Test 04: Information Retrieval

Submitted By – Srijana Raut (101134199)

1. How information retrieval is different from database managements system? Explain with example. Explain with examples.

   A database management system (DBMS) is a software application that empowers users to store, edit, and retrieve data from databases. This software utilizes a universally accepted approach to categorize, search for, and execute queries on data. Additionally, it takes care of incoming data, arranges it systematically, and offers means for other users and programs to make alterations and retrieve the data.
   Examples of DBMS are MySQL, Microsoft Access, SQL Server, File Maker, Oracle, and Clipper.

   Information Retrieval (IR) refers to a software system designed to manage, store and assess information within document collections, with a primary focus on textual content. It involves the process of seeking and obtaining unstructured material, typically in the form of text, to meet specific information requirements from extensive digital repositories. **For example – Info retrieval can be when user enter a query into a system.**
   Some of the key highlighted difference between Info Retrieval and database management system are.

Info Retrieval	DBMS
Data are free text which can be unstructured	DBMS are structured
Not semantic structure i.e., not tabular format	Semantic structure i.e., tabular format
Queries can be keyword, NLP	Keywords can be sql, relational algebra
It focuses on finding relevant information from a vast repository of unstructured data.	DBMS is used for storing, managing, and retrieving structured data
Application: Search Engine **Example: Google Search**	Application: Online Retail Store **Example: Amazon's Database System**

2. Discuss a Boolean retrieval model. Discuss the limitations of term-incidence document matrix. Why is indexing chosen?
[Hint: Consider term-document incidence matrix to highlight the importance of inverted indexing.]
Draw both.
a) term document incidence matrix, and
b) inverted index that would be built
for the following document collection:
• Doc 1: new home sales top forecasts

• Doc 2: home sales rise in July
• Doc 3: increase in home sales in July
• Doc 4: July new home sales rise

Boolean Retrieval model is simply an information retrieval task based on predefined Boolean logical operator. The Boolean retrieval model is a model for information retrieval in which we can pose any query which is in the form of a Boolean expression of terms, i.e., in which terms are combined with the operators AND, OR, and NOT. The model views each document as just a set of words.
The limitation of term-incidence document matrix is.

i. Only find the keyword is present or not. It doesn't check how many time the word is repeated.
ii. Location of words is not finding.
iii. Doesn't work for large repository.
iv. Doesn't maximize the information.

To resolve the above limitations, the inverted indexing is chosen over term -incidence document matrix.

The tdim and inverted index of the above docs are given below.

Terms	**Inverted Index**	**Term document incidence matrix (tdim)**
New	1->4	1001
Home	1->2->3->4	1111
Sales	1->2->3->4	1111
Top	1	1000
Forecasts	1	1000
Rise	2->4	0101
In	2->3	0110
July	2->3->4	0111
Increase	3	0010

3. For the same work (see above), what are the returned results for queries:
a) home AND July, and
b) top OR sales

home = 1111
July = 0111
Thus, Home and July = 0111

As a result, Document 2->3->4

Similarly, Top = 1000
Sales = 1111

Top Or Sales = 0000

Thus, No Document

4. For the same work (see above), how can you optimize the following query (provide capital O notation along the process):
home AND sales and July

Before optimizing, home AND Sales AND July = 0111
Ans – d2, d3, d4

After optimizing, according to the posting length.
In general, query optimization can be done in the following:
• Intersection the two smallest posting lists.
This way, the intermediate results will be no longer than the smallest of the two (i.e.,min(list1, list2)).
• The intermediate results will be intersected (processed) with the remaining one.

Length of July < Length of home < length of sales (Because posting length matters)
Ans - d2, d3, d4
Big O =| N |+ |M| where, N and M are respective length.