iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** XII **Month of publication:** December 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue XII December 2017- Available at www.ijraset.com

A Novel Biomedical Knowledge Base for Genomic and Proteomic Analysis using Graph Clustering and Collaborative Filtering

K. Venkatasubramanian[1], Dr. S.K.Srivatsa[2], Dr. C. Parthasarathy[3]
[1]*Research Scholar, SCSVMV University, Kanchipuram*
[1]*Associate Professor, Ramachandra College of Engineering, Eluru*
[2]*Retd. Prof. Anna University, Chennai - 602 025.*
[3]*Assistant Professor, Dept. of IT, SCSVMV University Kanchipuram*

Abstract: ***Gene analysis has a huge scope in identifying the genetic disorders early and preforms the respective diagnosis. Gene regulatory modules micro RNA (miRNA) and transcription factor (TF) play a very important role in gene regulation. Clustering is a main challenge in gene analysis. This reflects huge impact on genetic field. Thus in existing system the multiple genomic and proteomic analysis are scattered in multiple distributed systems. In our proposed architecture, we try to develop a common knowledge base for genomic and proteomic analysis using graph clustering, collaborative filtering (CF) and depth first search (DFS). Clustering is used to group the genes and regulatory modules for each and every gene expressions. Finally the challenge of deriving taxonomy for a particular gene id is resolved using Bayesian Rose Tree (BRT).***
Keywords: Gene Ontology, Regulatory Modules, Graph Clustering, Collaborative Filtering, Depth First Search, Bayesian Rose Tree

I. INTRODUCTION

Usage of computer and technology, huge amount of medical data's creates a huge scope of data mining techniques. Data mining techniques are widely used and popular among the medical research groups. Data mining technique are applied to obtain the solution from large amount of knowledge base, relationship / association among the variables, predict a specific disease based on historical datasets, assigning weightage to the variable etc. The objective of this research article is to develop a common knowledge base for genomic and protein patterns to identify the genetic disorders invoking regulatory modules as well. Also integration of collaborative filtering, graph based clustering, depth first search and Bayesian rose tree representation would provide an efficient and easy solution for representing the gene terms and identifying the associated diseases for a particular gene ID.

Increasing huge amount of bimolecular valuable data and information in life sciences there is large scope for gene analysis. The DNA compromises of gene and proteins. Gene analysis focus on identifying the association between the biomolecular entities. Thus in existing system the multiple genomic and proteomic analysis are scattered in multiple distributed systems. In our proposed architecture, we try to develop a common knowledge base for genomic and proteomic analysis, which can be accessed by doctors, scientists, researchers and others to provide solution for more genetic disorders. Thus analysis of gene and protein data provides vital opportunity for bioinformatics domain which produces biologically meaningful data and solutions. To systematize the knowledge base ontology methods and techniques are used to handle the gene terms.

Our proposed architecture help in understanding complex biological patterns and associations. For grouping of same gene information for a particular gene disorder graph clustering (for clustering regulatory modules like miRNA, Transcription Factor (TF) and gene), Collaborative filtering and depth first search (for gene ontology - Molecular Function (MF), Biological Process (BP), and Cellular Component (CC)) approachesare been used. Clustering is defined as a group of similar data elements or data elements somewhere interconnected. Graphs are nothing but structures, combination of set of vertices and set of edges. Graph clustering is defined as identifying and grouping the vertices from the input graph into clusters. Graph clustering technique is quite popular and widely used for data clustering. Graph clustering is used to identify the associations and cluster the gene regulatory modules miRNA, TF and gene for the respective gene ID. Graph is usually defined as vertices and weighted edges.

Gene ontology is a collaborative process of work to identify the need and descriptions about the gene products along its databases. The process of collaboration started with three model organisms. The GO process has grown on a high rate by incorporating many databases, which includes world's major repositories for plant, animal, and microbial genomes. There are separate aspects for

maintenance of all the gene products. Every structure of the gene products is assigned with separate properties. The scope of GO is that protein- protein interactions. And also anatomical features above the level of cellular components include all cell types. There are different types of Ontology to elucidate different fields. GO is the largely used frameworks around the world.

There are three main Go sub ontology's and they are Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). Each ontology has separate storing and organizing biological concepts called the GO Terms which are mainly used for describing the functions. There is separate code for every term of Gene and also a textual description for the same. The biological process of the Gene are performed through a process known as annotation which are stored in databases which is known as the Gene Ontology Annotation database (GOA). GO evaluates the annotation consistency to avoid the inconsistent of the process. In this project weighted-association rule mining was introduced to identify the gene associations. Thus our proposed system with mining and identifying the associations for a particular gene disorder among large set of gene ID would provide us an effective knowledge base for genetic disorders. Also proposing a integration or fusion technique of both gene ontology and regulatory modules would provide more accurate results. The microRNA (miRNA), gene and transcription factor (TF) are considered for gene regulation. In this for a particular genetic disorder the relationship between miRNA and TF are identified. Finally the results are represented as a tree integrating all the diseases associated with a particular gene ID using Bayesian Rose Tree (BRT).

