Information Retrieval
SUB: 19CS523
SL.NO.:
Reg: 212220040009
Name: Anitha.R
Dept & year: CSE & IV yr.

6.a) Vector space model & its advantages over Boolean model
6.b) Diff b/w Data Retrieval & Info. retrieval

7.a) Inverted Index
7.b) Privacy Issues

8.a) Pre-processing techniques with examples
8.b) Web spamming & detecting techniques

6.b) Data Retrieval & Information Retrieval

Information Retrieval:

Information Retrieval is the process of retrieving information that is exact to or relevant to the query written by the user and extracting the relevant information from the documents.

This information retrieval has an huge impact in the field of web searching, web spamming and many more techniques.

Why Information Retrieval?:

Suppose if we wanted to retrieve a data from a document, the user would provide relevant queries to extract information from the document

the user query where in the data retrieval only the exact specified terms are extracted / retrieved based on the user query.

So, in this case Information Retrieval plays a vital role in extracting information.

Some models of IR system include,

=> The Boolean Model
=> Vector Space model
=> Language Model
=> Probabilistic Model

Data Retrieval:

Data Retrieval is the explicit process of retrieving only the exact terms specified in the user query.

For example,

If a user writes a query to extract words like humor & happiness, the Data Retrieval Mechanism would only retrieves data that exactly matches the user query. Whereas in Information Retrieval, the relevant terms like synonyms in thesaurus section would also be retrieved in the case of Information Retrieval.

>> Query: "happy", "humor"
>> Document_content: All happy and cheerful people tend to possess great humor sense
>> Information Retrieval:>> happy, cheerful, humor

Information Retrieval	Data Retrieval
i) Information Retrieval tend to retrieve all the relevant information from the user query	Data Retrieval retrieves/ extracts only the exact information from the user query.
ii) This explicity doesn't cause any implications	This might cause certain implications
iii) This method cannot be used in Database system.	This method can be used in Database system
iv) The main goal is to retrieve relevant information.	The main goal is to retrieve exact data based on user query
v) This approach would end up as too few or too many responses.	This approach would end up as exact matches that are found from user query.
vi) Relevant information can be found from the thesauri like synonyms of the given terms.	only exact matches of data are extracted not the relevant information.

Information Retrieval	Data Retrieval
=> Many IR models are available for retrieving the information.	Limited Data Retrieval Methods are available.
=> The success rate is comparatively larger	The success rate is comparatively limited to Data information retrieval.
=> Retrieval process is slower when compared to Data Retrieval.	Retrieval speed is higher as only the matching terms of the user query are extracted.
=> The results of Information retrieval are larger in quantity	Precise quantity of results in Data Retrieval.
=> Thesauri terms like synonyms and more similar terms are extracted in the results	Thesauri results are not included in Data Retrieval.
=> Data base managing users could'nt efficiently use this approach	Data base managing users could use this data retrieval approach.

Inverted Index

7. b) Privacy Issues:

In this world of Web, people tend to use many websites and surf internet for many purposes. We also tend to upload many private information like phone number, address in many forms and in social media. This would probably end up in privacy issues. These privacy web search issues would end in opening up many threats related to security to the users.

Some privacy issues include with the preserving measures:

⇒ Virus/ Malware attacks:

Intruders would tend to infect the user's pc with virus and Malwares like Trojan, Ransomware to the personal computers and would try to hack the personal information of the users.

They might also wanted to know the sensitive information like passwords inorder to intervent the system and hijack it.

This would be so crucial if it is being attempted in a large organization with elevated economic status.

⇒ Web spamming:

Web spamming is the process of intruding unwanted information to the relevant information and hiding the relevant searches from the query by intervening with the page rank. This page rank would interfere with the quality of the search results as well.

⇒ Identity theft:

These attacks include duplicating one's identity by the theft of personal information like password, username and acting upon one's role in platforms like social media and many more.

⇒ Antivirus installation:

Antivirus installation would hinder most of the privacy attacks being from arising. As these special softwares that are specifically designed to identify viruses and malwares that causes serious security impacts. By using these, 90-t of the malwares are detected and dipeted from the system.

⇒ Trojan Attacks:

Trojan Attacks include continuous attempts made to hijack systems by providing fake information and extracting

8. a) Pre-processing Techniques:

Pre-processing is the process of removing unnecessary noice from the data before providing the data as input for data analysis or any other manipulation techniques.

The pre-processing Techniques in Information Retrieval

=> Tokenization
=> Stop-word Elimination
=> Stemming
=> Index selection
=> Text searching Thesauri

Tokenization:

Tokenization is a pre-processing technique that extracts the exact tokens from the document. This would involve the retrieval of specific data from the user query.

Eg: Document: "The quick brown fox jumps over the lazy dog".

Query: "quick fox"

So, this query would lead to the extraction of the

Here in this technique, the exact words, "quick" and "fox" are retrieved from the document.

Stop-word Elimination:

In the stopword elimination the common stopwords like prepositions, adjectives, pronouns, adverbs and kind of parts of speech are eliminated to retrieve the relevant information from the document. Articles are also removed in this approach.

Document: The quick brown fox jumps over the lazy dog.

After stopword-elimination: brown fox jumps over dog

Here, the terms, "the", "quick", "lazy" are eliminated and the required key terms alone are generated.

"the" – article
"quick" – adjective
"lazy" – adjective

Here after removing articles and adjectives, the key terms are retrieved so under which the noise in the data is cleared from the document.

Stemming:

Stemming is the process of removing unwanted parts of a word. We would have different forms of words for a single word, it could be adjective, adverb or noun. So the suffix of the word or the prefix of the word are eliminated and only the root word is extracted in the process of Stemming.

Document: "He attempted the task much quicker and faster than the one who previously attempted it"

After Stemming:

After the process of stemming, the root words extracted are,
"He", "attempt", "the", "task", "muck", "quick", "ans", "fast", "than", "one", "who", "previous", "attempt", "it". So, in this approach the unwanted noise in data is removed.

Index Selection:

Index selection is the process of selecting the most relevant and important keywords from the sections of the document. These terms are selected as index terms of the document such that from these terms, the important

Manual Indexing:

Manual Indexing deals with the selection of the index terms by manual intervention of humans.

Automatic Indexing:

This Automatic Indexing would involve, selecting / retrieving index terms by machines automatically without human intervention.

This would involve web searching and web indexing easier and calculating separate page ranks of the webpages.

Thesauri:

Thesauri method of pre-processing involves retrieval of the relevant data like synonyms of words from the document. This would result in the retrieval of relevant data from the user query.

For example,

For retrieving the document data "He sat down in a place where he felt calm and peaceful".

By using Thesauri, the synonyms "calm", "peace" both are retrieved as they are relevant to each other leading the retrieval of relevant information.