UNIT-II

Cataloging and Indexing: Objectives, Indexing Process, Automatic Indexing, formation Extraction.

Data Structures: Introduction, Stemming Algorithms, Inverted file structures, N-gram data structure, PAT data structure, Signature file structure, Hypertext data structure.

CATALOGING AND INDEXING:

The transformation from received item to searchable data structure is called indexing.

- Process can be manual or automatic.
- Creating a direct search in document data base or in direct search through index files.
- Concept based representation: instead of transforming the input into a searchable format some systems transform the input into different representation that is concept-based Search and return item as per the incoming items.

History of indexing: Shows the dependency of information processing capabilities on manual and then automatic processing systems.

- Indexing originally called cataloguing oldest technique to identity the contents of items to assist in retrieval.
- Items overlap between full item indexing, public and private indexing of files.

Objectives:

The public file indexer needs to consider the information needs of all users of library system. Items overlap between full item indexing, public and private indexing of files.

- Users may use public index files as part of search criteria to increase recall.
- They can constrain the research by private index files
- The primary objective of representing the concepts with in an item to facilitate users finding relevant information.
- Users may use public index files as part of search criteria to increase recall.
- They can constrain there search by private index files.
- The primary objective of representing the concepts with in an item to facilitate users finding relevant information.

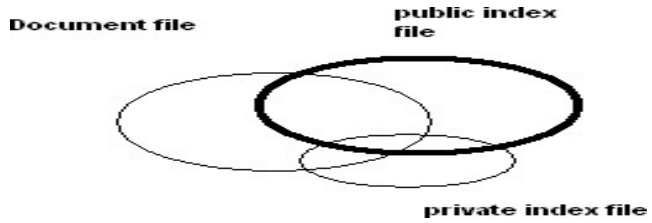
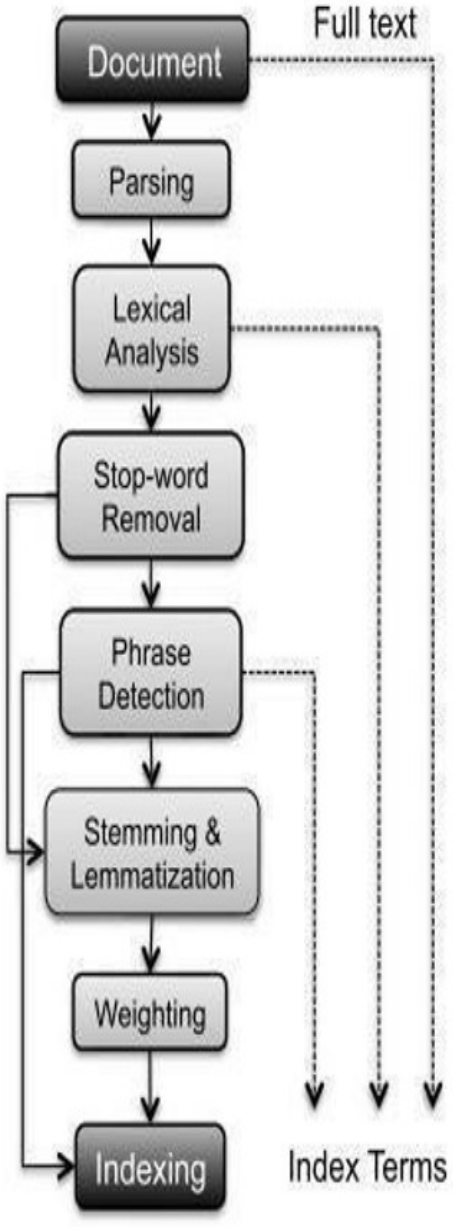


Fig: Indexing process

1. Decide the scope of indexing and the level of detail to be provided. Based on usage scenario of users.
2. Second decision is to link index terms together in a single index for a particular concept.

