DEG_B)$$

$$SIM(QUERY_{OR}, DOC) = C_{OR1} * \max(DOC1_1, DOC_2, \ldots, DOC_n) + C_{OR2} * \min(DOC_1, DOC_2, \ldots, DOC_n)$$

$$SIM(QUERY_{AND}, DOC) = C_{AND1} * \min(DOC_1, DOC_2, \ldots, DOC_n) + C_{AND2} * \max(DOC1_1, DOC_2, \ldots, DOC_n)$$

$$SIM(QUERY\ DOC) = \sum_{i=1}^{n} r^{i-1} d_i \Big/ \sum_{i=1}^{n} r^{i-1}$$

$$Q_{OR} = (A_1, a_1)\ OR\ (A_2, a_2)\ OR\ \ldots\ OR\ (A_n, a_n)$$

$$Q_{AND} = (A_1, a_1)\ AND\ (A_2, a_2)\ AND\ \ldots\ AND\ (A_n, a_n)$$

SearchingtheINTERNETandHypertext

The Internet has multiple different mechanisms that are the basis for search of items. The primarytechniquesareassociatedwithserversontheInternetthatcreateindexesofitemsontheInternetandallowsearchofthem.SomeofthemostcommonlyusednodesareYAHOO,AltaVistaandLycos.Inallofthesesystems there are active processes thatvisita large numberof Internetsitesandretrieve textual datawhichthey index. The primary design decisions are on the level to which they retrieve data and their generalphilosophyonuseraccess.

LYCOS (http://www.lycos.com) and AltaVista automatically go out to other Internet sites and return thetextatthesitesforautomaticindexing(http://www.altavista.digital.com).Lycosreturnshomepagesfromeach site for automatic indexing while Altavista indexes all of the text at a site. The retrieved text is thenusedtocreateanindextothesourceitemsstoringthe

UniversalResourceLocator(URL)toprovidetotheuser to retrieve an item. All of the systems use some form of ranking algorithm to assist in display of theretrieved items. The algorithm is kept relatively simple using statistical information on the occurrence ofwordswithin theretrievedtext

Closelyassociatedwiththecreationoftheindexesisthetechnique

foraccessingnodesonTherearesixkeycharacteristicsofintelligentagents(Heilmann-96):

1. Autonomy - the search agent must be able to operate without interaction with a human agent. It musthavecontroloveritsowninternalstatesandmakeindependentdecisions.Thisimpliesasearchcapabilitytotraverseinformationsitesbaseduponpre-establishedcriteriacollectingpotentiallyrelevantinformation.
2. Communications Ability - the agent must be able to communicate with the information sites as ittraversesthem.Thisimpliesauniversallyacceptedlanguagedefiningtheexternalinterfaces(e.g.,Z39.50).
3. CapacityforCooperation- thisconceptsuggeststhatintelligentagentsneedtocooperatetoperformmutuallybeneficialtasks.
4. Capacity for Reasoning - There are three types of reasoning scenarios (Roseler-94): Rule-based - whereuserhasdefinedasetofconditionsandactionstobetakenKnowledge-based- wheretheintelligentagentshave stored previous conditions and actions taken which are used to deduce future actions Artificialevolution based - where intelligent agents spawn new agents with higher logic capability to perform itsobjectives.
5. AdaptiveBehavior- closelytiedto1and4,adaptivebehaviorpermitstheintelligentagenttoassessitscurrentstateandmakedecisionsontheactionsitshouldtake
6. Trustworthiness- theusermusttrustthattheintelligentagentwillactontheuser’sbehalftolocateinformationthattheuserhasaccesstoand isrelevanttotheuser.

InformationVisualization

Functionsthatareavailablewithelectronicdisplayandvisualizationofdatathatwerenotpreviouslyprovided are:

- modifyrepresentationsofdataandinformationorthedisplaycondition(e.g.,changingcolorscales)
- usethesamerepresentationwhileshowingchangesindata(e.g.,movingbetweenclustersofitemsshowingnewlinkages)
- animatethedisplaytoshowchangesinspaceandtime
- Createhyperlinksunderusercontroltoestablishrelationshipsbetweendata

Information Visualization addresses how the results of a search may be optimally displayed to the users tofacilitate their understanding of what the search has provided and their selection of most likely items ofinteresttoread.Cognitive(thementai actionorprocessof acquiringknowledge and understanding throughthought,experience, and thesenses)engineeringderives designprinciplesfor visualization techniquesfromwhatweknowabouttheneuralprocessesinvolved

withattention,memory,imageryandinformationprocessingofthehumanvisualsystem.
Cognitive engineering results can be applied to methods of reviewing the concepts contained in itemsselected by search of an information system. Visualization can be divided into two broad

classes: linkvisualizationand attribute(concept)visualization.Linkvisualizationdisplays relationships among items.Attributevisualizationrevealscontent relationshipsacrosslargenumbersofitems. Therearemanyareasthatinformationvisualizationandpresentationcanhelptheuser:

a. reducetheamountoftimetounderstandtheresultsofasearchandlikelyclustersofrelevantinformation

b. yieldinformationthatcomes fromtherelationshipsbetweenitemsversustreatingeachitemasindependent

c. performsimpleactionsthatproducesophisticatedinformationsearchfunctions

Visualizationisthetransformationofinformationintoavisualformwhichenablestheusertoobserveandunderstandtheinformation.

Cognition (thementalactionorprocessof acquiringknowledge andunderstanding through thought,experience,and thesenses)

Perception (the ability to see, hear, or become aware of something through the senses)Proximity -nearbyfiguresaregroupedtogetherSimilarity-similarfiguresaregroupedtogether
Continuity-figuresareinterpretedassmoothcontinuouspatternsratherthandiscontinuousconcatenationsof shapes (e.g., a circle with its diameter drawn is perceived as two continuous shapes, a circle and a line,versustwo halfcirclesconcatenated together)

Closure - gaps within a figure are filled in to create a whole (e.g., using dashed lines to represent a squaredoesnotpreventunderstandingitasasquare)Connectedness-uniformandlinked spots,linesorareasareperceivedasasingleunit.

AspectsoftheVisualizationProcess

Oneof thefirst-levelcognitiveprocesses is preattention,thatis,taking thesignificantvisual informationfrom the photoreceptors and forming primitives. In Figure 8.1 the visual system detects the difference inorientations between the left and middle portion of the figure and determines the logical border betweenthem.AnexampleofusingtheconsciousprocessingcapabilitiesofthebrainisthedetectionofthedifferentshapedobjectsandtheborderbetweenthemshownbetweentheleftsideandmiddleoftheFigure8.1.Thereadercanlikelydetectthedifferencesinthetimeittakestovisualizethetwodifferentboundaries. Thepreattentiveprocesscandetecttheboundariesbetweenorientationgroupsofthesameobject.Aharderprocess is to identify the equivalence of rotated objects. For example, a rotated square requires more effortto recognize it as a square. As we migrate into characters, the problem of identification of the character isaffectedbyrotatingthecharacterinadirectionnotnormallyencountered.It iseasiertodetectthesymmetrywhentheaxisisvertical.Figure8.2demonstratestheseeffects.

