An Enhanced Data and Information Retrieval System in Distributed Environment Using Divide-and-Conquer Algorithm Augmented with Firefly Algorithm

Alawode, J.A. & **Hammed,** M.
Computer Science Department, Federal Polytechnic, Ilaro
ademola.alawode@federalpolyilaro.edu.ng & mudasiru.hammed@federalpolyilaro.edu.ng

ARTICLE HISTORY

Received: May 10, 2023
Revised: May 19, 2023
Accepted: June 10, 2023

Abstract

In the advent of rapid increase in the worldwide data and information with its usage, effective and faster retrieval of information are becoming important for business organizations and innovations. The literature revealed that information retrieval is a challenging task which has attracted various information retrieval models, algorithms, and techniques from researchers. But only an effective and faster system can handle a large-scale database in a distributed environment where concurrent requests can be made by the user or groups of users irrespective of whether data are sorted or not. This study used divide-and-conquer augmented with firefly algorithm for effectiveness and better performance of retrieval system and the result shows that the system used in this study proved to be scalable for any form of retrieval system.

***Keywords*:** Retrieval system, database system, distributed environment, divide-and-conquer, firefly algorithm

Citation
Alawode, J.A. & Hammed, M. (2023). An Enhanced Data and Information Retrieval System in Distributed Environment Using Divide-and Conquer Algorithm Augmented with Firefly Algorithm. *International Journal of Women in Technical Education and Employment*, 4(1), 56 – 62.

Introduction

Data is an asset of importance in everyday life for every user in every field. Important concern about the data is that it should be stored at a safe place and that they should be easy to access (Binita & Blessy, 2022). The roles that the database play in today's IT-based economy is much significant and industries, organizations and systems depend on the database accuracy to perform most of their operations (Ahmed, 2007). Current growth of electronic resources such as journals, textbooks, articles etc. increase the number of scientific resources. The emergence of web sites, blogs, efficient domain with specific information making retrieval system to becoming very important (Claire & Jacques, 2010). Information has become a critical asset for individual and any innovative company, with the growth in the field of information and communication technology (Alok & Sagar, 2017). The research data such as research results, projects results and information about organizations that researchers publish on the web page play very important roles in modern research (Niranjan, *et. al.*, 2016). There is high dependence on modern research and research results produce requirements for research capability to retrieve information about research in efficient manner (Niranjan, *et. al.*, 2016). Information retrieval system can be viewed as a process where user capable of converting its information needs into a list of documents stored in the base. Retrieval system can be seen as an act of finding or discovering data or information that is stored in a base. Information retrieval is not limited to only information stored in the database, but the objective of information retrieval system is to enable the user to find relevant information from an organized collection of documents in the database. The retrieval system searches and retrieves specific data or an information such as salary of a particular employee, group of employees in a particular department and group of employee in more than one department. But, conventional database management systems, such as Access, Oracle, MySQL, and so on, deal with structured data. One of the main problems to deal with the management of such information is that the interoperability between various databases and information systems is very weak (Niranjan, *et. al.*, 2016). The literature revealed that there some issues in most of an existing information retrieval model. Among the models that were examined include document and query indexing, query evaluation, and system evaluation. Thus, some of these issues cause delay when searching and retrieving data or information from a pool large database. Meanwhile, effectiveness of the information retrieval system depends on the efficiency of the search and retrieval techniques (Francis & Mrindoko, 2019). Many studies have proposed different retrieval models which include Boolean retrieval model, ad hoc retrieval model, inverted index, vector space model and probabilistic model etc. But the literature revealed that some of these retrieval methods proposed in an existing study cannot quickly respond to concurrent retrieval of information at same time and therefore, some of the existing studies examined are not scalable for large scale information retrieval. Some of the studies examined include Priyanka, *et.al.*, (2016) proposed Boolean model, the model based on the Boolean algebra logic where the data to be retrieved and the query for retrieving the data must be indicated as a set of indexes. This model used different Boolean operators such as AND, OR, NOT for query formulation. The drawback of the model is that only precise matching is feasible while partial matching may not be feasible. Thus, formulation of the query expression is more complicated, and the retrieval of data are also not in ranked. Joydip & Pushpak, (2016), proposed probabilistic model which is capable to archives correct match in a set of documents and the strength of this model is that the query formulation is possible. But the weakness is that the probabilities which may be difficult to estimate and unrealistic assumptions of independence. Matsumoto & Hung, (2010) proposed fuzzy clustering, this model is a clustering algorithm which high performance achievement depends on the number of clusters. However, there is need for strong and faster algorithm for quick processing of large document irrespective of whether the data are structured or not and whether data are sorted or not. This study used divide-and-conquer with binary search algorithm augmented with firefly algorithm which enhances the process of searching and retrieving of information in the database. Thus, method is scalable for concurrent retrieval of data and information compared to Boolean and probabilistic model which require estimated of data. The method used appropriate and quickly combining the search results of each subset to form a result of original search and retrieving.

