IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

Intelligent Knowledge Extractor

Pratiksha Patil[1], Prof. Dr. A.D. Thakare[1], Kanchan Pawar[1], Poonam Kshirsagar[1], Rupa Solapure[1]

Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune[1]

Abstract: Now a days it is important to extract knowledge from structured i.e. relational databases, XML and unstructured i.e. text documents, images etc sources. The resulting knowledge needs to be in a human-readable format. The large amount of online repositories requires extracting useful knowledge. This extracted knowledge is then useful for further processing like analysis etc. There are some domains like Biomedical, E-Commerce, Banking, where the huge amount of data present. This huge amount of data needs analysis, in order to produce useful knowledge which will helpful for analysis like in biomedical domain analysis of Clinical data used to predict the Disease, Treatments etc. There are various techniques which are used for knowledge extraction. In this proposed work, we are using Clustering technique. Clustering analysis is an important field of artificial intelligence and data mining. The basic idea is to use words and characters from the documents for checking degree of similarities among documents and cluster those documents without prior knowledge. This paper introduces the proposed work based on clustering biomedical abstracts to extract the hidden knowledge. The document clustering of biomedical abstracts will be based on Genetic algorithm. The work will be based on Genetic Algorithm to find optimal cluster centre. Genetic algorithm performs simultaneous mutation. The proposed algorithm outperforms the K-Means and Simple GA.

Keywords: Genetic Algorithm, Clustering, TF-IDF, Biomedical abstracts.

I. INTRODUCTION

Genetic Algorithm is an optimization techniques inspired by natural selection and natural genetics. Unlike many search algorithms, which perform a local, greedy search, GA is a stochastic general search method, capable of effectively exploring large search spaces [2]. The Biomedical repository contains large set of abstract of several diseases like Cancer, MTB, and Malaria etc. This abstracts needs to be analyzed, to make the data useful.

In the analysis of a Biomedical abstracts an important step is the clustering of various abstracts as per their disease names. This clustering of abstract will make the searching operation on data easier. In this work, the each abstract is considered as a document. The quantity of documents that needs to be converted into digital format is thus increasing, creating the need for systems capable to extracting knowledge and understanding document automatically [4].

II. CLUSTERING TECHNIQUES

Clustering is a process of dividing data into groups of similar objects. This grouping is such that intra-cluster similarity should be high and inter-cluster similarity should be low. The clustering algorithm attempts to find natural groups of components, based on some similarity [5].

A. K-means algorithm

K-means is a greedy and hierarchical algorithm can produce overlapping clusters. Easily result in local optimization.
Steps of K-means -
1. Select dataset
2. Select number of mean manually
3. Find distance of every point of dataset from that mean by using Euclidian distance and form groups (clusters).
4. Again find centre of cluster
5. Repeat 3 and 4 until convergence has occurred.

B. Genetic Algorithm
Genetic algorithms are stochastic-based search techniques that comprise a population of individuals, where each individual encodes a candidate solution in a chromosome [6].

Genetic Algorithm is search algorithms based on the mechanics of natural selection and natural genetics. They combine survival of fittest among string structure [8]. An initial population of individuals is generated at random or heuristically.

The chromosome is a collection of genes where genes can generally be represented by different methods like binary encoding, value encoding, permutation encoding and tree encoding. In each generation, the population is evaluated using fitness function.

Next comes the selection process, where in the high fitness chromosomes are used to eliminate low fitness chromosomes. The selection alone does not produce any new individuals into the population.

Hence selection is followed by crossover and mutation operations. The new population generated undergoes the further selection, crossover and mutation till the termination criterion is not satisfied. Convergence of the genetic algorithm depends on the various criterions like fitness value achieved or number of generations [2].

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

Working of GA:

