



INFORMATION RETRIEVAL

By

Rohini Naik

OUTLINE

- Introduction
- Information versus Data Retrieval
- IR System Block Diagram
- Major challenges in IR
- Boolean Retrieval

Introduction

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

- Unstructured /semi structured data
- covers supporting users in browsing or filtering document collections or further processing a set of retrieved documents.

Introduction

Information retrieval (IR) is subfield of computer science that deals with automated retrieval of information (especially text) based on their content and context.

Calvin Moores (1950) : It is concerned with the representation, storage, and organization and accessing of information items

Introduction

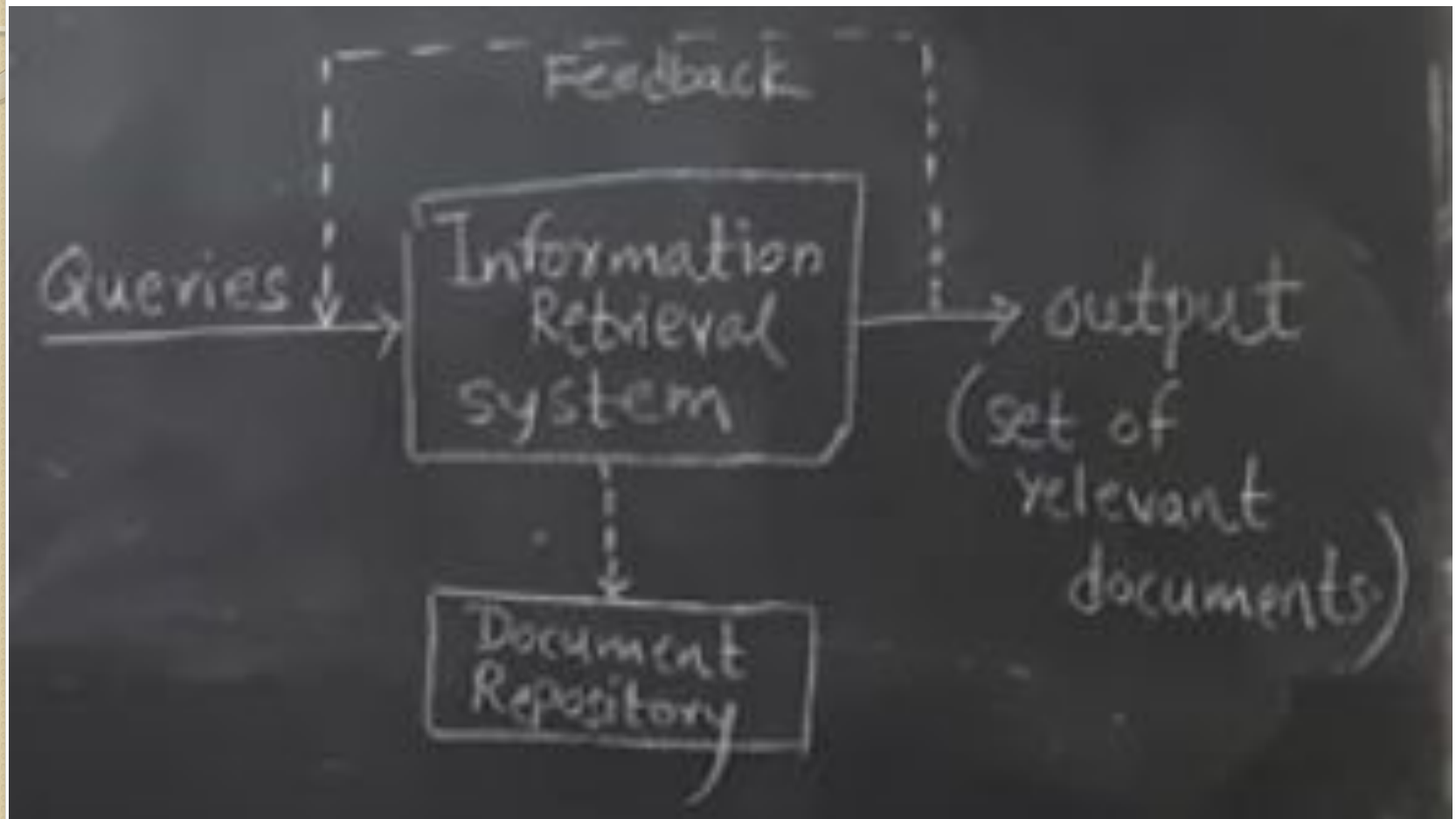
- Given a set of documents, clustering is the task of coming up with a good grouping of the documents based on their contents
- Given a set of topics, standing information needs, or other categories (such as suitability of texts for different age groups), classification is the task of deciding which class(es), if any, each of a set of documents belongs to.

Information versus Data Retrieval

	Information Retrieval	Data Retrieval
Data	Free text, unstructured	Database tables, structured
Queries	Keywords, Natural language	SQL, Relational algebras
Results	Approximate matches	Exact matches
Results	Ordered by relevance	Unordered
Accessibility	Non-expert humans	Knowledgeable users or automatic processes

Information Retrieval	Data Retrieval
The software program that deals with the organization, storage, retrieval, and evaluation of information from document repositories particularly textual information.	Data retrieval deals with obtaining data from a database management system such as ODBMS. It is A process of identifying and retrieving the data from the database, based on the query provided by user or application.
Retrieves information about a subject.	Determines the keywords in the user query and retrieves the data.
Small errors are likely to go unnoticed.	A single error object means total failure.
Not always well structured and is semantically ambiguous.	Has a well-defined structure and semantics.
Does not provide a solution to the user of the database system.	Provides solutions to the user of the database system.
The results obtained are approximate matches.	The results obtained are exact matches.
Results are ordered by relevance.	Results are unordered by relevance.
It is a probabilistic model.	It is a deterministic model.