II. RELATED STUDY

1) Giuseppe Agapito, Mario Cannataro, Pietro Hiram Guzzi and Marianna Milano, Extracting Cross-Ontology Weighted Association Rules from Gene Ontology Annotations paper introduces Gene Ontology-based Weighted Association Rules to analyse the weighted association among the gene annotated datasets. This approach describes about the classes of formulation namely extrinsic and intrinsic methods. Thus this proposed approach was able to extract the exact association rules without compromising accuracy and confidence.

2) Marco Masseroli, ArifCanakoglu, and Stefano Ceri, Integration and Querying of Genomic and Proteomic Semantic Annotations for Biomedical Knowledge Extraction, this research article briefs the scope of handling the complex biomedical information's comprising of genomic and proteomic annotations. Also these information's are distributed in various data sources, which leads to complex integration or accessibility challenges. Hence to develop and maintain a common Genomic and Proteomic Knowledge Base (GPKB) is very important for research, scientists and doctors to extract the solutions and provide useful knowledge information's regarding genetic analysis.

3) K. Venkatasubramanian, Dr.S.K.Srivatsa and Dr. C. Parthasarathy, A Graph theory algorithmic approach to data clustering and its Application, this research paper introduces notion of a dominant set of vertices in an edge-weighted graph. This paper also address the issues in deriving taxonomy using clustering techniques. In this paper clustering application in image segmentation, object recognition, and information retrieval are been explained.

4) R. Priscilla, C.N Prashantha and S. Swamynathan, Analysis of Gene expression data using MATLAB Software, this article elaborates and compares the clustering and bi-clustering algorithms for grouping the genes to identify the gene expression patterns. Clustering methods are so important to group the gene into clusters irrespective of genetic ID, taxonomy ID, family ID, treatments, tissue etc.

5) Satu Elisa Schaeffer, Graph clustering, Elsevier, this survey article explains about the graph clustering method. Graph clustering technique is quite popular and widely used for data clustering. Graphs are nothing but structures, combination of set of vertices and set of edges. Graph clustering is defined as identifying and grouping the vertices from the input graph into clusters.

6) K. Venkatasubramanian, Dr.S.K.Srivatsa and Dr. C. Parthasarathy, Graph–Theoretic Clustering for Image Grouping and Retrieval this research article briefs about the graph based clustering. It states that there are 2 types of clustering approaches in graph theoretic approach namely hierarchical and partitional clustering approaches. Minimal spanning trees (MST) is likely associated with graph based clustering method. MST advantage is to cluster groups of irregular boundaries. Basically graph clustering analyses the edges between two vertices with their weights. MST based algorithm identifies the differences between two subgraphs and within a subgraph.

7) Yangqiu Song, Shixia Liu, Xueqing Liu, and Haixun Wang, Automatic Taxonomy Construction from Keywords via Scalable Bayesian Rose Trees this paper studies the important problem of representing taxonomy from a set of keywords. This paper proposed a Bayesian rose tree (BRT) based approach to develop a hierarchical taxonomy for the input keywords. Thus this eliminates the complexity of the existing techniques and helps is identifying a specific taxonomy from one million keywords in an hour. Thus BRT is very simple to be implemented and provides easy understanding for the readers.

8) Pedro F. Felzenszwalb and Daniel P. Huttenlocher, Efficient Graph-Based Image Segmentation, this paper explains the challenge of segmenting an image into regions. In this paper graph based representation is been used for image segmentation. This approach constructs a graph using two different neighbourhood edges. Thus this paper provides solution for image segmentation and grouping.

9) ShwetaKharya, Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease, this paper has discuss the various data mining approaches to identify and provide diagnosis for breast cancer. This paper identifies the best predictor accuracy among all data mining techniques and states decision tree data mining technique provides 93.62% accuracy. Automatic decision tree based prediction with machine learning would be useful tool for medical research groups for predicting cancers. This papers also illustrates the scope of data mining technique in medical domain. Analysing large volume of medical data, prediction of disease analysing the historical data, association among the patterns would lead to huge demand for data mining techniques.

10) Huseyin Kaya and S. GündüzÖğüdücu, A new approach for mutation analysis using data mining techniques this proposed an efficient diagnostic method of genetic disorders to identify the mutations in the DNA sequence. This approach uses chromatograms, which mines for the possible mutations among the unknown knowledge base with regard to the reference chromatograms. Thus many research articles majorly focus on diagnostic solution for genetic disorder not concerned about the genomic and protein sequences.

