UNIT-I

Introduction to Information Retrieval

~ Pratik Patil

Q1] Difference between Data Retrieval & Information Retrieval.

Parameter	Data Retrieval	Information Retrieval
1) Data Type	Structured	Unstructured
2) Query Language	SQL, DB query	Natural language, Keyword
3) Result type	Exact match	Relevant Document
4) Matching logic	Precise matching	Approximate / Relevant based matching
5) Output format	Table of data	Ranked list of document.
6) Used in	RDBMS, Databases	Search engine.
7) Ranking of Result	No	Yes.
8) Query Specification	Complete	Incomplete.
9) Error Response	Sensitive	Insensitive.
10) Model	Deterministic	Probalistic
11) Example	select * from student where id = 1;	Google search → final year project ideas.

Q2 Explain information Retrieval process with the help of block diagram.

→ An information Retrieval system is a software system that helps users find relevant information from a large collection of unstructured data like text documents, web pages, PDFs, etc...

- Any system that is used to make the literature searching may be called as information Retrieval System
- Hence, IR is the study of finding needed information. It helps user to find information that matches their information needs.
- It locates based on user input such as keyword. for example find document containing "Database Management System"

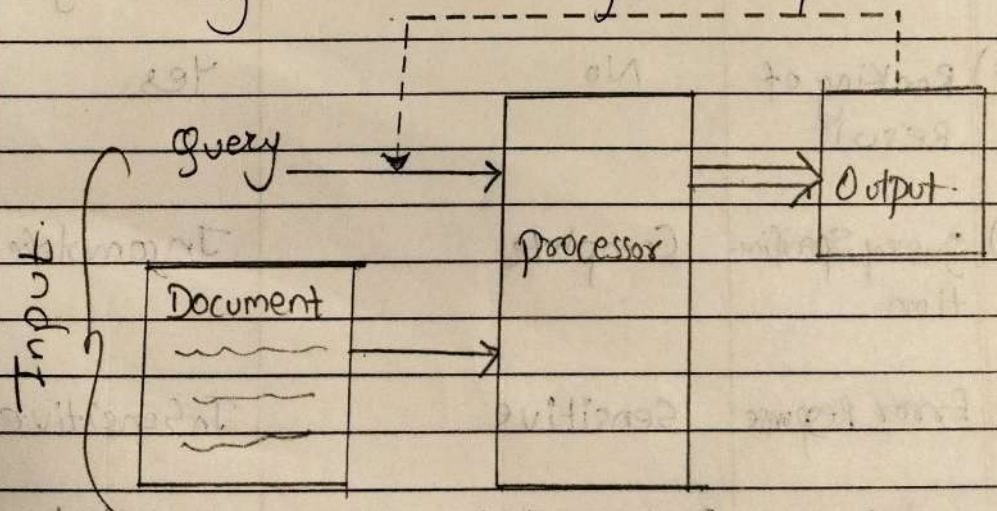


fig → IR block diagram.

a) Input → The process begins when the user enters a query. This query is in natural language. A text of document is lost once it has been processed.

b) A document representation → It is a list of extracted words considered to be significant.

c) Processor: It involve in performing actual retrieval function, executing the search,

d) feedback → Improves the subsequent run after sample retrieval

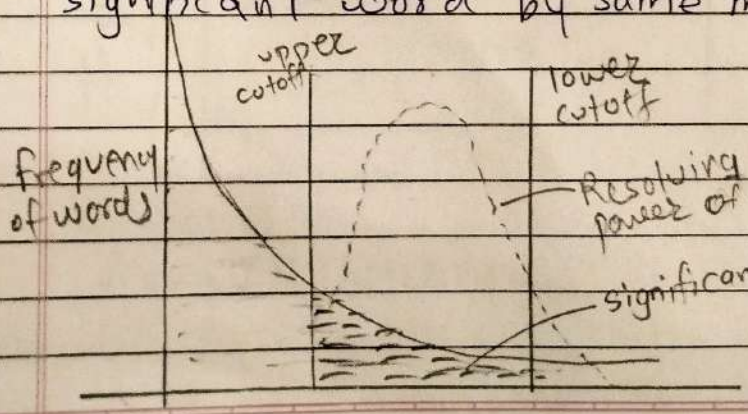
e) output → A set of document number. Like google shows most relevant links at the top.