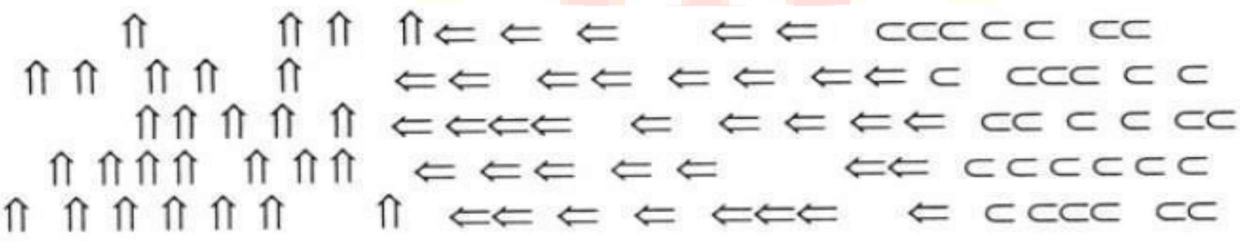


Figure 8.1 Preattentive Detection Mechanism

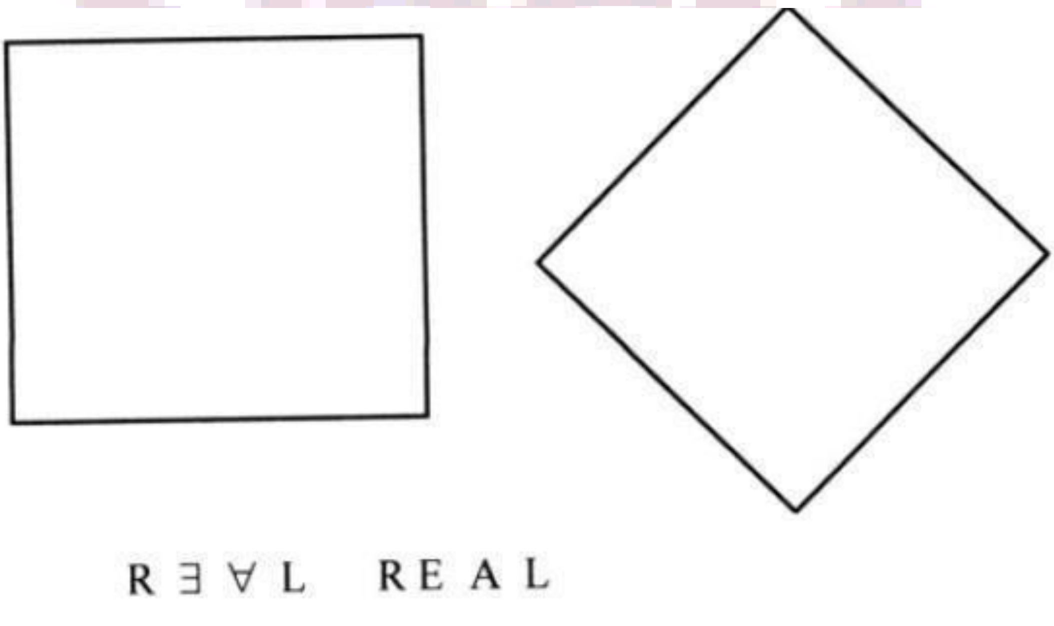


Figure 8.2 Rotating a Square and Reversing Letters in “REAL”

Colorisoneofthemostfrequentlyusedvisualizationtechniquestoorganize,classify,andenhancefeatures.Theg oalsfordisplayingtheresult fromsearchesfallintotwomajorclasses:documentclusteringandsearchstatementanalysis.Thegoalofdocumentclusteringistopresenttheuserwithavisualrepresentationofthedocumentspaceconstrainedbythesearchcriteria.Withinthisconstrainedspacethereexistclustersof documentsdefinedbythedocumentcontent.Visualizationtoolsinthisareaattempttodisplaytheclusters,withanindicationoftheirsizeandtopic,asabasisforuserstonavigatetoitemsofinterest.The second goal is to assist the user in understanding why items were retrieved, thereby providinginformationneededtorefinethequery.Visualizationtechniquesapproachthisproblembydisplayingthetotalsetof terms,including additionalterms from relevance feedback orthesaurus expansion, alongwithdocumentsretrievedandindicatetheimportanceofthetermtotheretrieval andrankingprocess. Link analysis is also important because it provides aggregate-level information within an informationsystem.Onewayoforganizinginformationishierarchical.Atwo-dimensionalrepresentationbecomesdifficultforausertounderstand as

The hierarchy becomes large. One of the earliest experiments in information visualization was theInformationVisualizerdevelopedbyXEROXPARC.ItincorporatesvariousvisualizationformatssuchasDataMap,

InfoGrid, ConeTree, and the Perspective wall. The Cone-Tree is a 3-Dimensional representation of data,whereonenodeofthetreeisrepresentedattheapexandailtheinformationsubordinatetoitisarrangedinacircularstructureatitsbase.

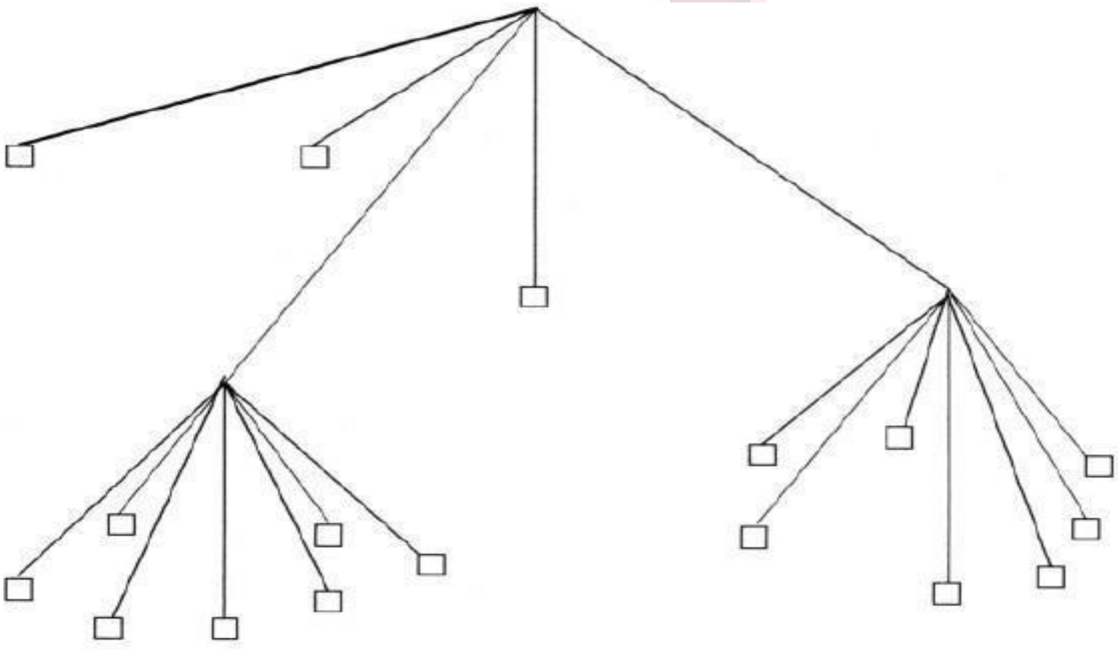


Figure 8.4 Cone Tree

Thusasix-dimensionalcoordinatespacemayhavethreeofthecoordinatesdefinedasasubspacewithintheotherthreecoordinatespaces.ThishasbeencalledFeiner’s“worldswithinworlds”approach