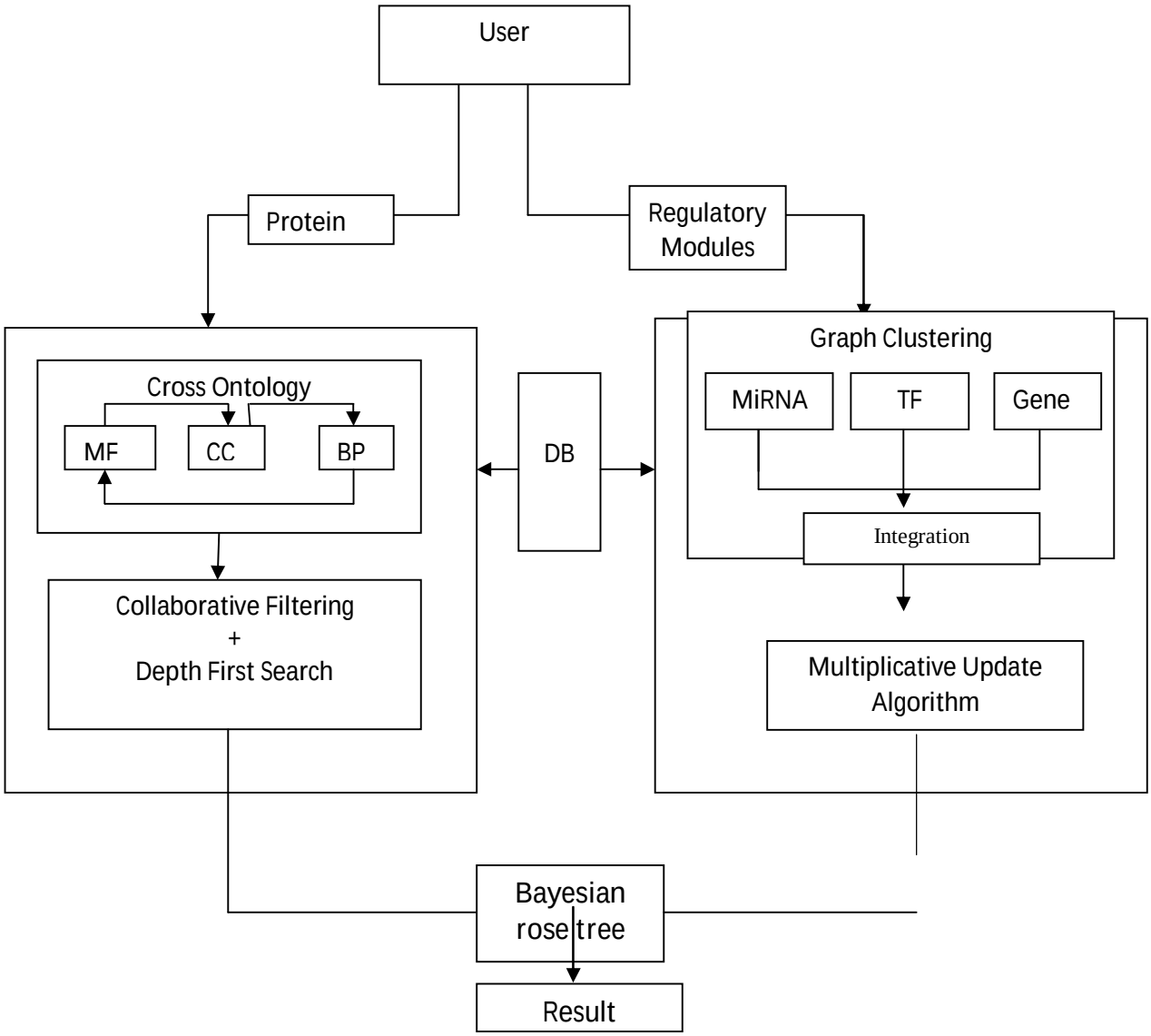
11) JiaweiLuo, Gen Xiang and Chu Pan, Discovery of microRNAs and transcription factors co-regulatory modules by integrating multiple types of genomic data, explains about the regulatory modules for a particular gene id. Thus regulatory modules play a very important part in gene regulation. The regulatory modules are microRNA (miRNA) and transcription factor (TF). However it's a big challenge to identify the relationship between miRNAs and TFs.

Gene id	MiRna	TF	Gene	Weighted Confidence	Cross ontology category
64324	MI0000060	chr11	NSD1	0.50	BP-CC-MF
1028	MI0000061	chr22	CDKN1C	0.35	BP-CC-MF
105259599	MI0000062	chr22	H19-ICR	0.74	BP-CC-MF
100506658	MI0000064	chr21	OCLN	2.00	BP-CC-MF
55630	MI0000065	chr9	SLC39A4	0.64	BP-CC-MF
1130	MI0000066	chr19	LYST	0.75	BP-CC-MF
2517	MI0000067	chr9	FUCA1	0.89	BP-CC-MF
2629	MI0000068	chrX	GBA	0.67	BP-CC-MF
4688	MI0000263	chr12	NCF2	0.84	BP-CC-MF
2720	MI0000265	chr19	GLB1	0.98	BP-CC-MF

Table 1 shows miRNA, TF, gene, weighted confidence and cross ontology for few gene id's.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue XII December 2017- Available at www.ijraset.com

III. SYSTEM ARCHITECTURE



A. Gene Ontology

Gene Ontology is a widely used framework for the model of biology. The GO defines the concepts used to describe gene function and its relationships between these concepts.Gene Ontology includes three main sub-ontologies namely Biological Process, Molecular Function, and Cellular Component. Each ontology stores and organizes biological concepts, called GO Terms, used for describing functions, processes and localization of biological molecules. Different ontologies are been proposed and used to analyze different type of fields. The gene ontology functions are classified into three aspects, they are molecular function of gene products, molecular activities of gene products and cellular component.