Page No.
Date

Q3 Explain Luhn's Ideas & method.

→ Luhn's idea was that words with high frequency & low frequency are too common & too rare to contribute significantly to the content of a document & that only words with medium frequency are significant.

- The idea behind this was based on the fact that a writer normally repeats certain words when writing on a subject.
- The more a specific word is found in the document the more significance may be assigned to these words.
- He also took common words into ~~consideration~~ consideration & ~~present~~ preserve such ~~common~~ words together, but significance does not reside under this words.
- The most common words must put together and excluded from beign considered as significant word by same method.



Page No.
Date

Luhn's method for each document

Step 1) Eliminate very common words

Step 2) Apply stemming
eg → differentiation, different, differently → differ.

Step 3 → frequency calculation.
→ Count how often each normalized word appears.

Step 4 Keep only medium-frequency-words.
Remove non-frequent terms.

Page No.
Date

34 Explain Conflation algorithm in detail with its advantages & disadvantages.

→ • Information Retrieval is the matter of deciding which documents in collection should be retrieved to satisfy a user's need for information.

• The retrival decisions are made by comparing the terms of the query with the index terms appearing in the document itself

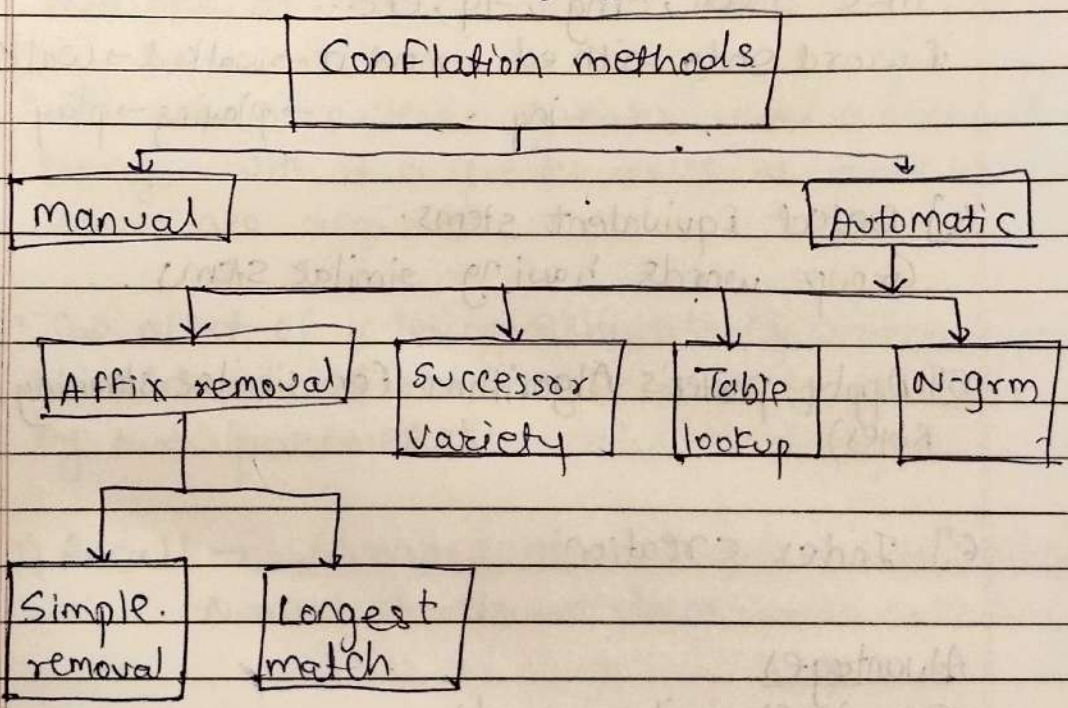
• for example Computational, Computer, Computes, computing etc... are generally the most common, with other source including valid alternative spellings, mis-spellings etc.