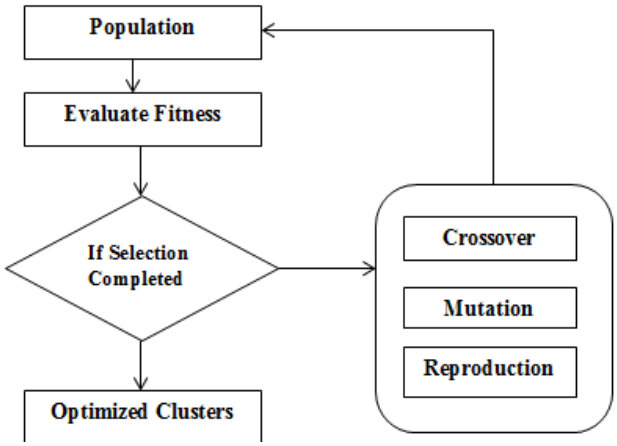

Figure 1 Genetic Algorithm

III. CLUSTERING OF BIOMEDICL ABSTRACTS

The three sub-problems addressed by the clustering process are
(i) Defining a similarity measure to judge the similarity (or distance) between different elements
(ii) Implementing an efficient algorithm to discover the clusters of most similar elements in an unsupervised way and
(iii) Derive a description that can characterize the elements of a cluster in a succinct manner [10].

The proposed system will utilize the capability of GA to form the clusters of Biomedical Abstracts. The main objective of our system is to find potential and hidden knowledge from repositories for decision making. Document clustering of Biomedical abstracts using Genetic Algorithm. Biomedical abstracts will be grouped into different domains such as Cancer, malaria and MTB etc according to similarity of keywords using algorithm. When given a set of running text document objects from a biomedical abstracts, we call Disease based clustering the problem of partitioning the document objects into disjoint sets of Diseases. The system forms cluster on the basis of keywords.

Procedure for grouping of keywords:
i. Identify keywords from different domains.
ii. Keywords are arranged in different groups according to similarity.
iii. Group centre is formed.
iv. Process is repeated until optimized clusters are formed.
v. Analyzing clusters to extract hidden knowledge and decision making.

Proposed system should be able to scan and store the information present on the document. In the existing scenario biomedical researcher has to read whole text document. So it is a time consuming task. In proposed system researcher's effort are reduced in such a way that system automatically processes biomedical abstracts and extract the hidden knowledge.

IV. ARCHITECTURE OF PROPOSED SYSTEM

Proposed system contains set of biomedical abstracts as input. The documents are clustered based on similarity of keywords in to different domains. Genetic Algorithms perform the same operations on the population of possible targets with only those that fit the solution better surviving.
Even though there is no formal definition of GAs, all of them consist of four elements. The first is the population of chromosomes which represent the possible solutions of the problem.
Selection is the second element and it refers to the part of the population that will evolve to the next generation. Selection is performed based on a fitness function. The selection process is applied to each generation produced.
Crossover refers to the combination or exchange of characteristics between two members of the group defined by selection, by which offspring is produced. Then mutation is performed. The process repeats until optimized clusters are formed. [7]

The result of genetic algorithm is useful for researcher. The researcher can easily extract the knowledge by searching. This searching operation is efficient, because of we have already used Genetic Clustering algorithm. So, the proposed work aim towards reducing the efforts and time of biomedical researcher, by extracting useful knowledge automatically.
There are users in this system administrator and researcher. The administrator has access to add document, delete document, preprocess document and perform clustering on given input data set using genetic algorithm. The researcher can search through this set of clusters. The following figure shows system architecture.

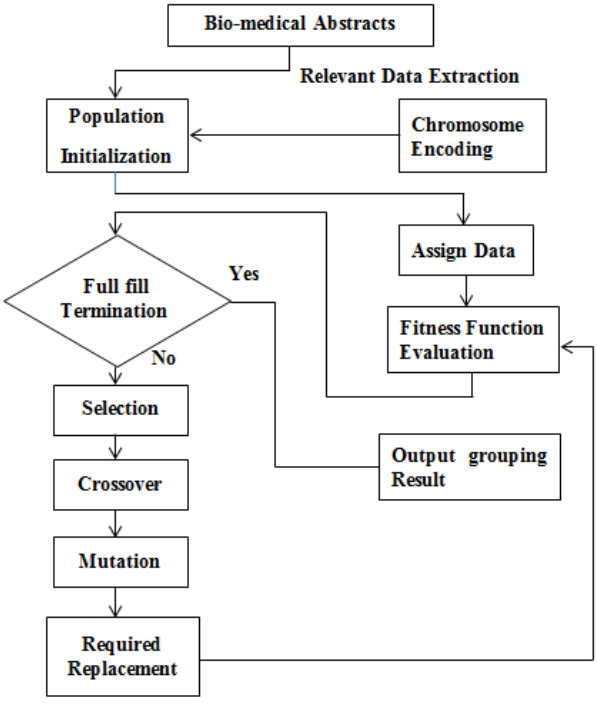

Figure 2: Architecture of system

DOI 10.17148/IJARCCE.2016.55243

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

V. EXPERIMENTAL RESULTS

The accuracy of GA can reach over 98% and generate better clustering results than K-Means [1] Experiments results will be examined with biomedical document. Optimized clusters are formed and result is compared with K-means algorithm. Articles will be grouped in different domains such as cancer, MTB, malaria etc diseases.