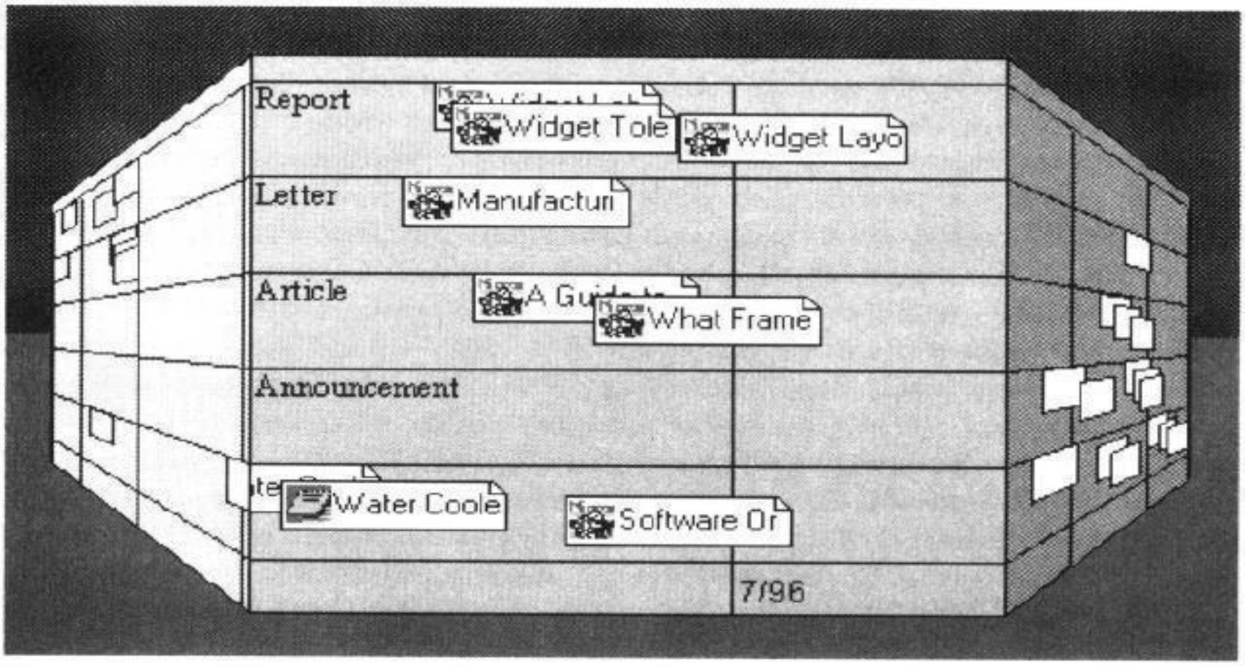


Figure 8.5 Perspective Wall
From inXight web site - www.inxight.com

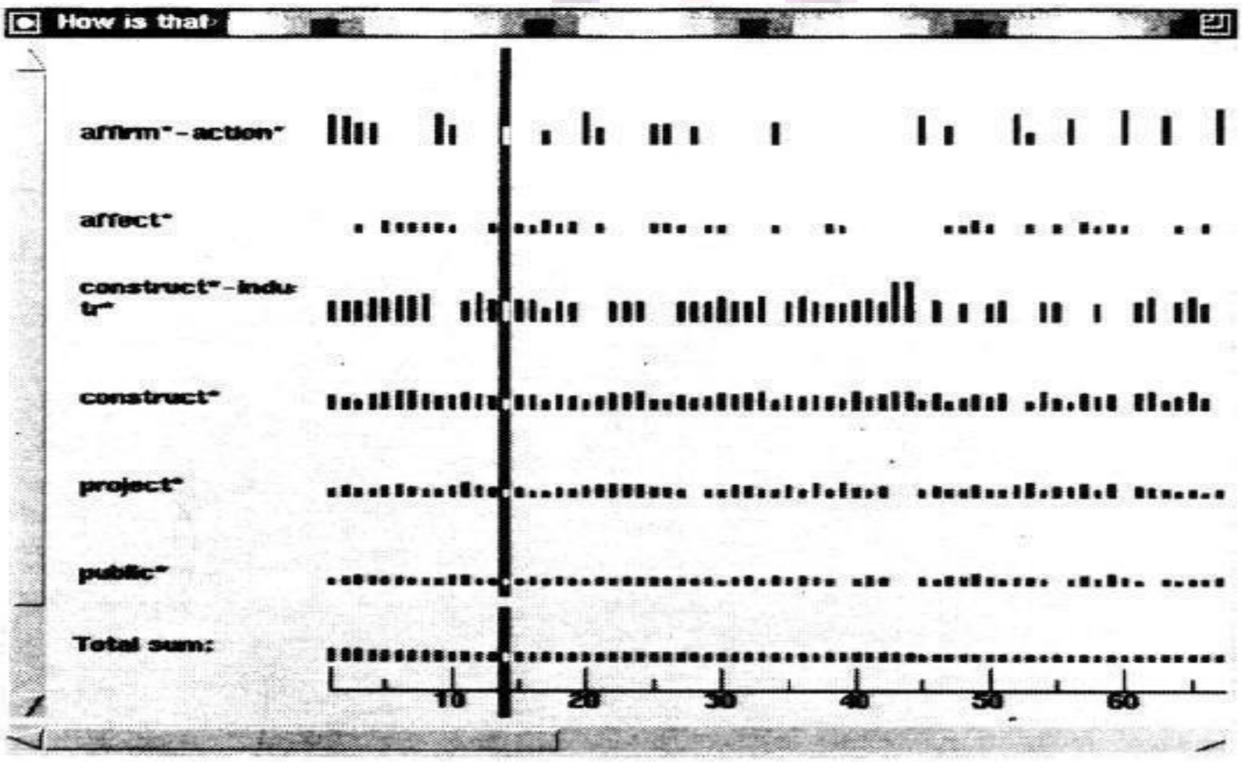


Figure 8.8 Visualization of Results
(from SIGIR 96, page 88)

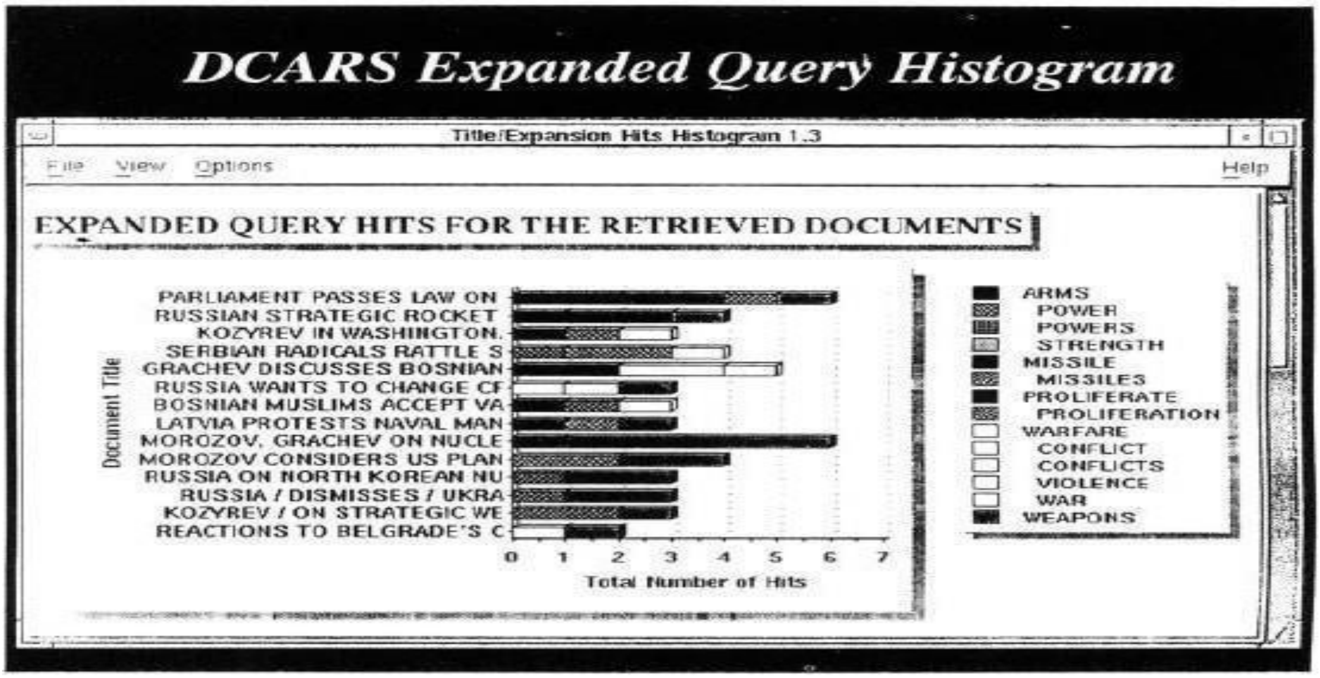


Figure 8.9 Example of DCARS Query Histogram
(from briefing by CALSPAN)

A slightly different commercial version having properties similar to the systems above is theDocument Content Analysis and Retrieval System (DCARS) being developed by CalspanAdvanced Technology Center.Theirsystem isdesigned to augmenttheRetrievalWaresearchproduct.Theydisplaythequeryresultsasahistogramwiththeitems asrowsandeachterm'scontributiontotheselectionindicatedbythewidthofatilebarontherow(seeFigure8.9). DCARSprovidesafriendlyuserinterfacethatindicateswhyaparticularitemwasfound,butitismuchharder tousetheinformationindetermininghowtomodifysearchstatementstoimprovethem.

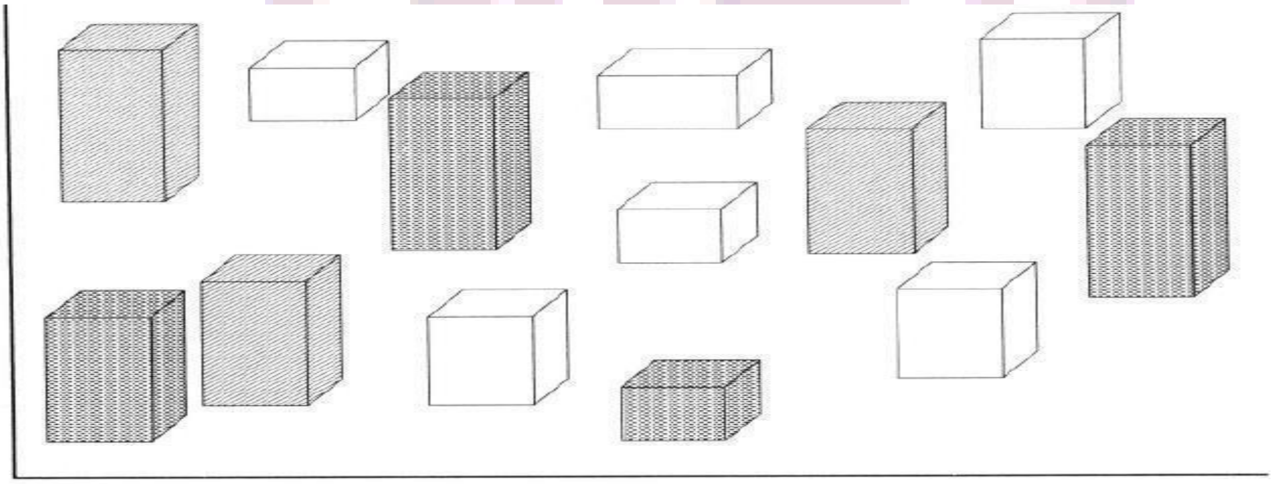


Figure 8.10 CityScape Example

Questions:

1. WriteaboutSearchStatementsandBinding?
2. WriteaboutsimilarityMeasuresandRanking?
3. WhatisRelevanceFeedback?Explainwithexample?
4. ExplainaboutInformationVisualization?
5. ExplainaboutCognitionandPerception?InformationVisualizationTechnologies?