The association or relationship between various GO terms and biological concepts are processed using annotation technique. Annotations for each genetic disorders are stored in the common database, thus this database is termed as Gene Ontology Annotation (GOA) database. Thus this proposed system is a collaborative effort to obtain consistent descriptions of genes in various database or sources. For this approach, Gene Ontology-based Weighted Association Rules Mining (GO-WAR) is proposed to extract the association rules among the gene id's with high level information content.

The information content (IC) can be defined into two classes namely extrinsic and intrinsic techniques. The extrinsic information content involves the annotation data and whereas intrinsic information content involves structural information extracted from the GO terms. The gene ontology results are transmitted to collaborative filtering and depth first search techniques for providing solution for the genetic disorders.

For providing accurate solutions, our proposed system provides a fusion of both Gene ontology approach and gene regulatory modules. Thus the fusion of both gene ontology and gene regulatory modules provides tremendous accurate solution for genetic disorders. The co-regulation study between microRNA (miRNA) and transcription factor (TF) has become an important issue recently. In our proposed system, we identity and analyse miRNA and TF for genetic disorders by integrating various types of genomic data. Graph clustering technique is used to obtain the associations between miRNA and TF for a particular gene id.

B. Collaborative Filtering

Each gene ID refers a particular diseases. Collaborative filtering is applied to the results obtained using Gene Ontology-based Weighted Association Rules Mining algorithm for optimization. Accuracy is been increased by applying Collaborative filtering technique. The information contents (extrinsic and intrinsic) and gene ontology results are been compared to obtain more accurate recommendation of gene to the patients who has genetic disorders.

C. Depth First Search

Usually while processing the graph it is complex to visit the entire chain, sometimes it's infinite. In our proposed system we proposed DFS algorithm to perform a depth search for a limited resources.

When DFS is performed to a limited depth the time is still linear in terms to number of expanded vertices. Edges although this number is not the same as the capacity of the entire graph because some vertices may be searched more than once and others not at all but the space complexity of this variant of DFS is only proportional to the depth limit. As a result this requires much smaller than the space needed for searching to the same depth using BFS.

Depth-first search applies DFS repeatedly with a sequence of increasing limits. In artificial intelligence mode of analysis, with a branching factor is greater than one factor. Iterative deepening increases the running time by only the same factor over the case in which the correct depth limit is known due to the geometric growth of the number of nodes per level. DFS may also be used to collect a sample of graph nodes. Mostly incomplete DFS similarly to incomplete BFS is biased towards nodes of high degree.

Depth first search (DFS) is an algorithm for searching tree or graph based data structure. DFS implemented for searching the most affected disease for a given gene ID. While using DFS we can able to identify the entire diseases for a particular gene ID's. DFS provides more efficient search results. The diseases are searched on the basis of chronological order. By using DFS search algorithm we can able to figure all associated diseases associated with each and every gene id.

DFS is an iterative approach. The DFS implementation are as follows

```
DFS(G, u)
u.visited = true
for each v ∈G.Adj[u]
ifv.visited == false
DFS(G,v)
init () {
For each u ∈ G
u.visited = false
For each u ∈ G
DFS(G, u)
}
```

D. Graph Clustering

Graph clustering is used to identify the associations and cluster the gene regulatory modules miRNA, TF and gene for the respective gene ID. Because of the high dimensionality problem, k-nn algorithm was used for clustering. When compared with k-nn, graph clustering provides more accuracy. Graph is usually defined as vertices and weighted edges. Graph clustering is defined as, a cluster with a set of alike entities and a cluster with a set of non-alike entities. Cluster should able to satisfy two basic conditions, 1. The cluster should contain high internal homogeneity. Homogeneity means being same. 2. The cluster should contain high in homogeneity between entities inside and outside the clusters. The output of the entities are projected as edge weighted graph where the weights on the edges inside the cluster must be large and weights on the edges connecting the cluster nodes to the external nodes must be small. Finally the extracted regulatory modules clusters should be close to the input gene id. Thus cluster with highest entities would be obtained. Sometimes, many clusters with highest entities are obtained, in that particular case the highest total weight of the edges in those clusters is been selected.

Also understanding the behavior of the features is important. This helps to understand the effectiveness of both the features and the distance between similar entities.

