An Efficient Searching Algorithm for Data Mining in Bioinformatics

Md. Selim Hossain[1], Md. Biplob Hossain[2], Md. Mahasin Ali[3], Md. Dulal Haque[4] and Md. Abubakar Siddik[5]

[1, 2, 3, 4, 5] Department of Telecommunication and Electronic Engineering, Hajee Mohammad Danesh Science and Technology University, Dinajpur-5200, Bangladesh

ABSTRACT

Bioinformatics is the application of computational techniques to analyze the Information associated with bimolecular on a large-scale. Bioinformatics is the science of storing, extracting, organizing, analyzing, interpreting and utilizing Information from biological sequences and molecules. Bioinformatics may be used in Sequence analysis, Genome annotation, Analysis of gene expression, Analysis of mutations in cancer etc .Data Mining (DM) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. It requires intelligent technologies and the willingness to explore the possibility of hidden knowledge that resides in the data. Data mining have some pre processing, processing and post processing activities. In this paper we analyze how data mining may help bio-medical data analysis and outline some research problems that may motivate the further developments of data mining tools for bio-data analysis However, one of the drawback of searching of data mining is its huge time complexity. The time complexity of linear search is linearly increased with increasing input. For binary search the time complexity is increased by logarithmic function. Moreover binary search require the input data to be sorted. So we proposed a new algorithm to solve these problems. In our proposed algorithm the input data is divided first by several time slots, then searching is performed by using the concept of tree. By using this algorithm we see the time complexity is reduced incredibly. For this the Bit Error Rate (BER) of the system is also reduced hence it increase the performance of the system.

Keywords: *Bioinformatics, Data Mining, Searching, BER, Efficient Algorithm etc.*

1. INTRODUCTION

Search advertising is typical heterogeneous multi-dimensional data sets whose click-through rate depends on query words, terms and other terms. Traditional data mining methods, limited to homogenous data source, Represent search ads as the vector space model, so they fail to sufficiently consider the search advertisements [1]. Data mining is the analysis or step of the Knowledge Discovery in Databases process an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets (big data) involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems[2,3]. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use for our desire result. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating to achieve goal and knowledge from large amount of data [4, 9].

2. SEARCHING IN DATA MINING

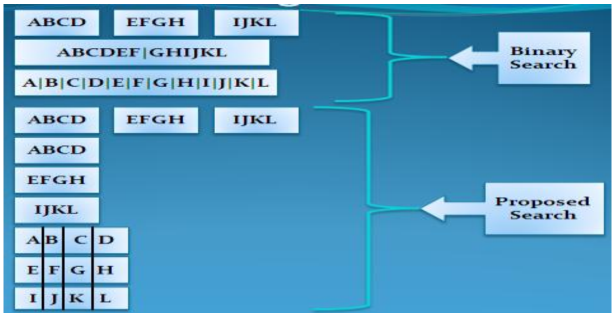
The sequential search algorithm iterates through each item in our data structure in search for a specific value. If the current item matches, we can return, else we must continue to the next item. In the worse case, this requires that we search through all items because in an unsorted structure, we cannot Say whether an untested value is the value we are searching for in specific values and more execution time is required. The binary search algorithm in data mining begins by comparing the target value to the value of the middle element of the sorted array. If the target value is equal to the middle element's value, then the position is returned and the search is finished. If the target value is less than the middle element's value, then the search continues on the lower half of the array; or if the target value is greater than the middle element's value, then the search continues on the upper half of the array. The techniques and methods and their applications in bioinformatics study, focusing on data integration, text mining and graph-based data analysis [2]. A model that is utilized to represent the master search schema, and an effective interface extraction algorithm based on the hierarchical structure of the web and pattern is developed to capture the rich semantic relationships of the online bioinformatics data sources. The k-means algorithm is a simple iterative method to partition a given dataset into a user specified number of clusters, k. This algorithm has been discovered by several researchers across different disciplines. Apriority is a seminal algorithm for finding frequent item sets using candidate generation [5]. One of the most popular data mining approaches is to find frequent item sets from a transaction dataset and derive association rules [10]. Moreover few others algorithm exist in data mining.