Fig. 2.3 Text processing phases in an IR system



Document Parsing. Documents come in all sorts of languages, character sets, and formats; often, the same document may contain multiple languages or formats, e.g., a French email with Portuguese PDF attachments. Document parsing deals with the recognition and "breaking down" of the document structure into individual components. In this preprocessing phase, unit documents are created, e.g., emails with attachments are split into one document representing the email and as many documents as there are attachments.

1. **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 on the language at hand: for example, the English language 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 its complexity and the extensive literature in the field for current approaches.

2. **Stop-Word Removal**. A subsequent step optionally applied to the results of lexical 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 by the Snowball stemmer,1 the effect of stop-word removal would be: "search engine most visible information retrieval applications".

3. **Phrase Detection**. This step captures text meaning beyond what is possible with pure bag- of-word approaches, thanks to the identification of noun groups and other phrases. Phrase detection maybe approached in several ways, including rules (e.g., retaining terms that are not separated by punctuation marks), morphological analysis, syntactic analysis, and combinations.

4. **For example**, scanning our example sentence "search engines are the most visible information retrieval applications" for noun phrases would probably result in identifying "search engines" and "information retrieval".

5. **Stemming and Lemmatization**. Following phrase extraction, stemming and lemmatization aim at stripping down word suffixes in order to normalize the word. In particular, stemming is a heuristic process that "chops off" the ends of words in the hope of achieving the goal correctly most of the time; a classic rule-based algorithm for this was devised by Porter [280]. According to the Porter stemmer, our example sentence "Search engines are the most visible information retrieval applications" would result in: "Search engine are the most visible inform retrieval".

Lemmatization is a process that typically uses dictionaries and morphological analysis 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, our sentence would result in "Search engine are the most visible information retrieval application" when lemmatized according to a WordNet-based lemmatizes

Weighting. The final phase of text preprocessing deals with term weighting. As previously mentioned, words in a text have different descriptive power; hence, index terms can be weighted differently to account for their significance with in a document and or a document collection. Such a weighting can be binary, e.g., assigning 0 for term absence and1 for presence.

SCOPE OF INDEXING

- When perform the indexing manually, problems arise from two sources the author and the indexer the author and the indexer.
- Vocabulary domain may be different the author and the indexer.
- This results in different quality levels of indexing.
- The indexer must determine when to stop the indexing process.
- Two factors to decide on level to index the concept in a item.
- The exhaustively and how specific indexing is desired.
- Exhaustively of index is the extent to which the different concepts in the item are indexed.
- For example, if two sentences of a10-page item on microprocessors discussion-board caches, should this concept be indexed Specific relates preciseness of index terms used in indexing.
- For example, whether the term "processor" or the term "microcomputer" or the term "Pentium" should be used in the index of an item is based upon the specificity decision.
- Indexing an item only on the most important concept in it and using general index terms yields low exhaustively and specificity.
- Another decision on indexing is what portion of an item to be indexed Simplest case is to limit the indexing to title and abstract (conceptual) zone.
- General indexing leads to loss of precision and recall.

PRE ORDINATION AND LINKAGES

- Another decision on linkages process whether linkages are available between index terms for an item.
- Used to correlate attributes associated with concepts discussed in an item. This process is called preordination.
- When index terms are not coordinated at index time the coordination occurs at search time. This is called post coordination, implementing by "AND" index terms.
- Factors that must be determined in linkage process are the number of terms that can be related.
- Ex., an item discusses the drilling of oil wells in Mexico by CITGO and the introduction of oil refine rise in Peru by the U.S."