UNIT-V

TextSearchAlgorithms: Introduction,Software text searchesAlgorithms,Hardwaretext searchsystems.

Multimedia Information Retrieval:SpokenLanguageAudio Retrieval,Non-SpeechAudioRetrieval,GraphRetrieval,ImageryRetrieval,VideoRetrieval

TextSearchAlgorithms

Threeclassicaltextretrievaltechniqueshavebeendefinedfororganizingitemsinatextual database,forrapidlyidentifyingtherelevantitemsand foreliminatingitemsthatdonotsatisfythesearch.

Thetechniquesare
1) Fulltextscanning(streaming)
2) Wordinversion
3) Multiattributesretrieval

Inadditiontousingtheindexesasamechanismforsearchingtextininformationsystems,streamingoftextwasfrequentlyfoundinthesystemsasanadditionalsearchmechanism.
Thebasicconceptofatextscanningsystemistheabilityforoneor moreuserstoenterqueries,andthetextto be searched is accessed and compared to the query terms. When all of the text has been accessed, thequeryiscomplete.

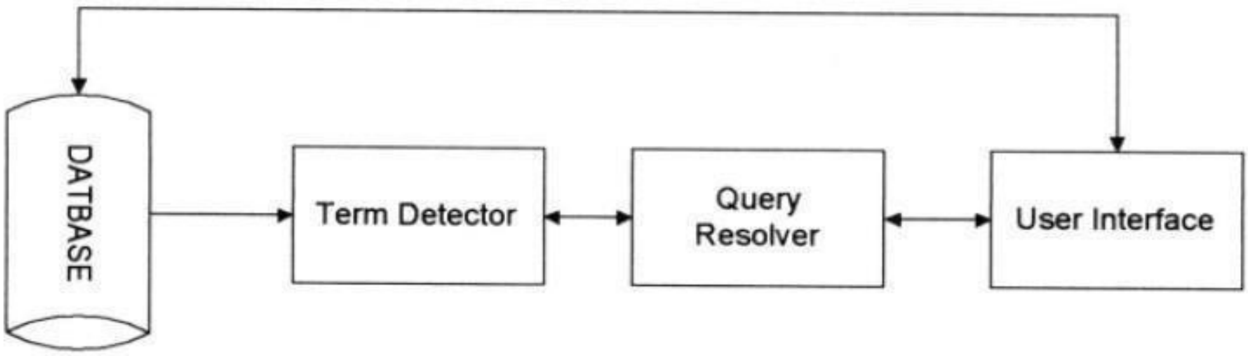


Figure 9.1 Text Streaming Architecture

Thedatabasecontainsthefulltextoftheitems.Thetermdetectoristhespecialhardware/software thatcontainsallofthetermsbeingsearchedforandinsomesystemsthelogicbetweentheitems.Itwillinputthe textand detecttheexistenceofthesearch terms.
Itwilloutputtothequeryresolverthedetectedtermstoallowforfinallogicalprocessingofaqueryagainstanitem.The queryresolverperformstwofunctions.

Itwillacceptsearchstatements fromtheusers,extractthelogicandsearchtermsandpassthe search terms to the detector. It also accepts results from the detector and determines which queries aresatisfied by the item and possibility the weight associated with hit. The Query Resolver will passinformation to the user interface that will be continually updating search status to the user and on requestretrieveanyitemsthatsatisfytheusersearchstatement.Theworstcasesearch forapatternof*m*charactersinastringof*n*charactersisatleast$n$-m+1oramagnitudeofO(n).

SomeoftheoriginalbruteforcemethodscouldrequireO(n*m)symbolcomparisons.Morerece ntimprovementshavereduced thetimetoO(n +m).

Inthecaseofhardwaresearchmachines,multipleparallelsearchmachines(termdetectors) mayworkagainstthesamedatastreamallowingformorequeriesoragainstdifferentdatastreamsreducing thetimetoaccessthecompletedatabase.Insoftwaresystems,multipledetectorsmay executeatthesametime.

Therearetwoapproachestothedatastream.Inthefirstapproachthecompletedatabaseisbeingsenttothedetector(s)functioningasasearchofthe database. In the second approach random retrieved items are being passed to the detectors. In thissecondcasetheideaistoperformanindexsearchofthedatabaseandletthetextstreamerperformadditionalsearchlogicthatisnotsatisfied bytheindexsearch.

Examplesoflimitsofindexsearchesare:Search
- forstopwords
- Searchforexactmatcheswhenstemmingisperformed
- Searchfortermsthatcontainbothleadingandtrailing“don’tcares”Searc
- hforsymbolsthatareontheinter-wordsymbollist(e.g.,“,;)

The full text search function does not require anyadditional storage overhead. There is also theadvantagewherehitsmaybereturnedtotheuserassoonasfound.Typicallyinanindexsystem,thecompletequerymustbeprocessedbefore anyhits aredeterminedoravailable.Streamingsystemsalso provide a very accurate estimate of current search status and time tocomplete the query. It isdifficulttolocateallthepossibleindexvaluesshortofsearchingthecompletedictionaryofpossibleterms.
Manyofthehardwareandsoftwaretextsearchersusefinitestateautomataasabasisfortheiralgorithms.A finitestateautomataisalogicalmachinethatiscomposedoffiveelements:
I-a setofinputsymbolsfromthealphabetsupportedbytheautomata
S-asetofpossiblestates
P-asetofproductionsthatdefinethenextstatebaseduponthe currentstateandinputsymbol
S0-aspecialstatecalledtheinitialstate
SF-asetof oneor morefinalstatesfromtheset**S**

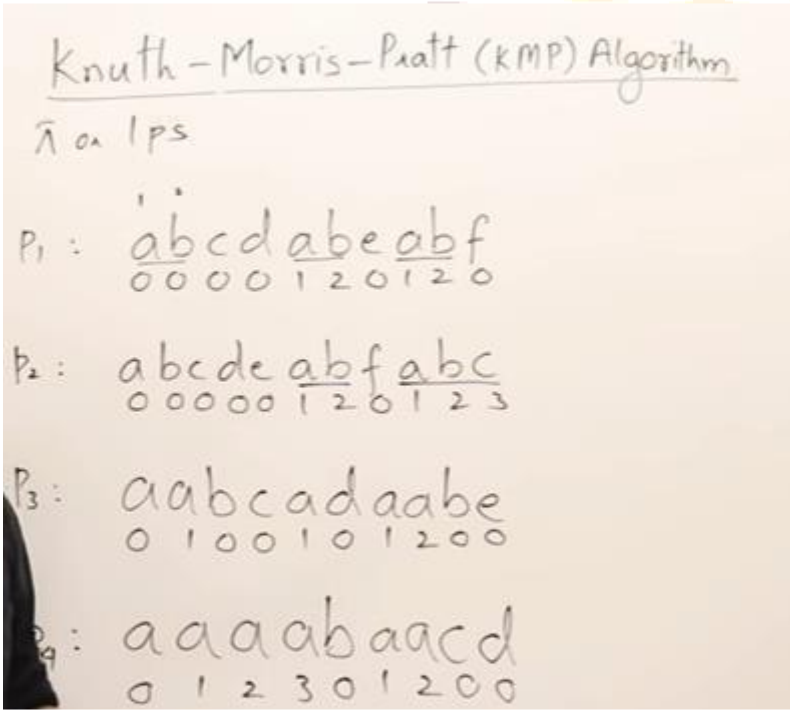
SoftwareTextSearchAlgorithms
Insoftwarestreamingtechniques,theitemtobesearchedisreadintomemoryandthenthealgorithmisapplied.
Therearefourmajoralgorithmsassociatedwith softwaretextsearch:
1) thebruteforceapproach
2) Knuth-Morris-Pratt
3) Boyer-Moore,Shift-ORalgorithm
4) Rabin-Karp.

