IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 5.731

Volume 10, Issue 1, October 2021

Analysis of the Suitability of Data Structures and its Application

Aakash Yadav[1], Mali Vivek[2], Malusare Priyanka[3]
Asst. Professor[1] and FYBCOM[2,3]
Uttar Bhartiya Sangh's Mahendra Pratap Sharda Prasad Singh College of Commerce & Science, Mumbai, Maharashtra

Abstract: *This article seeks to offer insight into the utilization of data structures in the field of information retrieval. With the growing demand and importance of sharing and analysing knowledge, information retrieval is becoming a prominent area of research. Data structures have long been a focal point of research in the discipline of computer science. With the rapid increase in data volume, it is imperative to have efficient data structures.*

Keywords: Data structures, Information retrieval

I. INTRODUCTION

Data has consistently been and is a valuable asset that ought to be utilized judiciously for the advantage of organizations and institutions. Due to the proliferation of social networking sites and technological improvements, there is currently a significant volume of data being shared. Information retrieval is the systematic process of finding and retrieving processed data from a pre-existing repository. Information can be defined as data that has been processed. Various computer science disciplines are actively engaged in the study of information retrieval. Information retrieval is a technology employed in various advanced computer science disciplines and incorporates several fundamental concepts in computer science. Information retrieval is an initial step in the text mining process that precedes the usage of other mining processes.

II. ACCESSING DATA

2.1 Information extraction and information retrieval

The terms information extraction (IE) and information retrieval (IR) are occasionally used interchangeably. These domains are completely separate and have different roles that lead to different outcomes. Information extraction lacks explicit objectives or predefined targets that must be achieved. It utilizes templates to provide organization to data that might otherwise lack structure. Information retrieval necessitates the employment of sophisticated techniques in order to meet the user's requirement of locating precise information from an established collection. Choosing a suitable index to enhance querying is an additional function of information retrieval. Moreover, an intelligent information retrieval system utilizes user feedback to enhance the present system and refine the methods. Data mining and information retrieval employ identical methods for summarization and grouping.

2.2 The efficacy of an information retrieval system

The success rate or performance of an information retrieval system is determined by considering the system's reaction time and quality. Regarding the outcome. Another aspect that can be evaluated based on user input is the quality of the information retrieval answer. Precision and recall are often used metrics for assessing the quality of a measurement. Recall metric is defined as the proportion of relevant documents retrieved out of the total number of relevant documents. The precision measure is determined by the ratio of relevant papers found to the total number of documents found. The significance of the materials is a subjective term, despite the fact that these metrics have unambiguous interpretations.

A quantitative parameter for an information retrieval system is its response time, which can be easily quantified. The factors that influence reaction time include the size and organization of the corpus being searched, the type of index being used, and the type of query being sent to the system. To reduce the response time of the IR system, we need to consider the nature and size of the corpus, the type of index, the query type, and the search methodology. Now, we will examine the implementation of data structures in the field of information retrieval and analyze how the choice of data structure affects the performance of retrieval systems.

III. DATA STRUCTURES

Data structures refer to the various techniques employed to store data in long-term memory. Each application possesses a distinct role pertaining to its data structure. The authors classify the data in into different structures according to the specific role it fulfils. Storage structures refer to data storage mechanisms, including arrays, linked structures, and hash tables. Process-oriented data structures, such as stacks, queues, and priority queues, are a set of data structures specifically designed for data processing purposes. There are certain data structures that have not been taken into consideration yet since they provide additional functionality beyond just storing data. These data structures also enable the organization of the data in a way that may be used to describe the data.

IV. TASKS RELATED TO THE RETRIEVAL OF INFORMATION

The primary objective of information retrieval is to respond to a user's inquiry. Researchers are not only seeking to provide answers to user inquiries, but also to predict the questions that users may pose to a collection of documents in the future. In their study, Fei Song and Bruce Croft computed the probability of the user's query phrases being produced for every document in the corpus. The effectiveness of any information retrieval system is contingent upon the feedback provided by the user. The authors in evaluate the efficacy of information retrieval by measuring the degree to which user preferences are fulfilled.

V. UTILIZING DATA STRUCTURES FOR THE STORAGE AND RETRIEVAL OF INFORMATION

Information retrieval utilizes word-centric indexing algorithms to acquire documents in response to a user query. Hash functions and hash tables are commonly used for indexing, while other structures like signature files, inverted files, etc. are also utilized. A hash data structure is used to establish connections between key values and data items. A hash function is used to map the search key to a corresponding key value. The key value often serves as an indicator of the bucket number to which the data item belongs. A bucket is simply a container for storing items. A hash table can serve as an in-memory data structure and is generally more efficient than most array formats. Avoiding collisions is achieved through meticulous selection of the hash functions. When comparing hashing and tree structures for range searches, it becomes evident that hashing is more suitable for equality searches. The hash functions generate an index that facilitates the retrieval of the pages that match the user's query. A signature file, also referred to as a hash file, is used for the purpose of filtering. The filtering process frequently identifies the sites that closely correspond to the query. The hash function is employed consistently during the filtering process to produce a distinct signature for every document. An inverted file consists of a hash file containing a collection of sorted words, each of which is associated with its respective page through a sequence of pointers.