IR System Block Diagram





Major challenges in IR

- **Information Overload:** The exponential growth of digital content on the internet and other sources has led to information overload. Users are often overwhelmed by the sheer volume of search results, making it difficult to find the most relevant information quickly.
- **Relevance and Precision:** Ensuring that the retrieved documents are relevant to the user's query is a critical challenge. IR systems need to strike a balance between providing comprehensive results and avoiding irrelevant or low-quality documents.
- **Query Ambiguity:** Queries from users can often be ambiguous or imprecise, making it challenging for IR systems to accurately understand the user's intent. This problem is particularly acute when dealing with natural language queries.

Major challenges in IR

- User Intent Understanding: Understanding the user's underlying intent behind the query is crucial for providing relevant results. However, accurately capturing user intent remains challenging, especially for complex or long-tail queries.
- Multilingual Information Retrieval: As the internet connects people from different linguistic backgrounds, IR systems must handle queries and content in multiple languages. Translating queries and retrieving relevant documents across languages presents unique challenges.
- Scalability and Efficiency: With the exponential growth of data, IR systems need to handle large-scale collections efficiently. Real-time retrieval and quick response times are essential to meet user expectations.

Major challenges in IR

- Personalization: Users have become accustomed to personalized experiences on the web. IR systems need to consider user preferences, behavior, and context to provide personalized search results and recommendations.
- Diversity in Information Sources: Modern IR systems must retrieve information from various sources, such as web pages, social media, images, videos, and more. Integrating and ranking information from diverse sources pose additional challenges.
- Trust and Credibility: With the abundance of misinformation and fake news on the internet, users seek reliable and credible information. Ensuring the trustworthiness of retrieved content is a growing concern.

Major challenges in IR

- Semantic Gap: The semantic gap refers to the mismatch between the user's query and the representation of content in the documents. Bridging this gap to better match user intent is a continuing challenge.
- Dynamic Content: Content on the web is continuously changing and evolving. Keeping IR systems up-to-date with the latest information poses challenges in maintaining freshness and accuracy.
- Long-Tail Queries: Handling infrequent or specific queries, known as long-tail queries, is challenging because they may have limited data for learning relevant patterns.
- Evaluation and Metrics: Developing appropriate evaluation metrics that reflect the real-world utility of IR systems is an ongoing challenge. Existing metrics, such as relevance, may not fully capture user satisfaction or the quality of the retrieved information.

Boolean Retrieval

- 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
- That is, in which terms are combined with the operators AND, OR, and NOT.
- The model views each document as just a set of words.
- Exact match retrieval

Boolean Retrieval

- Basic Assumption of Boolean Model
- An index term is either present(1) or absent(0) in the document
- All index terms provide equal evidence with respect to information needs.
- Queries are Boolean combinations of index terms.
 - X AND Y: represents doc that contains both X and Y
 - X OR Y: represents doc that contains either X or Y
 - NOT X: represents the doc that do not contain X

Boolean Retrieval

Example Queries:

- Lincon
- President AND Lincon: both words present anywhere in the document
- President AND Lincon AND NOT (automobile OR Car)

Processing Boolean Queries

- Using inverted index
 - Eg: Brutus AND Calpurnia
1. Locate Brutus in dictionary
 2. Retrieve its postings
 3. Locate Calpurnia in dictionary
 4. Retrieve its postings
 5. Intersect the two postings

Brutus →

1	2	4	11	31	45	173	174
---	---	---	----	----	----	-----	-----


Caesar →


1	2	4	5	6	16	57	132	...
---	---	---	---	---	----	----	-----	-----

Calpurnia →

2	31	54	101
---	----	----	-----

⋮

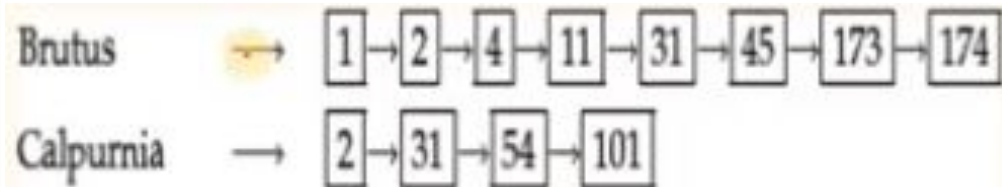
	
Dictionary	

	
Postings	

Implementation Boolean Queries

INTERSECT(p_1, p_2)

```
1  answer  $\leftarrow \langle \rangle$ 
2  while  $p_1 \neq \text{NIL}$  and  $p_2 \neq \text{NIL}$ 
3  do if  $\text{docID}(p_1) = \text{docID}(p_2)$ 
4      then  $\text{ADD}(\text{answer}, \text{docID}(p_1))$ 
5           $p_1 \leftarrow \text{next}(p_1)$ 
6           $p_2 \leftarrow \text{next}(p_2)$ 
7      else if  $\text{docID}(p_1) < \text{docID}(p_2)$ 
8          then  $p_1 \leftarrow \text{next}(p_1)$ 
9          else  $p_2 \leftarrow \text{next}(p_2)$ 
10 return answer
```





Thank You!!