Methods

A large database must be supported with very strong search algorithms that will allow faster retrieval of information from a pool of data in the base. The system architecture is depicted in figure 1.

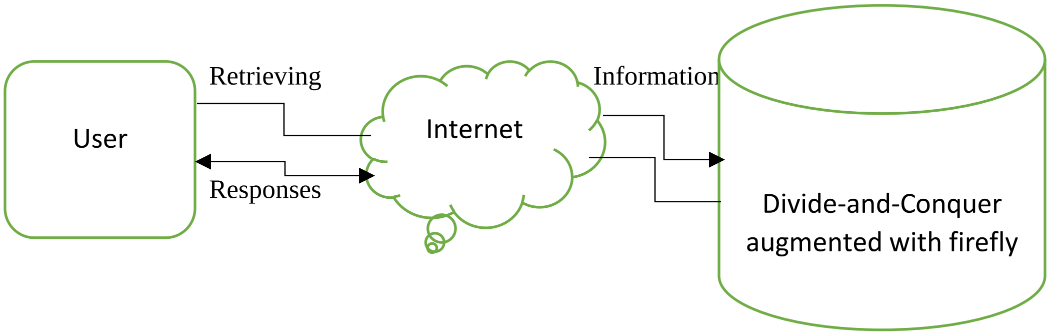


Figure 1: Information Retrieval System Architecture

This study considers retrieval system as a problem which is suitable to be solved with recursive algorithm. Thus, the study applied divide-and-conquer rules with binary search algorithm augmented with firefly algorithm which enhances the retrieval system. When user wants a piece of information or groups of users send their request concurrently for different information in the large database. Divide-and-conquer often follow a generic pattern to tackle a problem of size n by recursively dividing into a sub-problems of size n/b and then combining these answers in $0(n^d)$ time. That is, all information or set of data in the database were divided into sub-information or subsets of data and sub-information or subsets of data were searched recursively for target information or data. The results of search in sub-information or subsets of data were assembled to form the result of original search where the requested information or data by the user or group of users is retrieved. Divide-and-Conquer rule is depicted in the figure 2.