• Now, Number of ~~stremming~~ stemming algos. have been developed, which attempt to reduce a word to its steam or root form. stemming is the process for reducing the words to their stem, base or root form. This process of stemming is often called as Conflation. & These programs are called as stemming algorithms or Stemmers.

• word conflation is the process of grouping the words that have the same meaning and are reduced to single term for example: collect, collected, collecting, collection, collections can all.

- The Key terms of a query or document are represented by stems rather than by the original words. It also reduces the dictionary size. Smaller the dictionary size more saving of storage space & processing time.

- Conflation algorithms are classified into two main classes: stemming Algorithms which are languages dependent & string-similarity algorithm which are language independent



Steps in conflation Algorithm

1] Read Input Documents.
- Open & read each file, collecting words for processing

2] Stop word Removal.

- Eliminate high-frequency non-significant words like "the", "is", "of"
- Implements Luhn's upper cut-off technique.
- Reduce file size by 30-50 %.

3] Suffix Stripping (stemming)
- Apply rules to remove common suffixes like -ed, -ing, -ly, etc...

if word ends with ed remove it → walked → walk
ing → → playing → play

4] Detect Equivalent stems.
Group words having similar stems.

5] Apply porter's Algorithm (or similar stemming Rules)

6] Index creation.

Advantages
1] Combines similar words
2) Reduces storage
3) Speeds up searching
4) Improves recall.

Disadvantages:
1] May over-stem
2) No meaning check

3) Language specific.

Q Why is index term weighting used?

→

- An exhastive index is one which lists all possible index terms. Greater exhaustivity gives a higher recall, however this occurs at the expense of precision.
- This mean the user may retrieve larger number of irrelevant documents
- In manual system a greater level of exhaustivity brings with it a greater costs as more man hours are required.
- The effect of indexing exhaustivity and specificity on retrieval effectiveness can be explained by two parameters

1) Recall → Number of relevant document retrieved / Number of relevant docmt. in the collection

2) precision → Number of relevant document retrieved / Total Number of documents retrieved

Page No.
Date

Moreover Higher level of exhaustivity of indexing leads to high recall & low precision & low level of exhaustivity leads to low recall & high precision

- Using single words as index terms generally has good exhaustivity, but poor specificity due to word ambiguity.

- Good index is term that is used when index term is assigned to a collection of documents. then documents are rendered as disimilar as possible & bad term is one which render documents more similar

Q] what are the different measures of association ? Explain any five matching coefficients with suitable examples.

→

- In order to cluster the dataset we need to find degree of association between them. This may be distance measure of measure of similarity or dissimilarity

- Binary relationship between objects is used for classification.

- The relationship between document is described by.

1) Similarity → These values indicates how much two documents or objects are near to each other.

2) Association → Same as similarity but how they are ~~cat~~ characterized by discrete state attributes is considered.

3) Dissimilarity: It ~~is~~ shows that how much far the objects are.

- The measure of similarity is designed to quantify the likeness between objects.

- In ~~IRS~~ IRS, two document will be similar to each other if they have more number of common index terms. If two docs are having less number of common index then obviously they will be semantically far from each other.

Different matching coefficient →

Let's consider vector

$$X = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix} = Y$$

a ⇒ X=1 & Y=1 → 2 d = X=0 & Y=0 = 0

b ⇒ X=1 & Y=0 → 1

c ⇒ X=0 & Y=1 = 1

1] Jaccard coefficient:

measures similarity as the size of the intersection ~~be~~ divided by size of the union of the two sets.

$$J_{(X,Y)} = \frac{|X \cap Y|}{|X \cup Y|}$$

formula (set theory) = $\frac{a}{a+b+c}$

$= \frac{2}{2+1+1} = \frac{2}{4} = 0.5$

2) Dice's coefficient

Gives double weight to intersection. More sensitive to matches. Normalises for length by dividing by total no of non zero entries. we multiply by 2 so that we get a measure that ranges from 0 to 1.0

$$\frac{2|X \cap Y|}{|X| + |Y|}$$

$D = \frac{2a}{2a+b+c}$ $= \frac{2 \times 2}{2 \times 2 + 1 + 1} = \frac{4}{6} = 0.66$

3) Simple matching coefficient (SMC)

measure total agreement, including both 0-0 & 1-1 matches

$$SMC = \frac{a+d}{a+b+c+d}$$

$$SMC = \frac{2+0}{2+1+1+0} = \frac{2}{4} = 0.5$$

4) Overlap Coefficient

$$O(X,Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)}$$

$$O = \frac{a}{\min(a+b, a+c)}$$

$$O = \frac{2}{\min(3,3)} = \frac{2}{3} \approx 0.66$$

Q] Explain Rochhio's Algorithm.