Ofallofthealgorithms,Boyer-MoorehasbeenthefastestrequiringatmostO(n+ m)comparisons,Knuth-Pratt-MorrisandBoyer-MoorebothrequireO(n)preprocessingofsearchstringsTheBruteforceapproachisthesimpleststringmatchingalgorithm.Theideaistotryand matchthesearchstringagainsttheinputtext.Ifassoonasamismatchisdetectedinthecomparisonprocess,shift the input text one position andstart the comparison process over. The expectednumber of comparisonswhensearchinganinputtextstringof$n$charactersforapatternof$m$charactersis

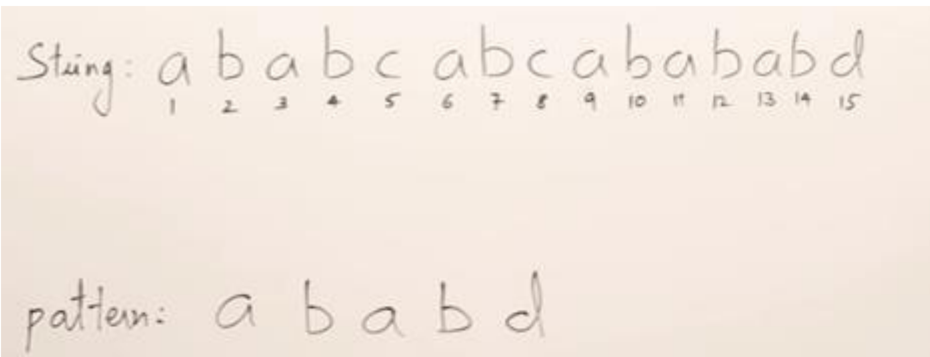
Nc=c/c-1(1-1/cm)*(n-m+1)+O(1)
WhereNcistheexpectednumber ofcomparisonsandcisthesizeofthealphabetfor thetext.

Knuth-Pratt-Morris(KPM)algorithm
Pattern:
abcdabC12
34567

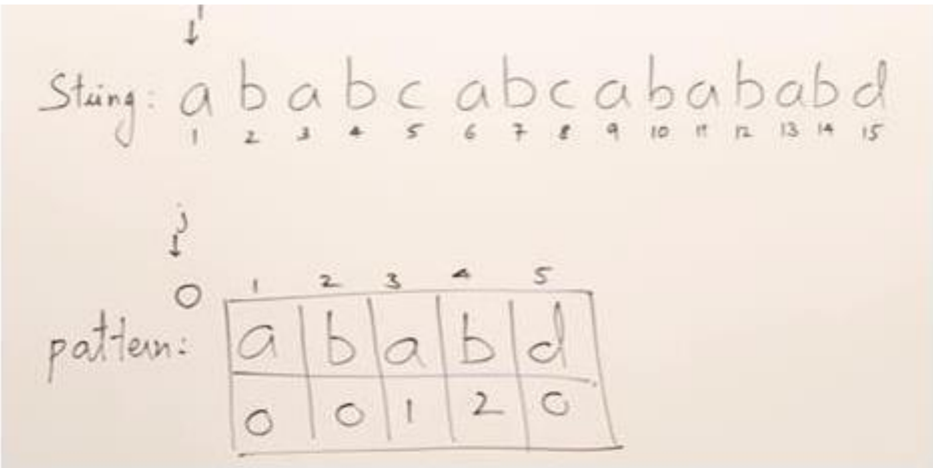
Nowfindoutsubstringsasprefix,suffixbytakinganynumberofcharactersfromlefttorightandrightt
o left.
Prefix:a,ab,abc,abcdetcS
uffix:c,bc,abc,dabcetc
Fromaboveprefix,suffixsubstringswecanobserveasubstring“abc”isthereinbothandalsothatisrepeate
dtwicein given pattern.



Example:
Givenstringandpatternis



Nowconstructtablewithrepeatedcharactersofpatternlikefollowing.



Here‘a’isrepeatedsokeepindex1belowit,‘b’isrepeatedsokeepindex2belowitandrestall ‘0’s.Nowstartsearchprocess.

Step-1:InitializestringindexasI,patternindexasj.Startjfromadding‘0’index.

Step-2:Comparestring[i]andpattern[j+1]i,e,‘a’and‘a’botharematchingsomovebothiandjtonextposition.

Step-3:Nowcomparestring[i]andpattern[j+1]i.e,bandbmatchingsomovebothi,jtonextposition Wheni=5andj=4string[i]=candpattern[j+1]=d.herenotmatchingthenmovejtoitsindexlocation.i.e,2.So nowjpositionispattern[2].

Nowcomparestring[i]=candpattern[j+1]=a.

Herenotmatchingthenmovejtoitsindexlocation i.e.,0.Sonowjpositionispattern[0].Jisnowon0positionsowecannotmovethereforemovenowIto nextlocation i.e.6.

Note:Herewecanobserveonlyjismovingbackbutnoti.Iismovingonlyintheforwarddirection.

Step-4:Repeattheprocesstillfindamatch

Boyer MooreAlgorithm:

Step-1:Construct'BadMatchTable'

Step-2:Comparerightmostcharacterofpatternwithgivenstringbasedonthe'value'ofbadmatchtable

Step-3:Ifmismatchthenshiftthepatterntotherightpositioncorrespondingtothe'value'ofbadmatch table

Whileconstructingbadmatchtableusefollowingformulaforvalue.value=lengthofpattern-index-1andlastvalue=lengthofpattern

Constructing Bad Match Table

T E A M M A S T   Length = 8

Index : 0 1 2 3 4 5 6 7

Letter	T	E	A	M	S	*
Value	7	6	5			

A = 8 - 2 - 1

Heretheletter'A'isoccurringtwicesoreplacethelatestvaluebyoldone.Inthes amewayforMalso.
Tisthelastcharacterinpatternsoitsvalue=8(lengthofpattern)

Constructing Bad Match Table

T E A M M A S T   Length = 8

Index : 0 1 2 3 4 5 6 7

Letter	T	E	A	M	S	*
Value	7	6	5	4		

M = 8 - 3 - 1

Constructing Bad Match Table

T E A M M A S T   Length = 8

Index : 0 1 2 3 4 5 6 7

Letter	T	E	A	M	S	*
Value	7	6	5	3		

M = 8 - 4 - 1

Constructing Bad Match Table

T E A M M A S T   Length = 8

Index : 0 1 2 3 4 5 6 7

Letter	T	E	A	M	S	*
Value	8	6	2	3	1	

T = 8   Last letter = length, if not already defined.