VI. DATA STRUCTURES DESIGNED FOR INFORMATION RETRIEVAL WITH A FOCUS ON THE PROCESSING ASPECT

A stack is a linear data structure that employs one end for the storage and retrieval of data elements. String matching in suffix arrays is performed using a stack in information retrieval techniques. A graph is a data structure comprising nodes and edges that connect them. It is a highly utilized data structure in various industries. It has been utilized to ascertain the correlation between two computer network nodes or to establish the linkage between two data items or components. Graphs are utilized in information retrieval to establish the correlation between user queries and the documents in the corpus. Graph structures are used as the framework for concept networks in fuzzy information retrieval. Each node represents a distinct thinking or written item. An edge is utilized in a concept network to connect two separate ideas, Ci, to a document, Di. The edge is allocated a real value between zero and one, representing the fuzzy weighting given to the relationship. Graphs are used in web-based information retrieval to calculate relevance scores by propagating relevance in document graphs [9]. Graphs in the realm of information retrieval have further uses such as collaborative filtering, categorization of the retrieved documents, and unified link analysis.

VII. DESCRIPTIVE DATA STRUCTURES FOR INFORMATION RETRIEVAL

A tree data structure is formed by a parent node that generates its subtrees, with each subtree having a root that contains a data item. Typically, the solution to a search tree is considered to be a leaf node. Trees can vary in their structure and traversal methods, resulting in a wide range of various types. An B tree is a type of binary search tree that has the additional capability of self-balancing. The advantage of using a B-tree is that the search operation requires just a logarithmic amount of time. A B+ tree is a height-adjustable, self-balancing tree structure that utilizes linked nodes as pointers. These pointers enable efficient execution of range searches in a B+ tree structure. When navigating a digital tree, the right subtree is visited when the bit value is 1, whereas the left subtree is examined when the bit value is 0. A search operation can be conceptualized as the creation of a new node in a search tree. Binary tree structures such as B trees and B+ trees are used to implement the index in the information retrieval process. A B-tree is used to implement the inverted files. A Prefix B-tree rebuilds the tree every time it is searched; it does not store the complete prefixes. This method incorporates the advantages of B-trees, digital search trees, and key compression techniques. Furthermore, it reduces the computational load associated with compression techniques. A string can be stored in a trie data structure by starting at the root node and traversing towards the leaf node. A PAT tree is a binary tree structure employed in the domain of information retrieval. PAT is an abbreviation for PATRICIA, which expands to "Practical Algorithm to Retrieve Information Coded in Alphanumeric." A PAT, which is a straightforward version on a trie, compresses every path where all internal vertices have exactly one child into a single edge. The trie data structure has a radix of 2, indicating that each node has two branches: left and right. Additionally, each bit of the key is compared individually. Contrary to other attempts, Patricia trees do not contain any nodes that have only one child. Each node possesses a minimum of two descendants or is a terminal node. This immediately implies that the proportion of internal (non-leaf) nodes to leaves is equivalent.

VIII. CONCLUSION

Despite having established a strong foundation, we are ready to move forward. Data Structures is a vast subject that encompasses more than just stacks, queues, and linked lists. Additional data structures include Maps, Hash Tables, Graphs, Trees, and more. Every data format has advantages and drawbacks, and its adoption should be determined by the requirements of the application. A computer science student should possess a comprehensive understanding of the fundamental data structures and their associated operations. Many high-level and object-oriented programming languages, such as C#, Java, and Python, provide a large number of these data structures. Consequently, it is crucial to comprehend the inner workings of objects. Dynamic data structures necessitate the use of dynamic storage allocation and reclamation. The task can be accomplished either explicitly by the programmer or implicitly by a high-level language. Understanding the fundamentals of storage management is crucial since these tactics have a significant impact on program behavior. The fundamental idea is to maintain a reservoir of memory fragments that can be used to store dynamic components of data structures as required. Once the need for allotted storage has ended, it can be returned to the pool. It can be utilized and repeatedly employed in this fashion.

Linear linked lists are data structures that facilitate efficient traversal of data elements. It utilizes pointers or references to aggregate multiple data items into a single data point. In the case of a doubly linked list, each node would include references to both the preceding and succeeding data elements. Linear linked lists are used to implement posting lists. A posting list is a data structure used to track documents that include a specific word. Usually, a word dictionary is generated, and for each phrase, a posting list is generated including a list of documents that contain that particular term.

REFERENCES

[1] S. Ceri et al., Web Information Retrieval, Data-Centric Systems and Applications, DOI 10.1007/978-3-642-39314-3_2, © Springer Verlag Berlin Heidelberg 2013

[2] Fei Song, W Bruce Croft. A general language model for information retrieval. Proceedings of the eighth international conference on Information and knowledge management (ACM) 1999/11/1. pp316-321

[3] Falley. P "Categories of Data Structures", Journal of Computing Sciences in Colleges - Papers of the Fourteenth Annual CCSC Midwestern Conference and Papers of the Sixteenth Annual CCSC Rocky Mountain Conference. Volume 23 Issue 1, October 2007. PP. 147-153, 2007-10-01

[4] B. Zhou and Y. Yao Evaluating information retrieval system performance based on user preference JIIS, 34:227–248, 2010

[5] Rudolf Bayer and Karl Undervaluer, “Prefix B Trees” ACM Transactions on Database Systems, Vol. 2, No. 1, March 1977, Pages 11-26.

[6] Morin, Patrick. "Data Structures for Strings", chapter 7, March 2012.

[7] Pogue, C. & Willett, P. (1987). Use of Text Signatures for Document Retrieval in a Highly Parallel Environment. Parallel Computing, 4, 259-268.

[8] Nicholas. B Elkin, W.B Roces Raft, Retrieval Techniques, Annual Review of Information Science and Technology, Volume 22. 1987. Martha E. Williams, Editor Published for the American Society for Information Science (ASIS) by Elsevier Science Publishers