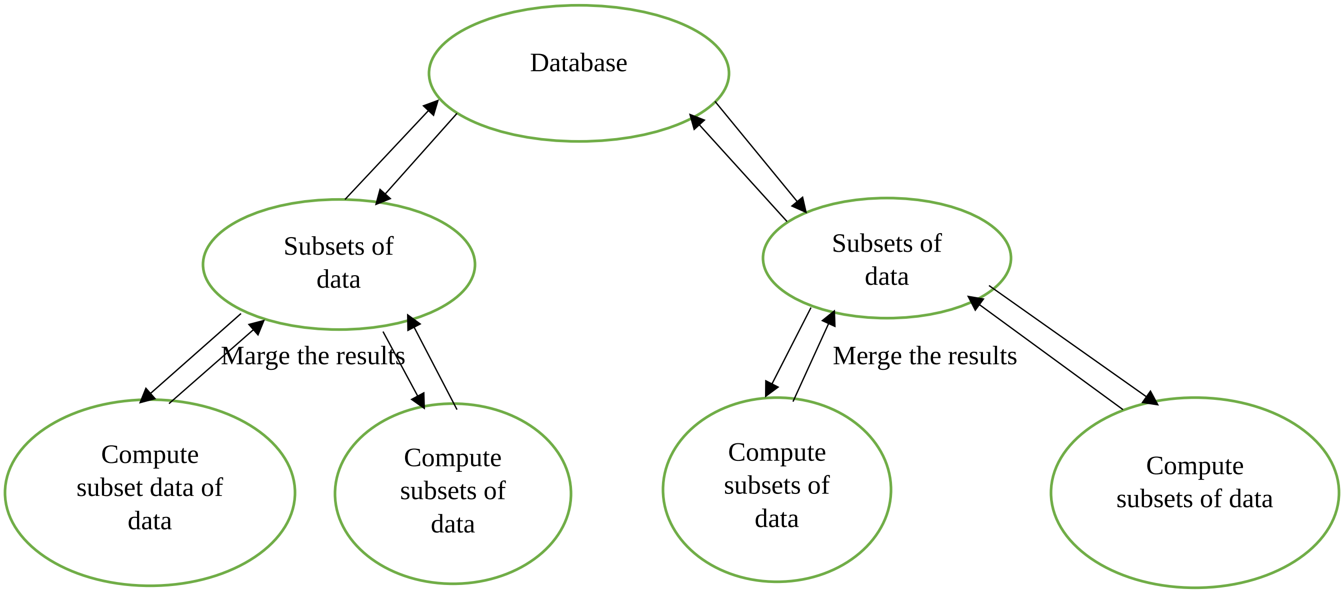


Figure 2: Logical Representation of Divide-and-Conquer for Information Retrieval

Algorithm 1

Step1: Breaking it into sub-problems of smaller instances with the same type of problem.

Step2: Recursively solving these sub-problems.

Step3: Appropriately combining the answers to output final result.

The divide-and-conquer augmented with firefly achieved combination of results from a large number of subsets through the flashing characteristics behavior of fireflies. This enhances the performance of the system to quickly search and retrieve data and information for large number of users. The database is divided into several subsets of data. In each subset there is a result of search. The fitness value of the data outlines the hierarchy, the data with the best fitness will be the result of search in each subset, the other data will be considered as data, but not the result of subsets. The results of subsets are appropriately combined as the information or data retrieval result which will be returned to the users. This process continues to update every time that user or group of users want to retrieve data or information from the database. The firefly strategies are depicted in algorithm 2.

Algorithm 2

Step1: Generate a random set of solutions, $\{x_1, x_2, \ldots x_n\}$

Step2: Calculate intensity for each solution member, $\{r_1, r_2, \ldots, r_n\}$

Step3: Move each firefly i towards other brighter fireflies, and if there is no other brighter firefly, move it randomly.

Step4: Update the set of solutions.

Step5: Terminate if a criterion is fulfilled, otherwise go back to step 2.

Mathematically, each firefly has a location $X = (x_1, x_2, \ldots x_n)$ in a n-dimensional space and a light intensity $I(x)$ or attractiveness $\beta(x)$ which are proportional to objective function $f(x)$.

Therefore, the attractiveness of a firefly is modelled with equation 1

$$\beta = \beta_0 e^{-kr2} \qquad (1)$$

The r is defined as the distance between any two firefly i and j at x_i and x_j respectively which is modelled in Cartesian distance in equation 2

$$r_{ij} = \|x_i - x_j\| \qquad (2)$$

For any given two fireflies x_i and x_j, the movement of firefly i is attracted to another more attractive (brighter) firefly j is modelled in equation 3

$$x_i^{t+1} = x_i^t + \beta_0 e_{ij}^{-kr2}(x_j^t - x_i^t) + \alpha(Rand - \frac{1}{2}) \qquad (3)$$

This algorithm works based on global communication among the set of fireflies used to achieve best information retrieval. Firefly approach efficiently optimized the divide-and-conquer to iteratively achieve the best response time when users are concurrently retrieving information from distributed database. The study used equation 1, 2 and 3 to model the iterative performance of the fireflies to enhance divide-and-conquer approach. The flowchart in figure 3 depicts the firefly algorithm processes.