For example, Graph with three clusters has been chosen. All cluster members are drawn in black and their internal edges are drawn thicker than other edges of the graph. The cluster on the left is of good quality and dense. The one in the middle has the same number of internal edges, but many more to outside vertices making it a worse cluster. The cluster on the right has very few connections outside, but lacks internal density and hence is not a good cluster.

1.1 Clustered graph

For individual N matches a query is been processed and the best N matches are obtained. S is the set consisting of images which are been obtained by processing the query. Thus we can able to construct a graph with the entities in S and draw edges between the query and retrieved entities. We term these edges as set R, where R = {(i,j) ∈ S × S (j is the retrieved entities when i query is been processed) [12].

The neighborhood of X is termed as Neighborhood(X) = {Y | (X, Y) ∈ R}.

Conditional density D(Y |X)is the number of nodes available in neighborhood of X which have Y as a neighbor, D(Y |X) = #{N ∈ S | (N, Y) ∈ R and (X, N) ∈ R}.

Given an integer K, a dense region Z around a node X ∈ S is defined as

Z(X, K) = {Y ∈ S | D(Y |X) ≥ K}.

Thus clustering reduces the search for a complete subgraph which is named as clique.

Z(X) = Z(X, J) is a dense region candidate around X where J = max{K | #Z(X, K) ≥ K} because if M is amajorclique of size L ,then X, Y∈ M implies that D(Y |X) ≥ L.

Thus M ⊆ Z(X, L) and K ≤ L ≤ #Z(X, K).

Association of anode X to asubset B of S is defined as

$$A(X|B) = \#\{Neighborhood(X) \cap B\} / \#B \quad (13)$$

Where 0 ≤ A(X|B) ≤ 1.

Compactness of a subset B of S is defined as

$$C(B) = 1 / \#B \sum_{X \in B} A(X|B) \quad (14)$$

Where 0 ≤ C(B) ≤ 1.

E. Multiplicative Algorithm

Multiplicative algorithm is used for optimisation. An objective function is initiated to after integrate the regulatory modules and gene ontology. After integration, optimal solution should be obtained. Hence to resolve the optimization concern, in this proposed system we propose multiplicative updating algorithm.

F. Bayesian Rose Tree (Brt)

For the final representation of results, we have used BRT based representation. Identifying and deriving a taxonomy from a keyword is always a challenge. BRT provides a rich knowledge when compared with the existing techniques like domain specific or binary clustering and hierarchical clustering. Also BRT is very simple and easy to understand [15]. Thus BRT can predominately reduce the complexity of clustering algorithms.