3. LIMITATION OF EXISTING METHOD IN SEARCHING OF DATA MINING

In sequential search the worse case, this requires that we search through all items because in an unsorted structure, we cannot say whether an untested value is the value we are searching for in specific values and more execution time is required [7,8]. The time complexity of sequential is O (n). In Binary search algorithm the data must be in sorted. The time complexity of binary search algorithm is O (Log (n)). In addition to being sensitive to initialization, the k-means algorithm suffers from several other problems. First, observe that k-means is a limiting case of fitting data by a mixture of k Gaussians with identical, isotropic covariance matrices (=σ2I), when the soft of data points to mixture components are hardened to allocate each data point solely to the most likely component. Similarity Search and Data Mining have become widespread problems of modern database applications involving complex objects [5, 6]. In addition to being sensitive to initialization, the k-means algorithm suffers from several other problems. First, observe that k-means is a limiting case of fitting data by a mixture of k Gaussians with identical, isotropic covariance matrices (=σ2I), when the soft assign of data points to mixture components are hardened to allocate each data point solely to the most likely component. So We Proposed an algorithm which is called Burst Search Algorithm that has less time complexity compared with other searching algorithm of data mining and time complexity is O(Log2(n*N)), where N is the number of slots of searching sentences.

4. PROPOSED CONCEPT AND ALGORITHM THEORY BASED PROCEDURE

1. Input the original message
2. Divide the original message into N words/slots
3. Rearrange the slots in parallel form
4. Input the characters that we want to search
5. Search each position of character of each time slot parallel at a time.
6. Search procedure will execute from left to right and continue until the last character of the slots.



4.1 Complexity Calculation

$1=n/2x$
$=>2x=n$
$=>\log2\ (2X) =\log2\ (n)$
$=>x*\log2\ (2) = \log2\ (n)$
$=>x*1=\log2\ (n)$
$=>x=\log2n$

Multiplying by Log on both side
n=Number of character of a word
N=Number of slots
$\log_2 (2) =1$

For N number slots of words $x=N*\log2n$
Complexity of our proposed Algorithm

$1=n*N/2x$
$=>2x=n*N$
$=>\log2\ (2X) =\log2\ (n*N)$
$=>x*\log2\ (2) =\log2\ (n*N)$
$=>x*1=\log2\ (n*N)$
$=>x=\log2\ (n*N)$

5. DATA MINING MODEL WITH PROPOSED ALGORITHM

So We Proposed an algorithm which is called Efficient Searching Algorithm that has less time complexity compared with other searching algorithm of data mining and time its time complexity is O(Log2(n*N)), where N is the number of slots of searching sentences. For the performing of this algorithm divide the searching sentence into word by word then Keep each word into individual time slot and Make individual tree for the letters of each time slot in which tree is visited from its root then the left most vertices, edges and so on. Then the edges are divided into forward, backward and cross edges and the vertices are ordering into preordering, post ordering and reverse post ordering as well as visit all the trees by using these procedures. At last the evaluating and interpreting the desired result is found.