Our proposed system, gives analysis of biomedical abstracts as follows

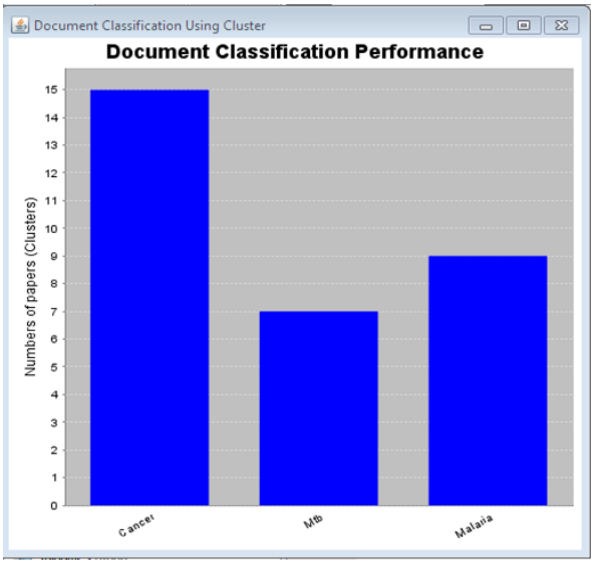

Figure 3: Performance Analysis

The above graph shows the performance of our proposed system. The graph represents number of clusters on x-axis and number of papers in each cluster on y-axis. In this proposed work we have used only three clusters cancer, MTB, Malaria etc.

VI. CONCLUSION

In this proposed work, we searched a solution for organization which is a need due to growing vast amount textual information. We used Genetic clustering algorithms in order to cluster documents. Biomedical disease abstracts are used as a dataset. The novelty of the approach is disease related terms are used as integrating resource for categorizing the retrieved abstracts.

K-means is a popular method for text documents clustering but its results are based on choice of cluster centers so it easily results in local optimization. So for optimized results we can use Genetic Algorithm. [3]

K-means is used for local optimization. Genetic Algorithm is used for global optimization. Genetic Algorithm can generate better results than k-means. More optimized clusters are formed using Genetic Algorithm. It is Hybrid Model and it can do automatic clustering. It is used for Better searching techniques. The results presented in this article open a number of issues for future investigation.

REFERENCES

[1] Application of Genetic Algorithm in Document Clustering Wei Jian-Xiang, Liu Huai, Sun Yue-hong, Su Xin-Ning, IEEE Computer Society Washington, DC, USA ©2009
[2] Application of Genetic Algorithm Optimized Neural Network Connection Weights For Medical Diagnosis of Pima Indians Diabetes Asha Gowda Karegowda1, A.S. Manjunath2, M.A. Jayaram3 1,3Dept. Of Master of Computer Applications, Siddaganga Institute of Technology, Tumkur, India International Journal Of Computer Applications (0975 – 8887) Volume 43– No.1, April 2012
[3] Text Documents Clustering using Genetic Algorithm and Discrete Differential evolution Yogesh Kumar Meena Hindustan Institute of Technology and Management Department of Information Technology, Agra, India Shashank Hindustan Institute of Technology and Management Department of Computer Science, Agra, India Vibhav Prakash Singh Anand Engineering College, Department of Information Technology, Agra, India
[4] Textual Article Clustering in Newspaper Pages Marco Aiello & Andrea Pegoretti Dep. of Information and Communication Technologies
[5] Genetic Algorithm for Document Clustering with Simultaneous and Ranked Mutation K. Premalatha (Corresponding Author) Kongu Engineering College Perundurai, Erode, TN, India
[6] Applying Genetic Algorithms to Decision Making in Autonomic Computing Systems Andres J. Ramirez, David B. Knoester, Betty
[7] Clustering Of News Articles to Extract Hidden Knowledge
[8] An Efficient Document Clustering by Optimization Technique for Cluster Optimality A.K.Santra Dean, Care School of computer Applications, C. Josephine Christy Research Scholar, Bharathiar University Coimbatore-638401.