→ • Rochhio's Algorithm was developed on the smart project. It operates in three stages.

1) Initial clustering:

- Select initial cluster centers.
- Assign remaining documents to nearest cluster or a "rag-bag"
- Recompute cluster centers
- Cluster may overlap

2) Refinement:

- Adjust parameters like similarity thresholds to better match desire cluster size

3) Tidying up:

- forcefully assign unclustered documents.
- Reduce overlap & finalize clean clusters.

Q Explain Single pass Algorithm.

→

- Single pass Algorithm is also called as incremental clustering
- Clusters are build in one pass over the data
- Each data point is compared to existing clusters, and added to closest one is if similarity is above a threshold, otherwise a new cluster is created.

~~eg~~ → Steps in single pass algorithm.

Step1 → ~~Obj~~ process the object serially

Step2 → Start with empty/zero (0) cluster

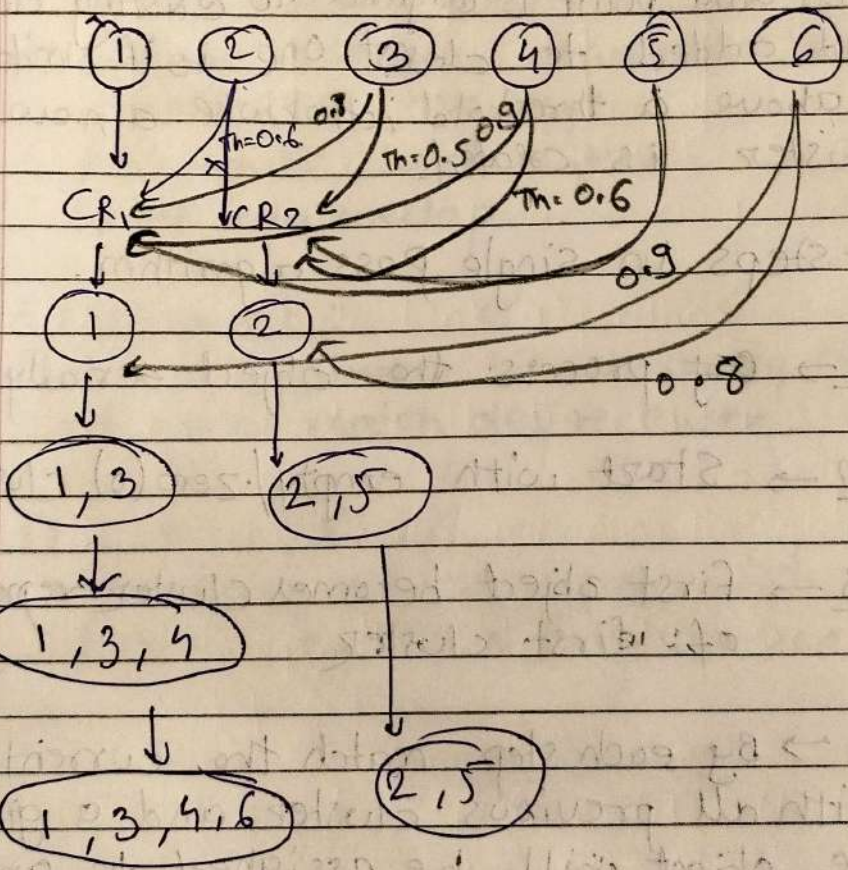
Step3 → First object becomes cluster representative of first cluster

Step4 → By each step match the current data with all previous cluster and a given the object will be assigned to one cluster according to some condition on it.

Step5 → when an new object is assigned to that cluster the representative of that cluster (CR) is recomputed.