DATA STRUCTURES

- Introduction to Data Structures
- Stemming Algorithms
- Inverted File Structure
- N-Gram Data Structure
- PAT Data Structure
- Signature File Structure
- Hypertext and XML Data Structures

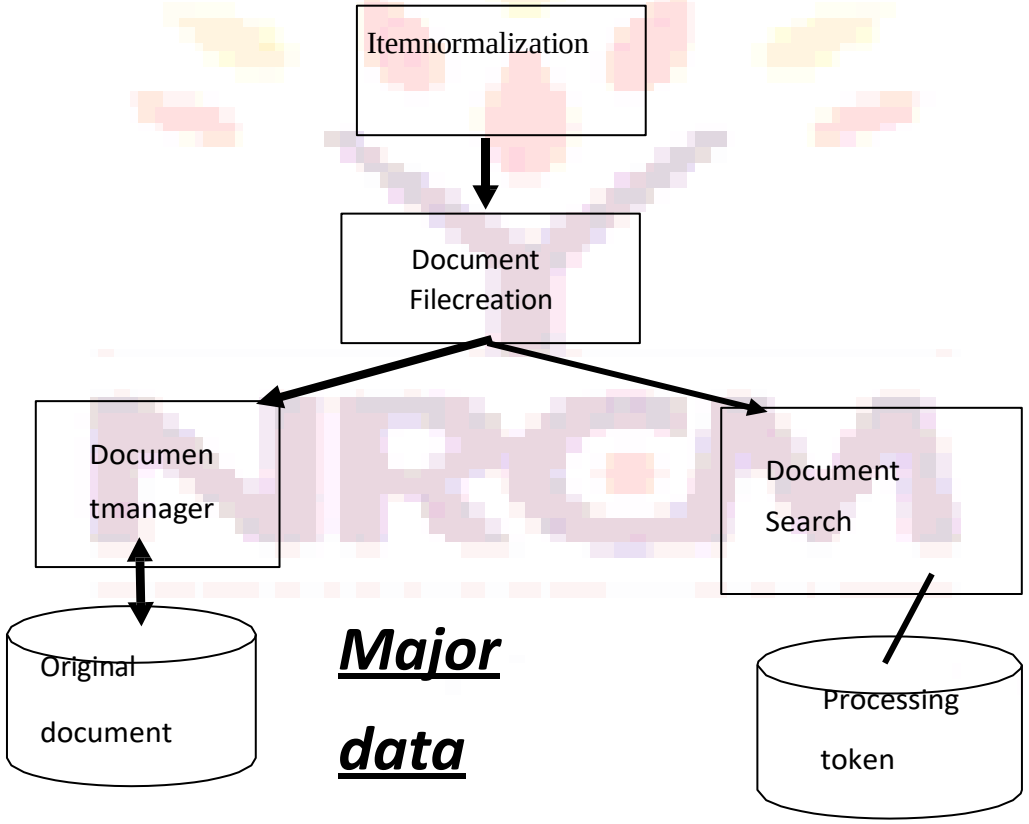
Datastructure:

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)isatermthatisusedtorefermappingmultiplemorpholo gicalvariantstosinglerepresentation(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(Awordformedfromtheinitiallettersofanamesay IARE …)shouldnothavestemmingapplied.

➢ Stemmingcanalsocauseproblems for natural languageprocessing NPLsystemsbycausing lossofinformation.

PORTERSTEMMINGALGORITHM

- Basedonasetconditionofthestem
- AconsonantinawordisaletterotherthanA,E,I,OorU,someimportants temconditionsare

1. Themeasuremofastemisafunctionofsequenceofvowels(V)follow ed 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.
- Theoriginaltermorstemmedversionofthetermislookedupinadictiona ryandreplaced bythestemthatbestrepresentsit.
- ThistechniquehasbeenimplementedinINQUERYandRetrieval waresystems-

INQUERYsystemusesthetechniquecalledKstem.

- Kstemisamorphologicalanalyzerthatconflateswordsvariantstoarootform.
- Itrequiresawordtobeinthedictionary
- Kstemuses6majordatafilestocontrolandlimitthestemmingprocess.