Boyer Moore Example

Letter	T	E	A	M	S	*
Value	8	6	2	3	1	8

W E L C O M E T O T E A M M A S T

T E A M M A S T

Mismatchheresomove8charactersrighthandside

Boyer Moore Example

Letter	T	E	A	M	S	*
Value	8	6	2	3	1	8

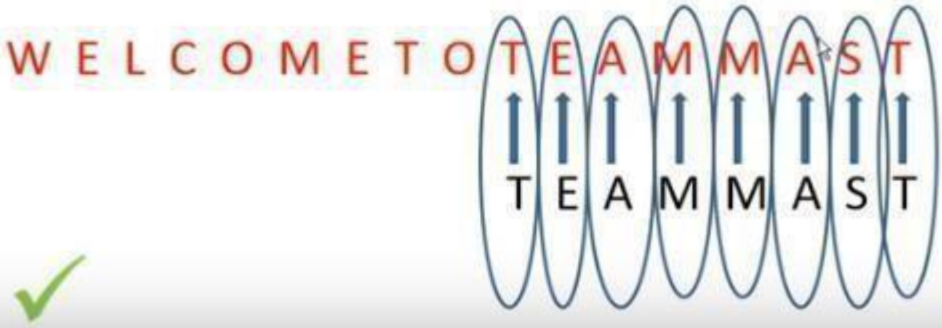
W E L C O M E T O T E A M M A S T

T E A M M A S T

Mismatchsomove1charactertotherighthandside

Boyer Moore Example

Letter	T	E	A	M	S	*
Value	8	6	2	3	1	8



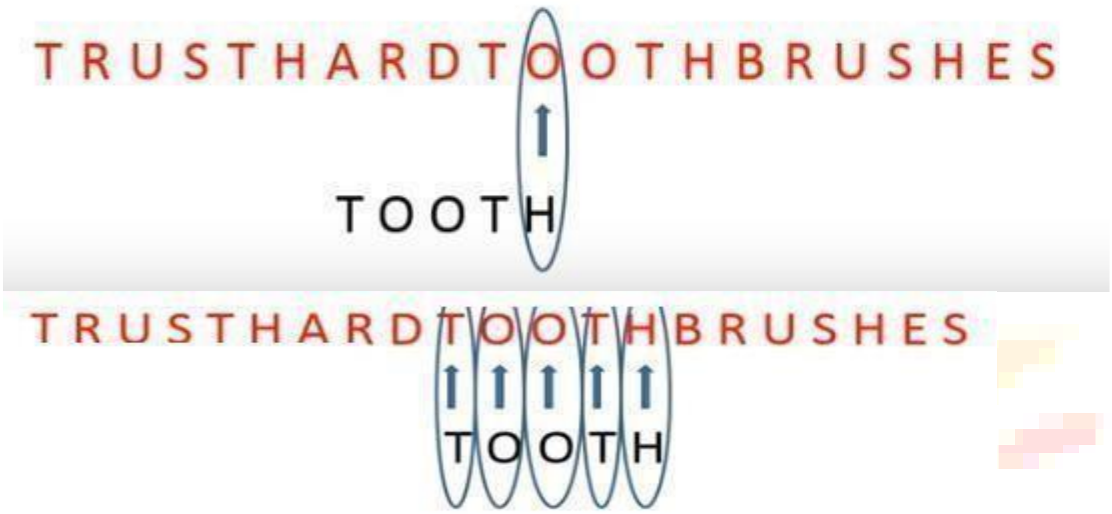
Letter	T	O	H	*
Value	1	2	5	5

Boyer Moore Example

Letter	T	O	H	*
Value	1	2	5	5

Boyer Moore Example

Letter	T	O	H	*
Value	1	2	5	5



HardwareTextSearchSystems
Softwaretextsearchisapplicabletomanycircumstancesbuthasencounteredrestrictionsontheability to handle many search terms simultaneously against the same text and limits due to I/Ospeeds. One approach that off loaded the resource intensive searching from the main processorswas to have a specialized hardware machine to perform the searches and pass the results to themain computer which supported the user interface and retrieval of hits. Since the searcher ishardwarebased,scalabilityisachievedbyincreasingthenumberofhardwaresearchdevices. Anothermajoradvantageofusingahardwaretextsearchunitisintheeliminationofthe indexthatrepresentsthedocumentdatabase.Typicallytheindexesare70%thesizeoftheactual items.Otheradvantagesarethatnewitemscanbesearchedassoonasreceivedbythesystemratherthanwaitingfor theindextobecreatedandthesearchspeed isdeterministic.

Thearithmeticpartofthesystemisfocusedonthetermdetector.Thereha

sbeenthreeapproachestoimplementingtermdetectors:
1. parallelcomparatorsorassociativememory,
2. acellularstructure,and
3. auniversalfinitestateautomata.

Whenthetermcomparatorisimplementedwithparallelcomparators,eachterminthe queryisassignedtoanindividualcomparisonelementandinputdataareseriallystreamedintothedetector.

Whenamatchoccurs,thetermcomparatorinformstheexternalqueryresolver(usuallyinthemaincomputer)by setting statusflags.

Specialized hardware that interfaces with computers and is used to search secondary storagedeviceswasdeveloped fromtheearly1970swiththemostrecentproductbeingthe**ParallelSearcher(previouslythe FastData Finder).**

ThetypicalhardwareconfigurationisshowninFigure9.9inthedashedbox.Thespeedofsearchisthenbased onthespeed oftheI/O.

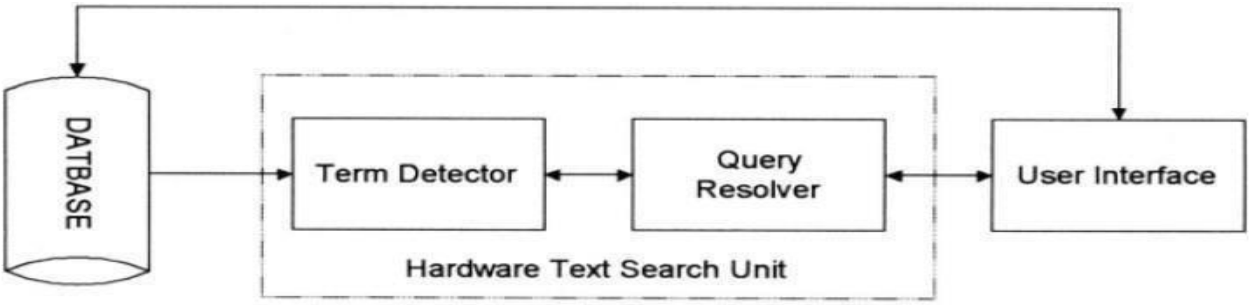


Figure 9.9 Hardware Text Search Unit

Oneoftheearliesthardwaretextstringsearchunitswasthe***RapidSearch***MachinedevelopedbyGeneralElectric. The machine consistedof aspecialpurpose search unitwhere asingle querywas passedagainst amagnetic tape containing the documents. A more sophisticated search unit was developed by OperatingSystemsInc.calledthe***AssociativeFileProcessor(AFP).***

Itiscapableofsearchingagainstmultiplequeriesatthesametime.Followingthatinitialdevelopment,OSI,using adifferentapproach,developedthe***High SpeedTextSearch(HSTS) machine***.ItusesanalgorithmsimilartotheAho-Corasick softwarefinitestatemachinealgorithmexceptthat it runs three parallel state machines. One state machine is dedicated to contiguous word phrases,anotherforimbeddedtermmatchandthefinalforexactwordmatch.

Inparallel with that development effort, GE redesigned their Rapid Search Machine into the***GESCAN unit***. The GESCAN system uses a text array processor (TAP) that simultaneouslymatches many terms and conditions against a given text stream the TAP receives the queryinformationfromtheuser'scomputeranddirectlyaccessthetextualdatafromsecondarystorage.

The TAP consists of alarge cachememory and an array of four to 128 query processors. The text is loadedintothecacheandsearchedbythequeryprocessors(Figure9.10).Eachqueryprocessorisindependenta ndcanbeloadedatanytime. Acompletequeryishandledbyeachqueryprocessor. Aqueryprocessorworkstwooperationsinparallel;matchingquerytermstoinputtextandBool

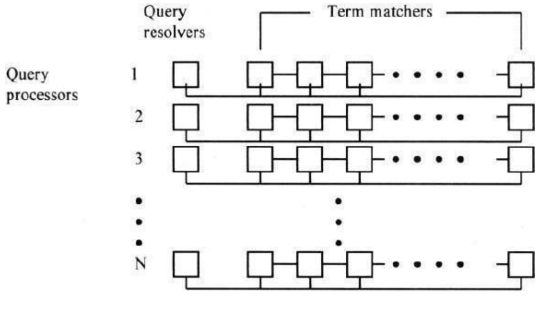


Figure 9.10 GESCAN Text Array Processor

eanlogicresolution.
Termmatchingisperformedbya seriesofcharactercellseachcontainingone characterofthequery. AstringofcharactercellsisimplementedonthesameLSIchipandthechipscanbeconnectedinseriesforlongerstr

ings.Whenawordorphraseofthequeryismatched, asignalis senttotheresolutionsub-processontheLSIchip.TheresolutionchipisresponsibleforresolvingtheBoolean logicbetweentermsandproximity requirements. Iftheitemsatisfiesthequery,theinformationistransmittedtotheuserscomputer.

ThetextarrayprocessorusesthesechipsinamatrixarrangementasshowninFigure9.10. Each row of the matrix is aquery processorinwhich thefirstchip performsthe query resolutionwhiletheremainingchipsmatchqueryterms.Themaximumnumberofcharactersinaqu eryis restrictedbythelengthofarowwhilethenumberofrowslimitthenumberofsimultaneousqueriesthatcanbeproce ssed.

Another approach for hardware searchers is to augment disc storage. The *augmentation is ageneralizedassociativesearch* elementplacedbetweenthereadandwriteheadsonthedisk.Thecontentaddressablesegment sequentialmemory(CASSM)systemusesthesesearchelementsinparalleltoobtainstructureddatafroma database.TheCASSMsystem wasdevelopedatthe UniversityofFloridaasa generalpurposesearchdevice.Itcanbeusedtoperformstringsearchingacrossthedatabase.

Anotherspecialsearchmachineisthe***relationalassociativeprocessor(RAP)***developedattheUniversityof Toronto. Like CASSM performs search across a secondary storage device using a series of cellscomparingdatain parallel.

The***FastData Finder(FDF)***isthemostrecentspecializedhardwaretextsearchunitstillin useinmanyorganizations.ItwasdevelopedtosearchtextandhasbeenusedtosearchEnglishandforeignlanguag es.

TheearlyFastDataFindersconsistedofanarrayofprogrammabletextprocessingcellsconnectedinseriesformingapi pelinehardwaresearchprocessor.

ThecellsareimplementedusingaVSLIchip.IntheTRECtestseachchipcontained24processorcellswitha typical system containing 3600 cells. Each cell will be a comparator for a single character limiting thetotalnumberofcharactersinaquerytothenumberofcells.

Thecellsareinterconnectedwithan8-bitdata pathandapproximately20-bitcontrolpath. Thetexttobesearchedpassesthrougheachcellina pipeline fashionuntilthe completedatabasehas been searched. As data is analyzed at each cell, the 20 control lines states are modifieddependingupontheircurrent stateandtheresultsfromthecomparator.