Step 6 → If ~~cluster~~ object does not ~~at~~ match previous clusters acc. to threshold value then make new cluster.

eg set = {1, 2, 3, 4, 5, 6}



assume Threshold = 0.8.
overlap → No.
No of cluster → 2.

Page No.
Date

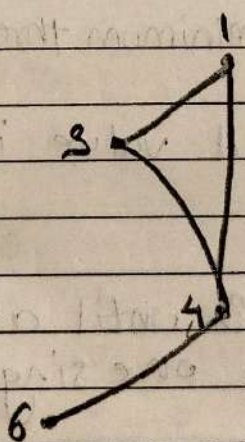
Advantages

1) Very fast
2) Efficient
3) Easy to implement

Disadvantages

1) Result depend on order of data.
2) No Quality Cluster

Scenario → Imagine you arranging the books in library as they come. You pick a book and assign it to an existing shelf if it matches the topic. If not, you create new shelf. You don't go back to reshuffle.



Q → Explain Single Link Algorithm.

- Single link Algorithm is also called as Nearest neighbour clustering.
- It is hierarchical clustering technique basically numerical levels called as dendogram.
- Begins with each data points as its own cluster, then repeatedly merges the two closests clusters based on minimum distance between any pair

Steps in Single Link algorithm:

Step1 → Start with no cluster

Step2 → Start with minimum threshold i.e. 0.1

Step3 → Keep threshold value increasing by 0.1

Step4 → Repeat step 3 until all the objects are in one single cluster

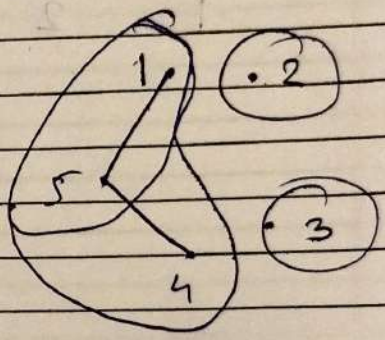
Step5 → Show hierarchy level according to iterator steps.

1					
2	.4				
3	.4	.2			
4	.3	.3	.3		
5	.1	.4	.4	.1	
	1	2	3	4	5

← → Dis-similarity matrix

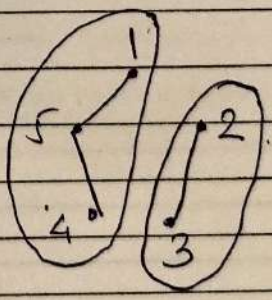
Step1 → Threshold value = 0.1 Th

1				
2	0			
3	0	0		
4	0	0	0	
5	1	0	0	1
	1	2	3	4 5



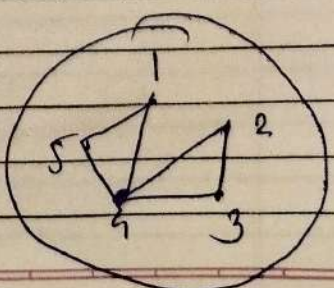
Step2 = 0.2 Th.

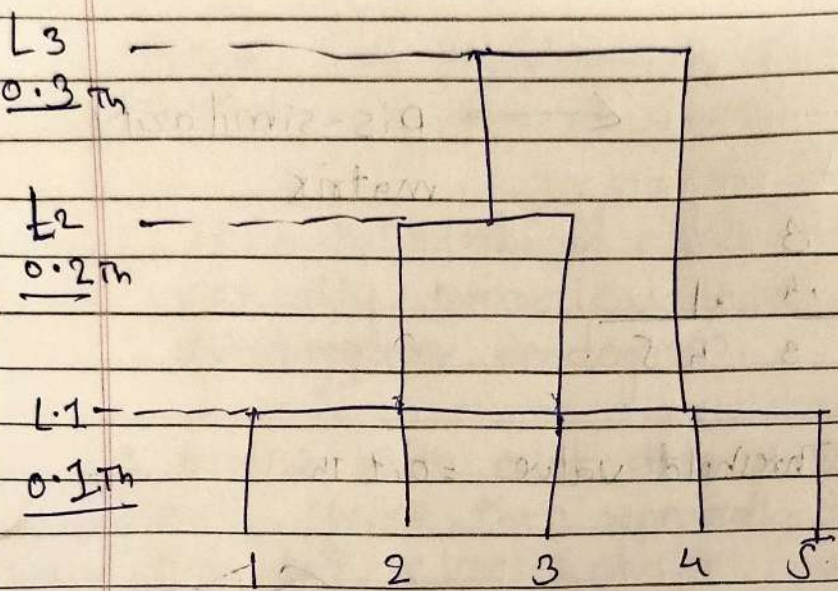
1				
2	0			
3	0	1		
4	0	0	0	
5	1	0	0	1
	1	2	3	4 5