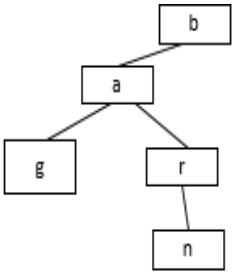
1. Dictionaryofwords(lexicon)
2. Supplementallistofwordsfordictionary
3. Exceptionallistofwordsthatshouldretaina„e‟attheend(e.g.,“suites”to“suite ”but“suited”to“suit”).
4. Direct_conflation-wordpairsthatoverridestemmingalgorithm.
5. County_nationality_conflation(BritishmapstoBritain)
6. Propernouns--thatshouldnotbestemmed

- 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

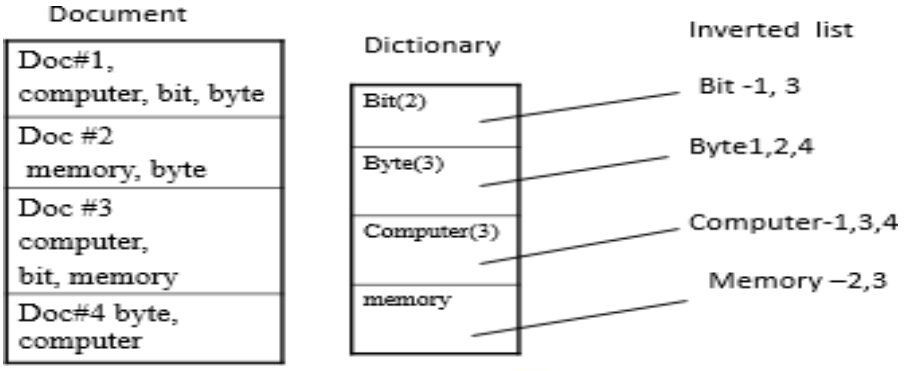
Invertedfilestructures are composedof three

filesThedocumentfile

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

Thisisasortedlist(processingtokens)inthesystemandapointertothelocationofitsinversionlist.

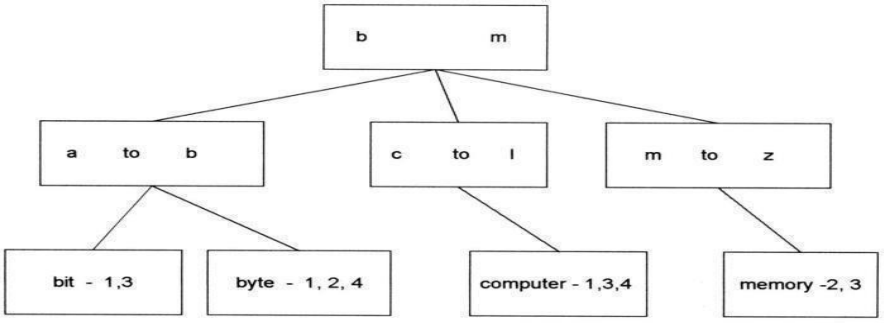
Dictionarycanalsostoreotherinformationusedinqueryoptimizationsuchaslengthofinver sionliststo increasetheprecision.



- 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.,bl
ank,period,semicolon,colon,etc.).

- Thesymbol#isusedtorepresenttheinterwordsymbolwhichisanyoneofase
t 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:

Theyplaceafinitelimitonthenumberofsearchable tokenMaxSeg

n=(□ □

)nmaximumnumberofuniquengramsthatcanbegenerated.“n”ist

helength ofn-grams

□ number of process able symbols Disadvantage:

longerthengramthesizeofinversionlistincrease.Performanceha

s85 %precision.

PATdatastructure(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.