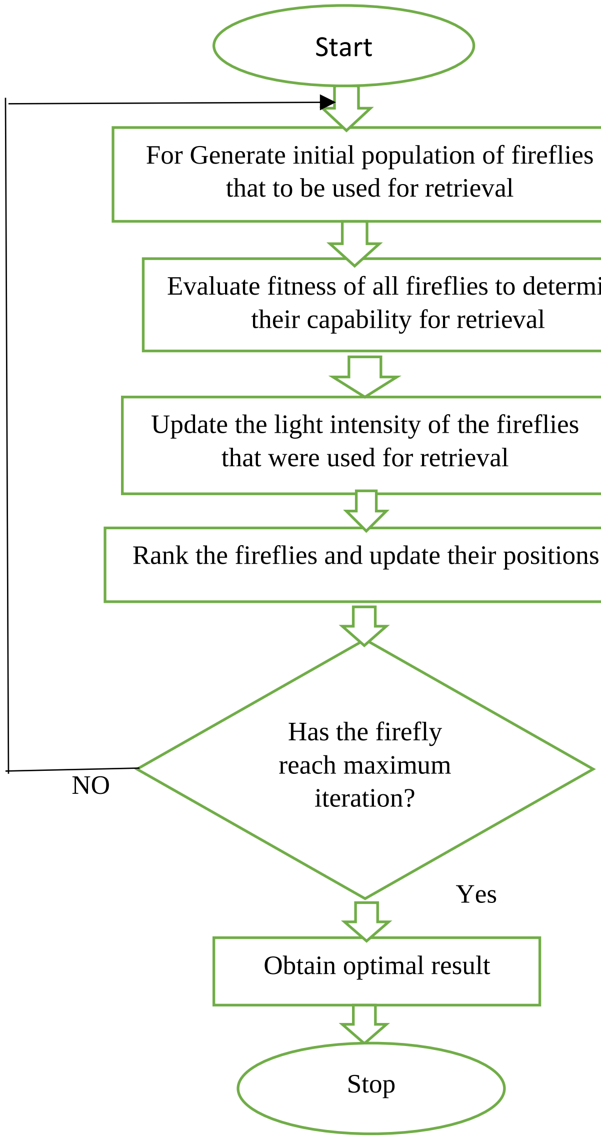
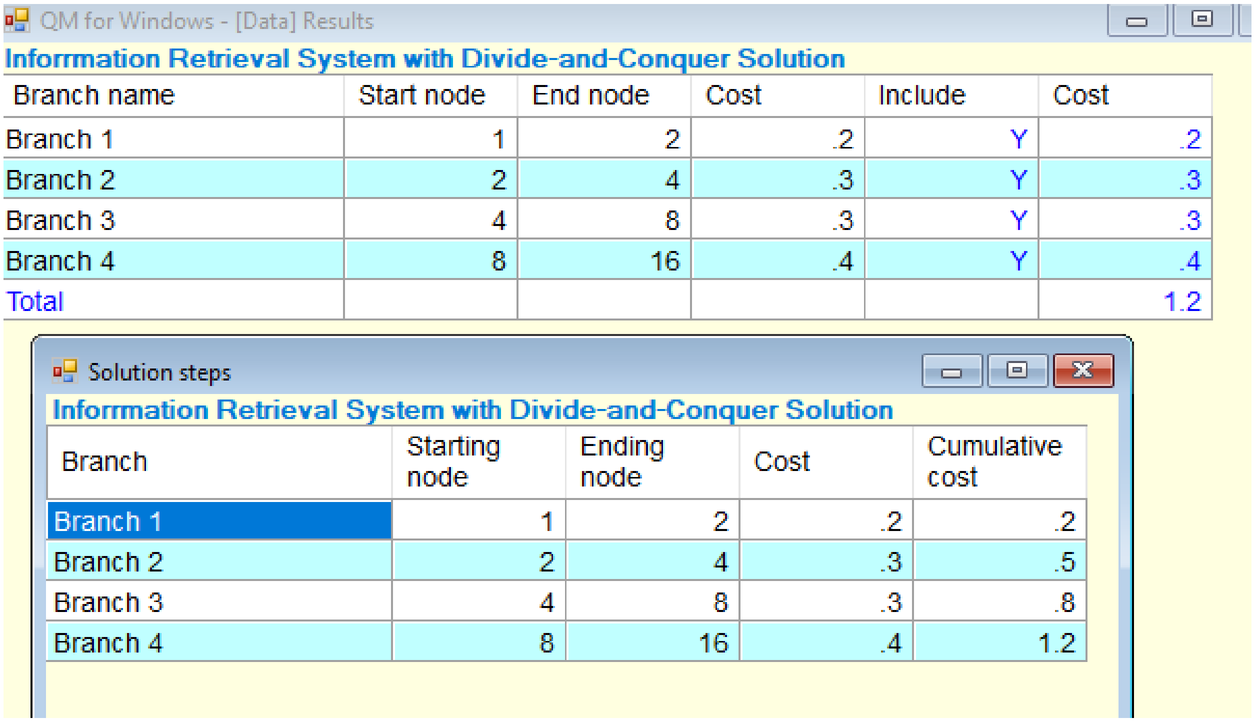