Step3 = 0.3

1				
2	0			
3	0	1		
4	1	1	1	
5	1	0	0	1
	1	2	3	4 5




L3
0.3 m
L2
0.2 m
L1
0.1 m
1
2
3
4
5

UNIT-2

Indexing & Searching Techniques.

Q Explain concept of inverted index file. How can it be used in information retrieval.

→

- Each document is assigned with a list of keywords or attributes. Each keyword is associated with operational relevance weights.
- The inverted file is the sorted list of keyword with each keyword having links to the document containing that keyword.
- Inverted file is composed of two element
  1) Vocabulary
  2) Occurrences
- Vocabulary is set of all different words in the text. For each such word a list of all the text positions where word appears is stored. The set of all those lists is called the occurrences.
- Every entry in the vocabulary has the word, a pointer into the postings structure and word metadata.

1 5 11 13 19 30 39
I want to bake something with chocolate.
It contains milk & sugar
55 58 68 74 79

Vocabulary	Occurrences.
Bake	13
Chochlate.	39
milk	68
sugar	79

- The space required for the vocabulary is smaller. it is $O(nB)$ & The occurences require much more space. i.e. $O(n)$

- Block addressing is required in order to reduce space requirement. The text is divided into blocks and occurences point to the blocks where the word appears. The classical indices which point to the exact occurences are called full inverted indices.

B-1 B-2

I wont to bake	something with chocolate
I contains milk and sugar	

Block-3

Vocabulary	Occurences
Bake	1
Chocolate.	2
Milk	3
Sugar	3

Searching algo on inverted index.

1) Vocabulary search :- The keywords from user query are searched for in the vocabulary

2) Retrieval of occurrences :- The list of occurrences of all the words found is retrieved.

3) Manipulation of occurrences → These lists are then processed to handle operations like Boolean searches (AND, OR, NOT), phrase queries or proximity searches.

Page No.
Date

Q Suffix Tree-
Explain Concept of Suffix tree in info. retrival.

→

- In 1973, Weiner introduced the concept of suffix tree. In 1990 Gene & Udi proposed suffix array.

- Suffix tree is special data structure that helps you find information quickly in a text.

- Suffix tree takes all the possible endings of that text and organizes them in a compressed way, like a tree with branches.

It has following properties

1) The tree has n leaves, labelled 1---n, one corresponding to each suffix of s.

2) Each internal node have atleast 2 children

3) Each ~~label~~ edge in the tree is labelled with a substring of s

4) If you follow path from very top of the tree to all way down to a leaf, you get one of the full suffixes of your original text.

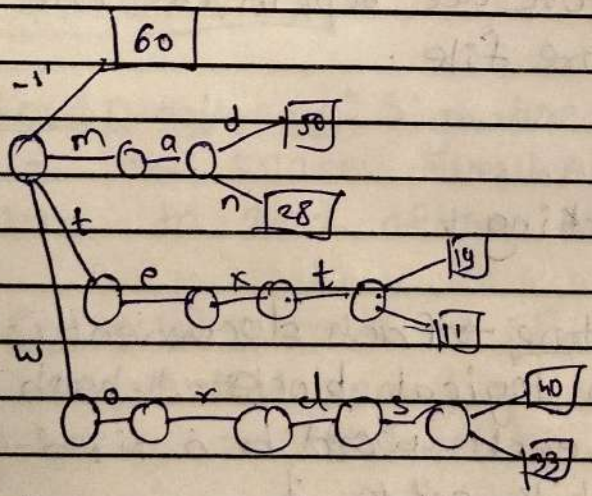
5) The labels of the edges connecting a node with children starts with different character.

