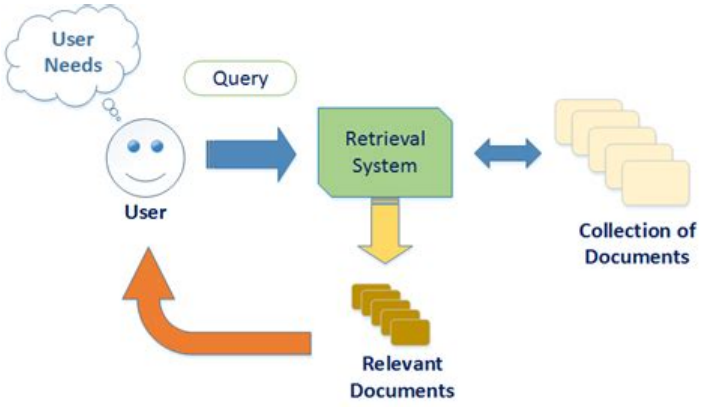
A STUDY ON THE USAGE OF DATA STRUCTURES IN INFORMATION RETRIEVAL

Information Retrival:

Information Retrieval (IR) can be defined as a software program that deals with the organization, storage, retrieval, and evaluation of information from document repositories, particularly textual information. Information Retrieval is the activity of obtaining material that can usually be documented on an unstructured nature i.e. usually text which satisfies an information need from within large collections which is stored on computers. For example, Information Retrieval can be when a user enters a query into the system.

- Finding something relevant in the large pool of imformation
- Index: compressed version of a document to summarize its contents and make it easier to search for
- Index is queried and goes through a ranking model which ranks based on relevance
- This model of ranking are trained by machine learning algorithms
- Information retrival used in libraries, file searching, search engine
- A crawler is used to fetch all relevant indexed documents to be further processed and ranked
- 'Inverted Index' popular way of indexing and efficient to answer user queries
- ' PageRank ' helps ranking web pages
- NDCG Metric helps measure the accuracy of relevance of the web pages
- The success rate or performance of an information retrieval system is calculated based on the response time of the system and also the quality of the output
- To decrease the response time of the IR system we need to concentrate on the type and size of corpus, type of index, and the type of the query along with the searching technique used.



Implementing Data Structures

- Data structures are the different techniques used to store the data in the persistent memory.
- The data structures like arrays, linked structures, hash tables are primarily used for storing the data and hence are classified as **storage structures**.
- Stacks, queues and priority queues are used for processing the data and these are classified as **process-oriented data structures**.
- Collections, sets, linear lists, binary trees, etc. are the data structures which describe the nature of the data held in it and hence we call them **descriptive data structures.**

Retrival Tasks

- Researchers are working towards not only satisfying the queries posed by the users but also to predict the queries that would be posed by the users to a document corpus.
- The effectiveness of any information retrieval system is based on the feedback from the user.

Storage Data Structures

- Information retrieval uses word oriented indexing techniques for retrieving the documents based on the query from the users.
- Primarily they employ the hash function and hash tables.
- The various data structures used for indexing in the information retrieval process. They discuss the effectiveness of hashing, B trees, B+ trees for implementing the index structures.

Hash

A hash data structure links key values to the data items. A hash function is used to map the search key to a key value.
The key value usually denotes the bucket number to which the data item belongs. A bucket is nothing but a memory area.
A hash table is most effective than most of the array structures and can be used as an in- memory data structure.
The hash functions are selected in such a way that it avoids collision.
Hashing as a searching technique is proven to be more apt for equality searches compared with tree structures that are better for range searches.

Process-oriented Data structures in Information retrieval

Stack

A stack is a linear data structure which uses one end of the data structure for storage and retrieval of data items. A stack is used in information retrieval algorithms for string matching in suffix arrays.

Graph

A graph is a data structure with nodes and edges connecting. It is one of the data structures that find wide application in multiple fields. It has been used to find the relationship between two data items or components or to find the connectivity between two different nodes in computer network.

> In information retrieval, graphs are used to find the relationship between the queries posed by the user and the documents present in the corpus. The implementation of semantic nets and frames look into the similarities in the structure of query and document graphs
>
> Graphs are also used to define the search space in the field of information retrieval. The graph structures frame the base for concept networks used in fuzzy information retrieval. Each node represents a concept or a document.
>
> The graphs are also used in web based information retrieval for providing relevance score based on the relevance propagation in document graph

---

Descriptive data structures in information

retrieval

Tree

A tree is a data structure has a data item as its root and the subtrees are generated with a node as a parent node. In a search tree generally the solution is found as leaf. There are various kinds of trees depending upon the way of arrangement and the way of traversal of the tree

B Tree

B tree is a binary search tree that has an additional property of self – balancing. The advantage of B-tree is that the searching needs only a logarithmic time

B+ Tree

A B+ tree is a self- balancing tree that can adjust its height and the nodes are linked used pointers. These pointers enable the range searches to be implemented efficiently in a B+ tree structure

---

Reasearch Analysis by:

- Het Joshi 1MS21CS053
- Hitesh Kumar 1MS21CS054
- Varun Balaji 1MS21CS057
- Malavika Dileep 1MS21CS069