IV. EXPERIMENTAL RESULTS

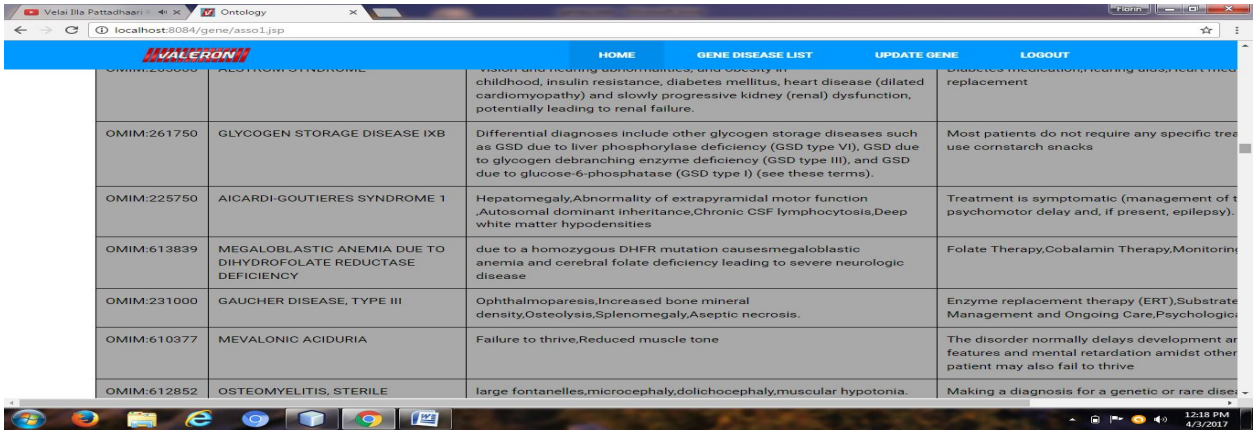


Figure1. Gene Ontology

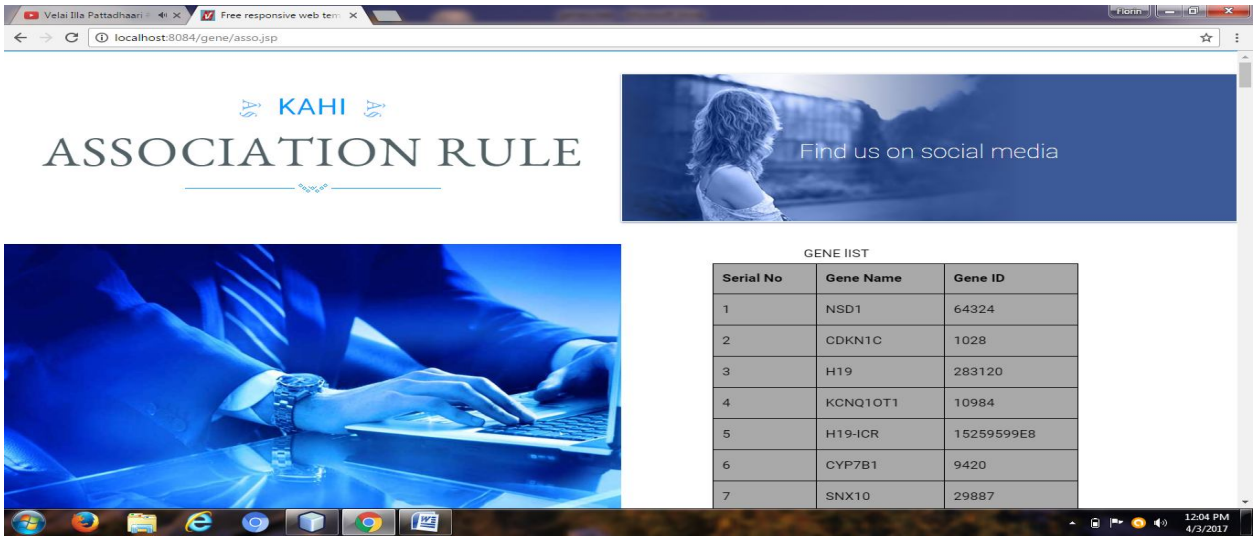


Figure2. Depth First Search

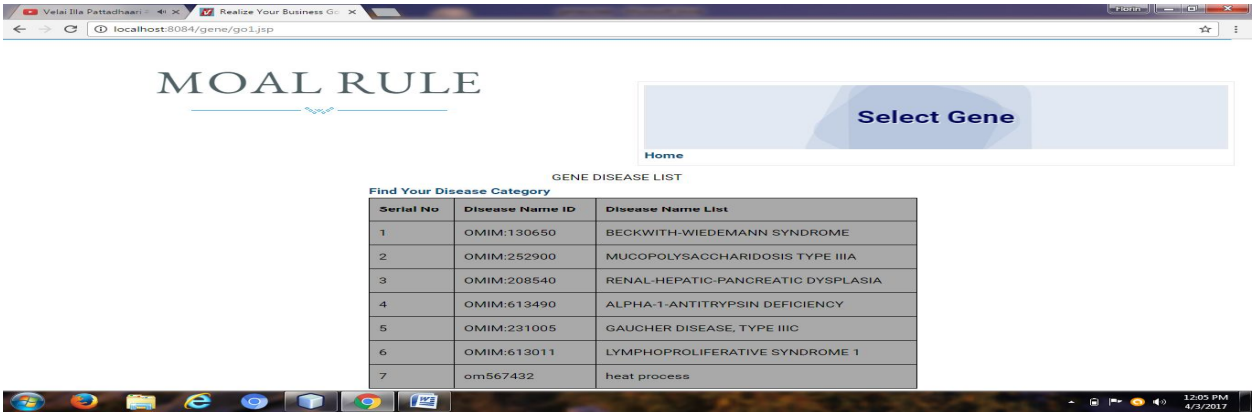


Figure3. Collaborative Filtering

Figure4. Multiplicative algorithm

Figure5. Graph Clustering

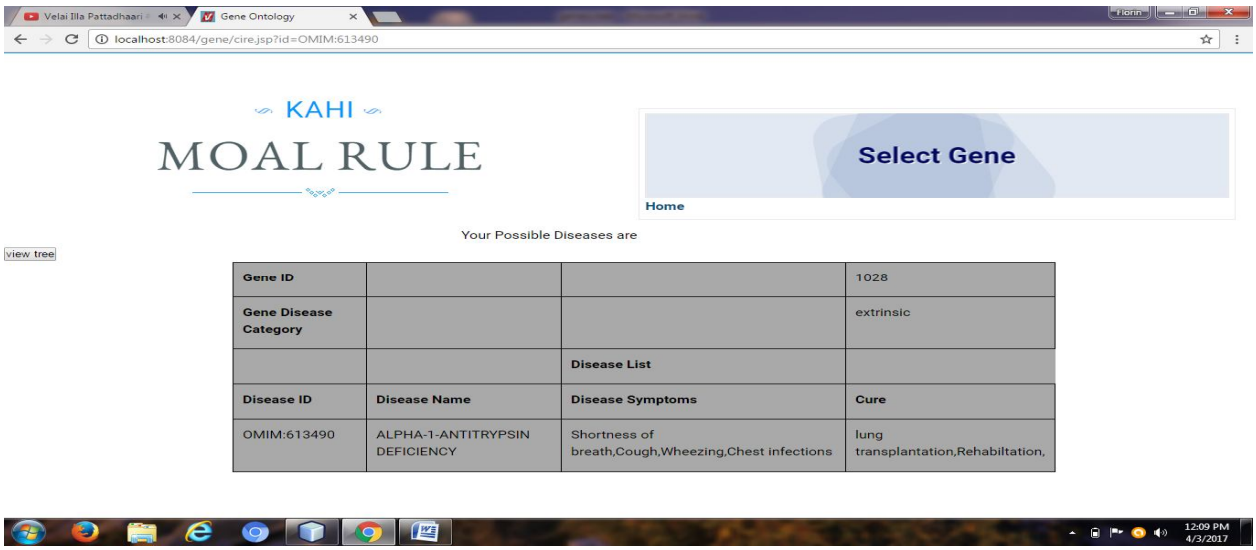


Figure6. Graph Clustering module

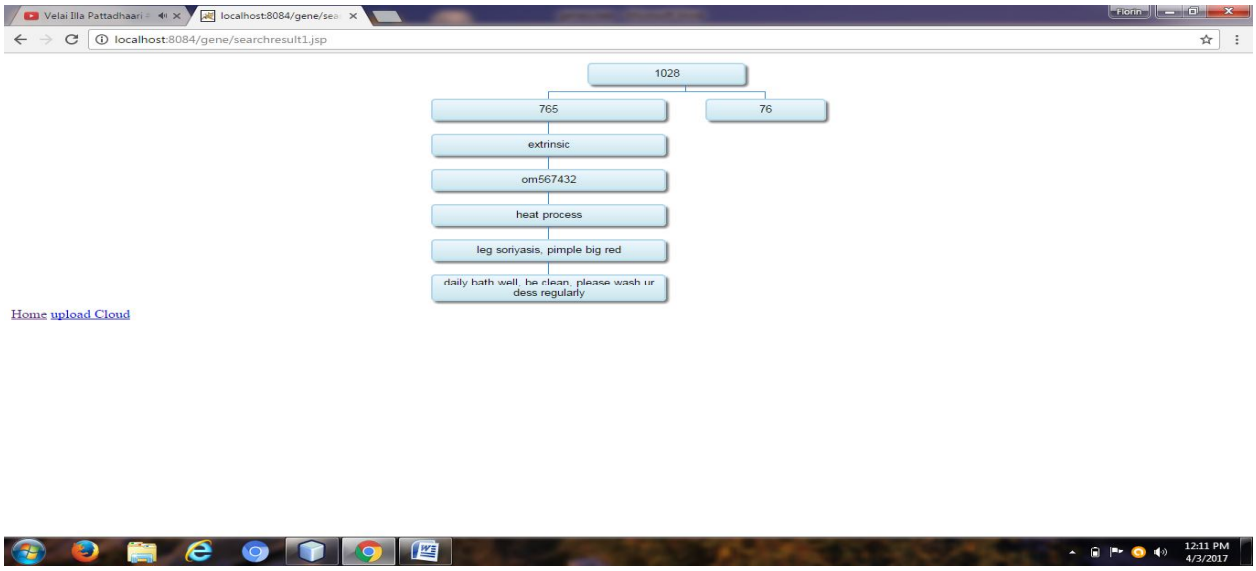


Figure7. Bayesian Rose Tree

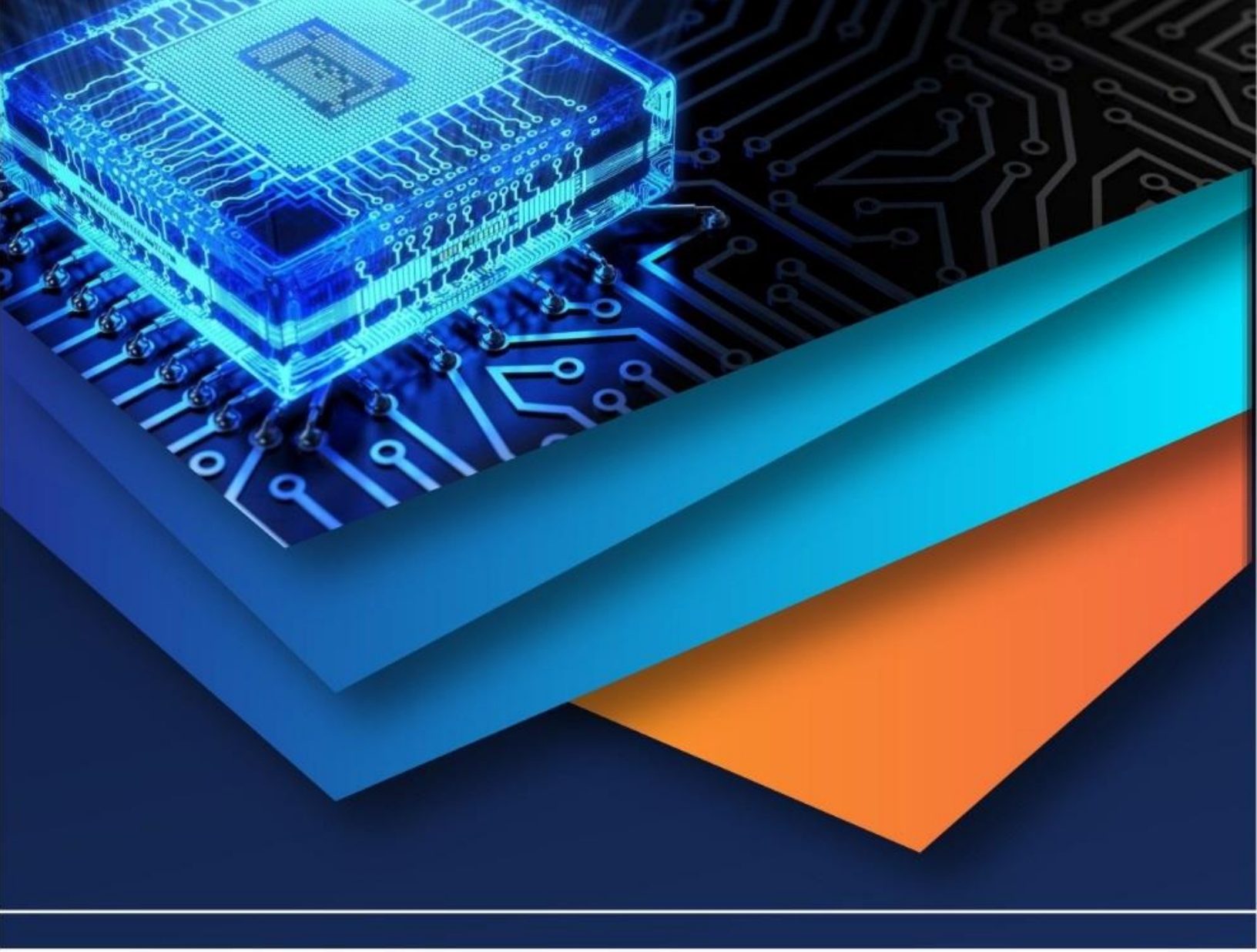
V. CONCLUSION

Recent progress in biotechnology created a huge scope for data mining and clustering techniques. Fusion architecture of both gene ontology and gene regulatory modules may provide useful and complex information for a particular gene which leads to provide efficient and accurate diagnosis. But this is of huge challenge.

Our proposed work results found satisfactory in identifying and analysing complex biological structures using graph clustering, collaborative filtering and depth first search. The