- The main problem with the suffix tree is their high space requirement. Depending upon implementation, each node can take up a significant amount of memory.

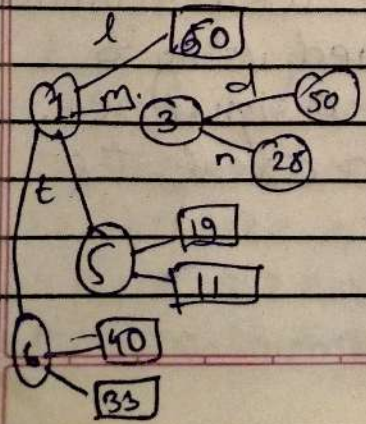
- Due to this reason suffix arrays are often used as a more space-efficient alternative that provides similar functionality.

1 6 9 11 17 19 24 28 33
This is a text. A text has many words.
40 46 50 55 60
words are made from letters.

suffix trie



suffix tree.



• Suffix array is simply an array containing all the pointers to the text suffixes listed in lexicographic order

60	50	28	19	11	40	33

Q] Explain the concepts of signature files in the information retrieval.

→

• Sequent Signature file is a word-oriented index-structure based on hashing, where a "signature" is created as an abstract representation of a document. All the signature are kept in the file called signature file.

• ~~Characteristics~~.
• Structure & working.

1) Word Signature → Each document is divided into logical blocks. A hash function maps each word to a fixed-length with a bits set to 1.

2) Block signature → The word signature for a block are combined using a bitwise or to form block signature. These are concatenated to create the document signature.

3) Searching → When user queries a word. a signature is generated for it & The system then matches it with each block.

example →

Word	Signature.			
free	001	000	110	010
text	000	010	101	001
block signature	001	010	111	011

Characteristics →

1) Low Overhead: Signature files typically have a low space overhead, requiring only 10% to 20% of the text size.

2) Sequential search → This method requires a sequential search over the index, which can be inefficient for very large ~~texts~~ texts

3) False Drop → It occurs when a document's signature matches query's signature but the query word does not match with word in the document.
It depends on
1) The size of signature
2) The size of bits set to 1
3) The number of words per-block

Q Explain exhaustituity & specificity with respect to Index term weighting

→

Exhaustivity

1) Exhaustivity means how much content is covered by a term/index

2) If a term talks about many topics in the document, it is called highly exhaustive.

3) Higher exhaustivity brings higher recall but results in low precision and low exhaustivity brings low recall but high precision.

4) It helps in getting more documents that may be relevent.

5) Useful when user wants more & broader information.

Specificity

1) specificity means how focused & narrow a term is for a document.

2) A specific term appears in fewer documents, but those are highly relevant.

3) High specificity means getting more accurate results. (High precision)

4) But too much specificity may miss some documents (Recall is low).

Q Explain the different kinds of search strategies.

→ Different kind of search strategies / Techniques are →

1) Boolean Search.
- Boolean logic have three operators: AND, OR, NOT.
- Boolean search strategy retrieves those documents which are 'true' for the query. it usually uses OR, AND & NOT logic.

1] AND → Retrieve document if both the condition i.e before and after condition is met.

2] OR → Retrieve document if any one condition or both conditions are met

eg
i] pages with whales
ii] Pages with iceland
iii] Pages with whales & iceland } Retrieves all three types.

3] Not → Only retrieves documen that is not in the given condition. using Not.
Eg → "pages not with whale."
only retrieve pages that does not contain whale.

example

$K_1 \rightarrow D_1, D_2, D_3, D_4$

$K_2 \rightarrow D_1, D_2$

$K_3 \rightarrow D_1, D_2, D_3$

$K_4 \rightarrow D_1$

$(K_1 \overset{AND}{\cancel{or}} K_2)$ OR $(K_3$ AND (Not $K_4))$

$K_1 \overset{AND}{\cancel{\&}\cancel{or}} K_2 \rightarrow D_1, D_2$

OR K_3 Not $K_4 \rightarrow D_2, D_3$

$\rightarrow D_1, D_2, D_3$

Advantages:

1] Simple to understand

2] easy to implement.