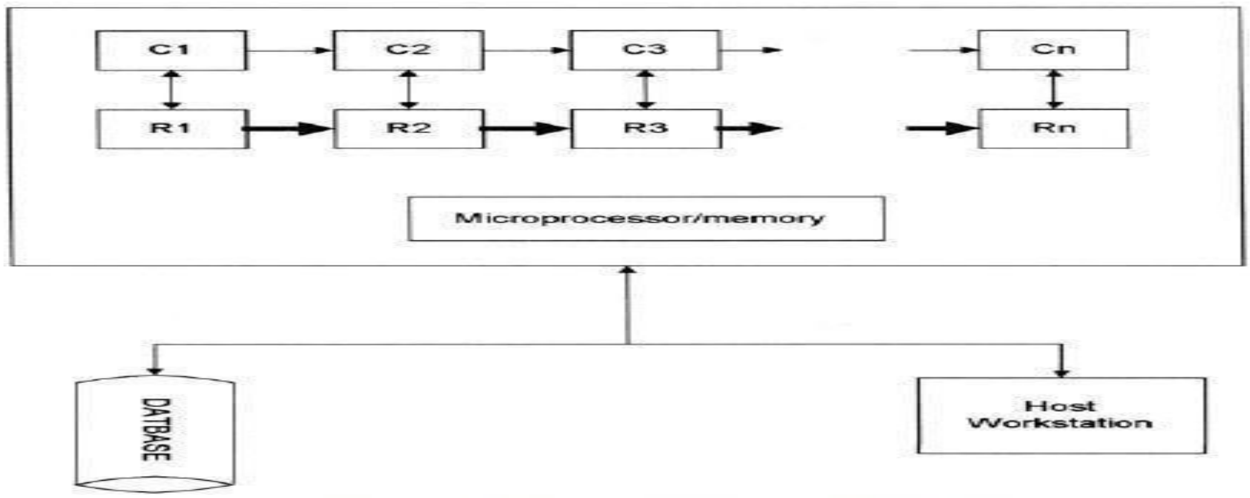


Figure 9.11 Fast Data Finder Architecture

Acelliscomposedofbotharegister cell(Rs)andacomparator (Cs).Theinputfromthe

Document database is controlled and buffered by the micro process/memory and feed through thecomparators. Thesearchcharacters are stored inthe registers. Theconnectionbetweenthe registers reflectsthecontrol linesthatarealsopassingstateinformation.Groupsofcellsareusedtodetectqueryterms,alongwithlogicbetweentheterms,byappropriateprogrammingofthecontrollines.

Whena patternmatchisdetected,ahitispassedtotheinternalmicroprocessorthatpassesitbacktothehostprocessor, allowingimmediateaccessby theusertotheHititem. ThefunctionssupportedbytheFastdataFinderare:

- BooleanLogicincludingnegation
- Proximity on an arbitrary
- patternVariable lengthex:“don’t
- cares”Term counting and
- thresholdsFuzzymatching
- Term
- weightsNumericranges

Multimediainformationretrieval(MMIRorMIR)

Datasourcesincludedirectlyperceivablemediasuchasaudio,imageandvideo,indirectlyperceivablesources such as text, semantic descriptions, bio signals as well as not perceivable sources such as bioinformation,stock prices,etc.

ThemethodologyofMMIRcanbeorganizedinthreegroups:

1. Methodsforthesummarizationofmediacontent(featureextraction).Theresultoffeatureextractionisa description.
2. Methodsforthefilteringofmediadescriptions(forexample,eliminationofredundancy)

3. Methodsforthecategorizationofmediadescriptionsintoclasses.

SpokenDocumentRetrieval:

A textual representation of the audio content from a video can be obtained through automatic speechrecognition.Informationretrievalfromspeechrecognitiontranscriptshasreceivedquiteabitofattentioninrecent yearsinthespokendocumentretrieval trackatTREC7,TREC8andTREC9.

The current 'consensus' from a number of published experiments in this area is that as long as speechrecognitionhasaword errorratebetterthan35% worderror,theninformationretrievalfromthetranscriptsof spoken documents is only 3-10% worse than information retrieval on perfect text transcriptions of thesamedocuments.

ImageSimilarityMatching.

Example-basedimageretrievaltaskhasbeenstudiedformanyyears.Thetaskrequirestheimagesearchenginetofindtheset ofimagesfromagivenimagecollectionthatis similartothegivenqueryimage. Traditionalmethodsforcontent-basedimageretrievalarebasedonavector model.

These methods represent an image as a set of features and the difference between two images is measuredthrougha(usuallyEuclidean)distancebetweentheirfeaturevectors.Whiletherehavebeennolarge-scale,standardizedevaluations of imageretrievalsystems,mostimage retrievalsystems arebasedonfeaturessuchascolor,texture,andshapethatareextractedfromtheimagepixels.

OCRdocumentretrieval:

A different, textual, representation is derived by reading the text that present in the video images usingoptical character recognition (OCR). At TREC 5, experiments have shown that information retrieval ondocumentsrecognizedthroughOCRwitha charactererrorrateof5%and20%degradesIReffectivenessby10%to50%dependingonthemetric, whencomparedtoperfect textretrieval

Incontrast,videoinformationretrieval muchmorecomplexandcombineselementsofspokendocuments,OCRdocuments,imagesimilarityaswellaso theraudioandimagefeatures.

Inthispaperwewillexaminetheeffectsofmultimodalinformationretrievalfromvideodocuments.Thereare only area of audio analysis that we examined was automatic speech recognition. While analyzing thevideo imagery, we considered the color similarity of images, and the presence of faces and text that wasreadableonthescreen.Weexploredthesedimensionsofaudioanalysisandimageanalysisseparatelyandincombinationinourvideoretrievalexperiments.Wewillpresentexperimentswitheachdifferenttypesofextracted metadata performed separately and also combined together in the context of the TREC VideoRetrieval evaluationperformedbytheNationalInstituteofStandardsandTechnology.

MethodsforExtractingTextualMetadata

SpeechRecognition

Theaudioprocessingcomponentofourvideoretrievalsystemsplitstheaudiotrackfromthe MPEG-1encodedvideofile, anddecodestheaudioanddownsamplesitto16kHz, 16bitsamples.

These samples are then passed to a speech recognizer. The speech recognition system we used for theseexperiments is a state-of-the-art large vocabulary, speaker independent speech recognizer [9]. For thepurposesofthisevaluation,a64000-wordlanguagemodelderivedfromalargecorpusofbroadcastnewstranscripts was used. Previous experiments had shown the word error rate on this type of mixeddocumentary-styledatawithfrequent overlapofmusicandspeechtobejust over30%.

Video OCR A different, textual, representation is derived by reading the text that present in the videoimages using optical character recognition (OCR). OCR technology has been commercially available formanyyears.However,readingthetextpresentinthevideostreamrequiresa numberofprocessingstepsinaddition to the actual character recognition. Our video optical character recognition system [5] uses thefollowingapproachtoidentifyandrecognizecaptionedtextthatappearsonthevideo.Giventhenumberofframes contained in typical broadcast news, it is not computationally feasible to process each and everyvideoframefortext.
Forthisreasona roughorquicktextregiondetectionisperformedfirst.Thenthetextmustbeextractedfrom the image, and converted into a binary black and white representation, since the commerciallyavailableOCRenginesdonotrecognizecoloredtextonavariablycoloredbackground.

Unliketextprintedonwhitepaper,thebackgroundoftheimagetendstobecomplex,withthecharacterhueandbrightnessverynearthebackgroundvalues.

InformationSystemEvaluation

The creation of the annual Text Retrieval Evaluation Conference (TREC) sponsored by the DefenseAdvancedResearchProjectsAgency(DARPA)andtheNationalInstituteofStandardsandTechnology(NIST)changedthestandardprocessofevaluatinginformationsystems.

Theconferenceprovidesastandarddatabaseconsistingofgigabytesoftestdata,searchstatementsandtheexpected results from the searches to academic researchers and commercial companies for testing of theirsystems.Thishasplacedastandardbaselineintocomparisonsofalgorithms.

InrecentyearstheevaluationofInformationRetrievalSystemsandtechniquesforindexing, sorting,searchingandretrievinginformationhavebecomeincreasinglyimportant. TherearemanyreasonstoevaluatetheeffectivenessofanInformationRetrievalSystem:

Toaidintheselectionofasystemtoprocure
□□ Tomonitorandevaluatesystemeffectiveness
□□ Toevaluatequerygenerationprocessforimprovements
□□ Toprovideinputstocost-benefitanalysisofaninformationsystem

- Todeterminetheeffectsofchangesmadetoanexistinginformationsystem.