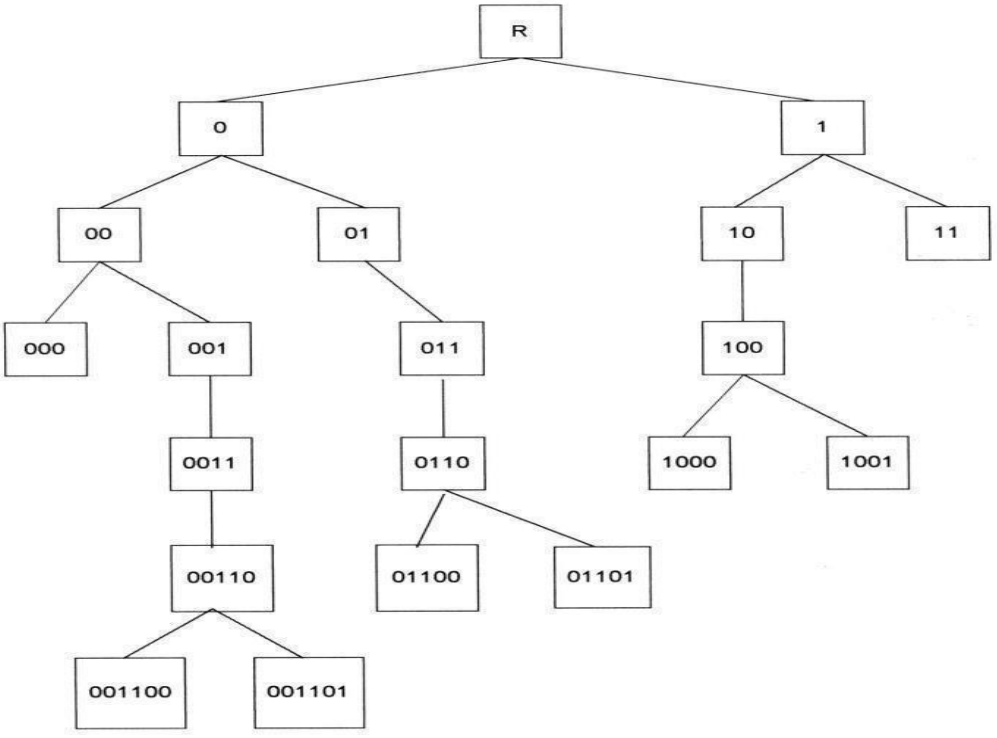
Examplesofsistrings

- 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.

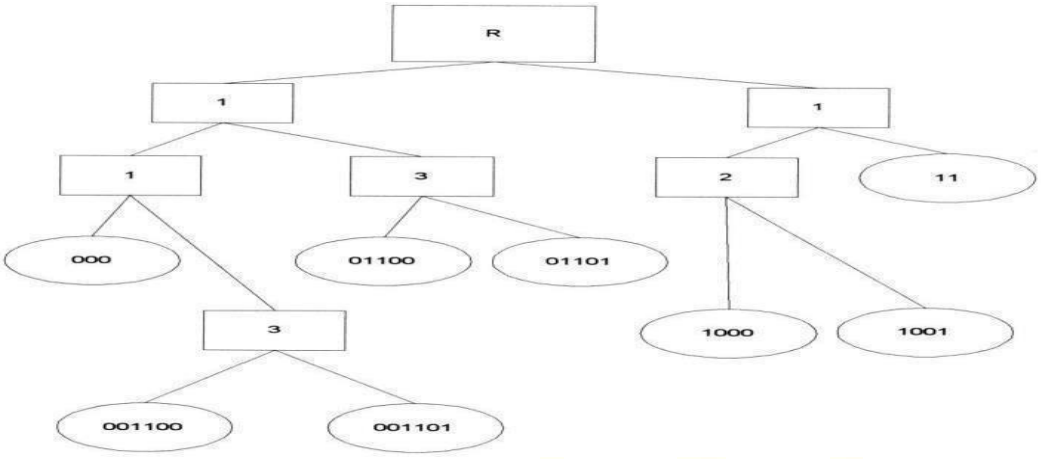
INPUT	100110001101
sistring1	1001....
sistring2	001100...
sistring3	01100....
sistring4	11.....
sistring5	1000...
sistring6	000.....
sistring7	001101
sistring8	01101

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?