Disadvantage:

1) Too much returned, or very little returned

2) Difficult to specify need

2) Serial search

In serial search, a system compares the query to every single document in the collection. It does this by calculating matching function. Documents are then retrieved in two ways

1) By Threshold: All documents that have a matching function value above threshold are retrieved.

By Ranking → The documents are ranked for most to least relevant based on their matching function values, and a cutoff point is used to retrieve a top-ranked of documents.

3) Cluster-Based Retrival.

→ Clustering is the process of grouping the similar things together. In this process the cluster hypothesis is used

Cluster hypothesis → ~~It is the process of~~ Documents in the same cluster behave similarly with respect to relevance to information needs.

Working

- The search starts from root node 0. It then calculates the matching function at the nodes ~~from~~ immediately next from node 0. i.e node 1 & 2

- In this search there are two prerequisits
1) Decision rule → how to move down. Here if max value of current set is greator than previous then continue.

2) Stop point → If max valve of corrent node is less than previous then forcely retrieved.

Q] List & explain the types of queries

→

1) Keyword Based Querying →
- Users enter one or more keywords to express their information need.
- The IR system retrieves documents containing at least one of the query keywords.
- Results are ranked by similarity or relevance.

2) Word Queries →
- This are simple, direct queries based on single word.
- The IR system splits documents into individual words and matches them to the query.
- These are effective for exact word matches but may miss context or variations.

3) Context-based Query
- search for group of words that appears together or close to each other in a document.
- If words appear together signal higher likelihood of relevance than if they appear apart.

4) Boolean Queries:
- Uses logical operators such as AND, OR or NOT to combine keywords.
- used in advanced search system or DB.

Database.

eg → AI & Robotics not automation · will exclude automation.

Q Explain various IR model in detail with their advantages & disadvantages.

→

1] Boolean model

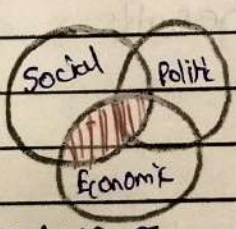
→ It uses Boolean logic to retrieve documents.

• A document is either relevant or not relevant (binary decision).

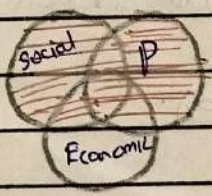
a) AND (∧) : The intersection of two sets

b) OR (∨) : The union of two set.

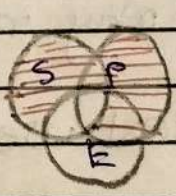
c) NOT (¬) : Set inverse.



S AND E



S ∨ P



(S ∨ P) ¬(S ∧ E)

Advantages

1) simple to implement.

2) Exact matching.

Disadvantage.

1) Not ranked.

2) Proper Knowledge.

Best use → User need exact matching

2) Vector model.

- Represent document & queries as vector in a multi-dimensional space
- measures similarity using cosine or other distance metrics.

Advantages

1) Documents are ranked
2) Supports partial matching.

Disadvantages

1) Complex.
2) Ignore dependencies.
3) May retrieve irrelevant results.

Best used when:

- Relevance ranking is important.
- In web search engine.

3] Probabilistic Model

- The probabilistic IR model tries to predict the probability that a document is relevant to a given user query.
- It uses mathematics & probability to rank documents based on how likely they are similar.

- Basic idea behind this.

- When you type query, the system assumes some documents in the collection are relevant & some are not.

- It calculates the probability of each document being relevant.

- Then it ranks the docs from most likely relevant to least likely.

How it works

1. Assume that we dont know which document are relevant yet.

2. For each document the system estimates probability that document is relevant to query → P(R|D) formula.

3. Rank according to relevance.

4) Feedback loop → If user marked relevant or not then model learn & update future.

Advantage.
1) Ranking is based on relevance probability.
2) Improve over time.
3) Flexible.

Disadvantage

1) Initial probability is hard to guess.
2) Requires training.
3) Complex.