MeasuresUsedinSystemEvaluations

Measurementscanbemadefromtwoperspectives:userperspectiveandsystemperspective.Techniquesforcollectingmeasurementscanalsobeobjectiveorsubjective. Anobjectivemeasureisonethatiswell-definedandbaseduponnumericvaluesderivedfromthesystemoperation. Asubjectivemeasurecanproduceanumber,butisbaseduponanindividual user'sjudgment

Measurementswithautomaticindexingofitemsarrivingatasystemarederivedfrom standardperformancemonitoringassociatedwithanyprogramina computer(e.g.,resourcesusedsuchasmemoryandprocessingcycles)andtimetoprocessanitemfromarrivaltoavailabilitytoa searchprocess.Whenmanual indexingisrequired, themeasuresarethenassociatedwiththeindexingprocess.

Response time is a metric frequently collected to determine the efficiency of the search execution.Responsetimeisdefinedasthetimeittakestoexecutethesearch.Inadditiontoefficiencyofthesearchprocess, thequalityofthesearchresultsarealsomeasuredbyprecisionandrecall.

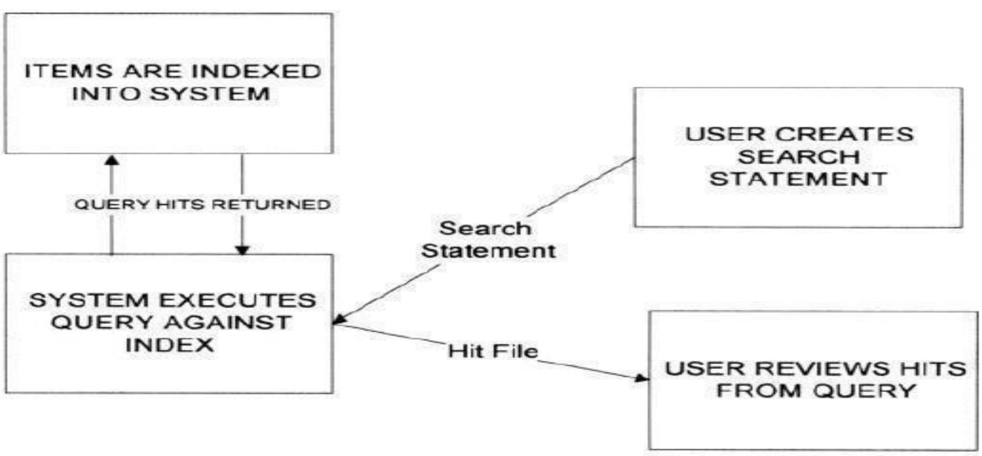


Figure 11.1 Identifying Relevant Items

$$\text{Fallout} = \frac{Number\_Retrieved\_Nonrelevant}{Number\_Total\_Nonrelevant}$$

where *Number_Total_Nonrelevant* is the total number of non-relevant items in the database. Fallout can be viewed as the inverse of recall and will never encounter the situation of 0/0 unless all the items in the database are relevant to the search. It can

Anothermeasurethatisdirectlyrelatedtoretrievingnon-relevantitemscanbeusedindefininghoweffectivean informationsystemisoperating. ThismeasureiscalledFalloutanddefinedas:

There are other measures of search capabilities that have been proposed. A new measure thatprovidesadditionalinsightincomparingsystemsoralgorithmsisthe“UniqueRelevanceRecall”(URR)metric. URRisusedtocomparemoretwoormorealgorithmsorsystems.

Itmeasuresthenumberofrelevantitemsthatareretrievedbyonealgorithmthatarenotretrievedbytheothers:

$$\text{Unique\_Relevance\_Recall} = \frac{Number\_unique\_relevant}{Number\_relevant}$$

Number_unique_relevant is the number of relevant items retrieved that were not retrieved by other algorithms. When many algorithms are being compared, the definition of *uniquely* found items for a particular system can be modified, allowing a small number of other systems to also find the same item and still be considered unique. This is accomplished by defining a percentage **(P_u)** of the total number of systems that can find an item and still consider it unique. *Number_relevant* can take on two different values based upon the objective of the evaluation:

VALUE	INTERPRETATION
Total Number Retrieved Relevant (TNRR)	the total number of relevant items found by all algorithms
Total Unique Relevant Retrieved (TURR)	the total number of unique items found by all the algorithms

A	B	C	D	E	F	G	H	I	J	K	L	M
3	4	2	22	1	100	200	22	100	10	500	6	15

Figure 11.2a Number Relevant Items

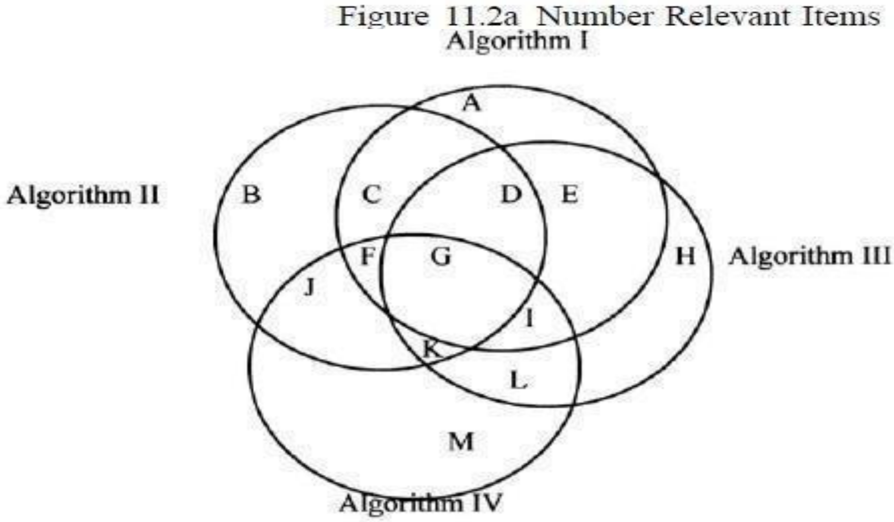


Figure 11.2b Four Algorithms With Overlap of Relevant Retrieved

actually found by the algorithm. Figure 11.2a and 11.2b provide an example of the overlap of relevant items assuming there are four different algorithms. Figure 11.2a gives the number of items in each area of the overlap diagram in Figure 11.2b. If a relevant item is found by only one or two techniques as a "unique item," then from the diagram the following values URR values can be produced:

Algorithm I - 6 unique items (areas A, C, E)
Algorithm II - 16 unique items (areas B, C, J)
Algorithm III - 29 unique items (areas E, H, L)
Algorithm IV - 31 unique items (areas J, L, M)

TNRR = A + B + C + • • • + M = 985
TURR = A + B + C + E + H + J + L + M = 61

Algorithm	URR_{TNRR}	URR_{TURR}
Algorithm I	6/985 = .0061	6/61= .098
Algorithm II	16/985 = .0162	16/61= .262
Algorithm III	29/985 = .0294	29/61 = .475
Algorithm IV	31/985 = .0315	31/61 = .508

Othermeasureshavebeenproposed forjudgingtheresultsofsearches:
<u>NoveltyRatio</u>:ratioofrelevantandnotknowntotheusertototal relevantretrieved<u>CoverageRatio</u>:ratioofrelevantitemsretrievedtototalrelevantbytheuserbeforet hesearch
<u>SoughtRecall</u>:ratioofthetotalrelevantreviewedbytheuserafterthesearchtothetotalrelevanttheuserwo uld havelikedto examine

In some systems, programs filter text streams, software categorizes data or intelligent agents alert users if important items are found. In these systems, the Information Retrieval System makes decisions without any human input and their decisions are binary in nature (an item is acted upon or ignored). These systems are called binary classification systems for which effectiveness measurements are created to determine how algorithms are working (Lewis-95). One measure is the utility measure that can be defined as (Cooper-73):

$$\mathbf{U = \alpha * (Relevant\_Retrieved) + \beta * (Non\text{-}Relevant\_Not\ Retrieved) - \delta * (Non\text{-}Relevant\_Retrieved) - \gamma * (Relevant\_Not\ Retrieved)}$$

where α and β are positive weighting factors the user places on retrieving relevant items and not retrieving non-relevant items while δ and γ are factors associated with the negative weight of not retrieving relevant items or retrieving non-relevant items. This formula can be simplified to account only for retrieved items with β and γ equal to zero (Lewis-96). Another family of effectiveness measures called the E-measure that combines recall and precision into a single score was proposed by Van Rijsbergen (Rijsbergen-79).

Questions:

1. WhatisTextSearch?ExplainTextStreamingArchitecture.
2. WhatisBruteForceApproach?
3. DiscussbrieflyaboutBoyer-MooreAlgorithm?
4. ExplaintheKNUTH-MORRIS-PRATTalgorithm?
5. ExplainvariousTypesofMultimediaData?WhatisSpokenDocumentRetrieval?
6. ExplainaboutVideoRetrieval?