Start
For Generate initial population of fireflies that to be used for retrieval
Evaluate fitness of all fireflies to determine their capability for retrieval
Update the light intensity of the fireflies that were used for retrieval
Rank the fireflies and update their positions
Has the firefly reach maximum iteration?
NO
Yes
Obtain optimal result
Stop

Figure 3: Logical Representation of Firefly for information Retrieval

Result

The techniques used in this study attained high degree of accuracy in terms of data and information retrieval in distributed environment. Divide-and-conquer rule was implemented in QM for windows environment, the environment shows the performance of divide-and-conquer as it is represented with spanning tree in figure 2. The cost shows the time taken the system to breakdown the data into subsets of data and merge the results from each subset. The result of implementation shows that time taken to get the original result is relatively low. When divide-and-conquer were combined with firefly algorithm, thus, effective retrieval system yield better results. The results of implementation shown in table 1. The analysis of implementation goes thus: the data analyzed has 16 subsets and the study used cost to measure the time taken the divide-and-conquer to retrieve its result from each subset. The time/per seconds was used by the study and total time taken (i.e. cumulative cost) the divide-and-conquer to combine results from 16subsets of data used was 1.2seconds which is relatively low. This shows that the retrieval system using divide-and-conquer optimized with fireflies achieved high degree of accuracy.

Table 1: Implementation of Divide-and-conquer



QM for Windows - [Data] Results

Inforrmation Retrieval System with Divide-and-Conquer Solution

Branch name	Start node	End node	Cost	Include	Cost
Branch 1	1	2	.2	Y	.2
Branch 2	2	4	.3	Y	.3
Branch 3	4	8	.3	Y	.3
Branch 4	8	16	.4	Y	.4
Total					1.2

Solution steps

Inforrmation Retrieval System with Divide-and-Conquer Solution

Branch	Starting node	Ending node	Cost	Cumulative cost
Branch 1	1	2	.2	.2
Branch 2	2	4	.3	.5
Branch 3	4	8	.3	.8
Branch 4	8	16	.4	1.2

Conclusion

In the advent of rapid increase in the worldwide data and information with its usage, effective and faster retrieval of information techniques are become important for businesses organizations and innovations. This study used divide-and-conquer as an effective technique to retrieve data and information in large scale database environment. The study enhances the system for better performance with firefly algorithm, thus, method used in this study outperform existing retrieval systems in the knowledge based, data mining and clustering approaches.

References

Ahmed, K., Panagiotis G. I. & Vassilios S. V. (2007). Duplicate Record Detection: A Survey. *IEEE Transactions on Knowledge and Data Engineering, 19(1).*

Alok, K., & Sagar, J. (2017). A Literature Review on Patent Information Retrieval Techniques. *Indian Journal of Science and Technology, 10(37)*

Binita, T. & Blessy, T. (2022). A Data Deduplication Approach for Eliminating Duplicate File

Upload over Cloud. *International Journal of Enhanced Research in Science, Technology & Engineering, 11(2)*.

Francis, A. R. & Mrindoko, R. N. (2019). Towards Enhancing Information Retrieval Systems: A Brief Survey of Strategies and Challenges. *IEEE, 2019 11th International Congress on Ultra-Modern Telecommunications and Control Systems and Workshops (ICUMT)*.

Joydip, D. & Pushpak, B. (2016). Ranking in Information Retrieval. *M.Tech Seminar Report, Department of Computer Science and Engineering, Indian Institute of Technology, Bombay*.

Niranjan, L., Samimul, Q. & Savita, S. (2016). Information Retrieval System and Challenges with Dataspace. *International Journal of Computer Applications (0975 – 8887) 147(8).*

Matsumoto, T. & Hung, E., (2010) Fuzzy clustering and relevance ranking of web search results with differentiating cluster label generation. *In: 2010 IEEE international conference on fuzzy systems (FUZZ), 1–8.*

Priyanka, M., Nidhi M., & Abhishek K. (2016). Document Ranking using Customizes Vector Method. *A Review, IJCSMC, 5(3).*