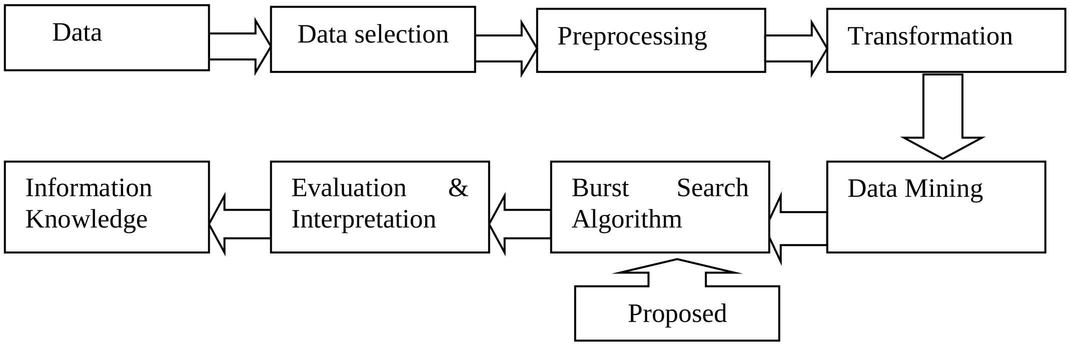


Fig. 5.1. Data Mining Model with Proposed

From the block diagram of data mining model, it can be described that data is to be selected by the selection group which is then processing and Transformed to the system model. Data Mining is an analytic process designed to explore data (usually large amounts of data typically business or market related also known as big data in search of consistent patterns or systematic relationships between variables, and then to validate the findings by applying the detected patterns.

6. SIMULATION AND RESULT

6.1 Time complexity of proposed and existing algorithms

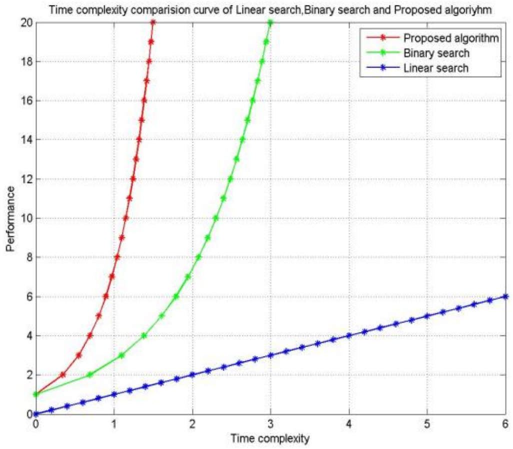


Fig. 6.1. Time complexity comparison curve of linear search, binary search and proposed algorithm.

Figure 6.1 shows time complexity of various algorithms. From this graph we see that the time complexity of linear search is linearly increased with increasing input and it is shown by blue color. Time complexity of linear search is denoted by O(n). For binary search the time complexity is increased by logarithmic function and it is shown by green color. Time complexity of binary search is denoted by O(logn). The red color shows the time complexity of our proposed algorithm. This is we get by partitioning input into several time slot and its complexity is denoted by O(log(n*N)).From this observation we also see that our proposed algorithm has lower time complexity than linear and binary search.

6.2 Bit Error Rate vs. Eb/No of proposed and existing algorithms

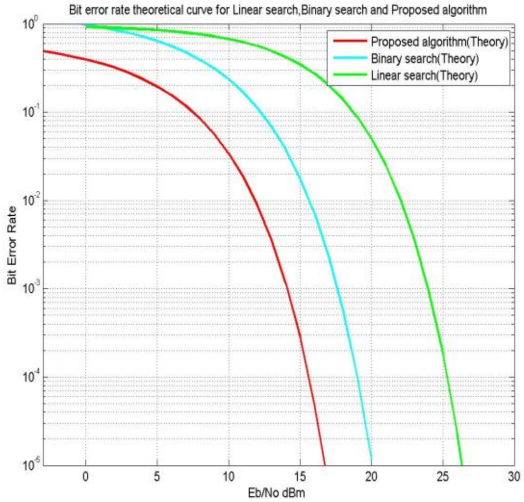


Fig. 6.2(a). Bit Error Rate vs Eb/No theoretical curve of linear search, binary search and proposed algorithm

From the MATLAB simulation of Bit Error Rate vs. EB/No curve, Figure 6.1 (a) shows the theoretical result while Figure 6.2 (b) shows the comparison of theoretical and practical result. In Figure 6.1(a), the green color indicate linear search, cyan color for binary search and red color for proposed algorithm. From this we can say that the theoretical bit error rate of proposed algorithm is lower than linear and binary search algorithm.

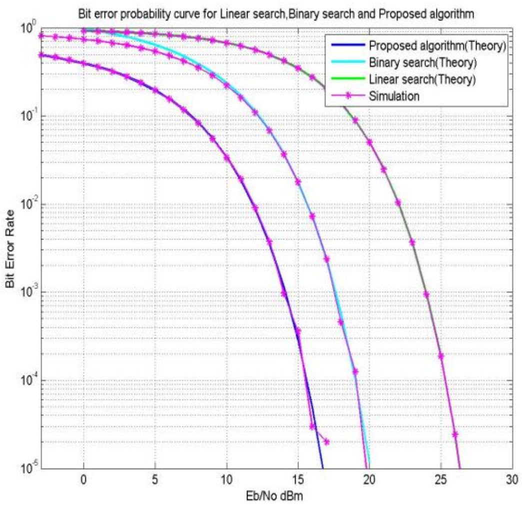


Fig. 6.2(b). Bit Error Rate vs. Eb/No theoretical and simulation curve of linear search, binary search and proposed algorithm.

In Figure 6.2(b), magenta color shows the simulation curve while the theoretical curve of linear search, binary search and proposed algorithm are represented by green, cyan and blue respectively. If we observe we will see the proposed algorithm, the theoretical blue color signal is mostly overlapped by the simulated magenta color signal and slightly changes in the last portion that is negligible. For binary search, there is a difference between theoretical cyan color and simulated magenta color. In the case of linear search the theoretical green color is almost overlapped by the simulated cyan color.

7. CONCLUSION

Bioinformatics is the science of managing, mining, and interpreting information from biological data. As tremendous progress has been made over the years, many of the fundamental problems in bioinformatics, such as protein structure prediction, gene-environment interaction, and regulatory pathway mapping, are still open. Data mining will play essential roles in understanding these fundamental problems and development of novel diagnostic solutions in post-genome medicine. The ongoing influx of these data, the inherent uncertainties in data collection processes, and the gap between data collection and knowledge gathering have collectively created exciting opportunities for data mining researchers. The inevitable rise of clinical data Warehouses will increase the potential for data mining technique to improve the quality and decrease the cost of bioinformatics.

REFERENCES

[1] Zhu Liping Ji Lianen Guo Wensheng,” Heterogeneous Data Mining in Search Advertisement Click Rates”, 2009 International Conference on Web Information Systems and Mining

[2] Xiaohua Hu,” Data Mining and Its Applications in Bioinformatics: Techniques and Methods”, 2011 IEEE International Conference on Granular Computing

[3] Jun Liu, Zheng Yan, Athanasios V. Vasilakos, Laurence T. Yang ““Data Mining in Cyber, Physical and Social Computing” IEEE Systems Journal on December 31, 2014

[4] N.M. Luscombe,D. Greenbaum,M. Gerstein “What is bioinformatics? An introduction and overview” Yearbook of Medical Informatics 2001

[5] John L. Pfaltz, Christopher M. Taylor “Closed Set Mining of Biological Data,” BIOKDD02:Workshop on Data Mining in Bioinformatics (with SIGKDD02 Conference).

[6] Matthew Herland, Taghi M Khoshgoftaar and Randall Wald “A reviewof datamining using big data in health informatics,” Herlandet al. Journal of Big Data 2014

[7] Manjunath T. N., Ravindra S. Hegadi, Umesh I. M., Ravikumar G. K “Realistic Analysis of Data Warehousing and Data Mining Application in Education Domain,” International Journal of Machine Learning and Computing, Vol. 2, No. 4, August 2012

[8] Terence Critchlow, Ron Musick, Tom Slezak “An Overview of Bioinformatics Research at Lawrence Livermore National Laboratory,” U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405-ENG-48. UCRL-JC-138042

[9] Jiawei Han “How Can Data Mining Help Bio-Data Analysis?” BIOKDD02: Workshop on Data Mining in Bioinformatics (with SIGKDD02 Conference).

[10] Jiawei Han, Micheline Kamber,Jian Pei “Data Mining Concepts and Techniques Third EditionMorgan Kaufmann Publishers is an imprint of Elsevier.225 Wyman Street, Waltham, MA 02451, USA.