intrinsic and extrinsic values are also calculated during gene ID analysis. The fusion results and its taxonomy are represented using BRT. To our knowledge, gene ontology and regulatory modules integration and addressing complex queries are not implemented and available in the existing system.

REFERENCES

[1] Giuseppe Agapito, Mario Cannataro, Pietro Hiram Guzzi and Marianna Milano, Extracting Cross-OntologWeighted Association Rules from Gene Ontology Annotations, ieee, 2016.
[2] Marco Masseroli, ArifCanakoglu, and Stefano Ceri, Integration and Querying of Genomic and Proteomic Semantic Annotations for Biomedical Knowledge Extraction, ieee 2016
[3] K. Venkatasubramanian, Dr.S.K.Srivatsa and Dr. C. Parthasarathy, A Graph theory algorithmic approach to data clustering and its Application, IJSEAT, vol 3, issue 9, 2015.Satu Elisa Schaeffer, Graph clustering, Elsevier, 2007
[4] K. Venkatasubramanian, Dr.S.K.Srivatsa and Dr. C. Parthasarathy, Graph–Theoretic Clustering for Image Grouping and Retrieval, 2017
[5] Yangqiu Song, Shixia Liu, Xueqing Liu, and Haixun Wang, Automatic Taxonomy Construction from Keywords via
[6] Pedro F. Felzenszwalb and Daniel P. Huttenlocher, Efficient Graph-Based Image Segmentation, 2010.
[7] ShwetaKharya, Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease, IJCSEIT, Vol.2, No.2, April 2012.
[8] Huseyin Kaya and S. GündüzÖğüdücu, A new approach for mutation analysis using data mining techniques, IEEE, 2010.
[9] JiaweiLuo, Gen Xiang and Chu Pan, Discovery of microRNAs and transcription factors co-regulatory modules by integrating multiple types of genomic data, IEEE 2017.
[10] SelimAksoy and Robert M. Haralick, Graph–TheoreticClusteringforImageGroupingandRetrieval
[11] S. Aksoy. Textural features for content-based image database retrieval. Master's thesis, University of Washington, Seattle, WA, June 1998. Online: http://isl.ee.washington.edu/aksoy/thesis.shtml.
[12] S. Aksoy and R. M. Haralick. Textural features for image database retrieval. In Proceedings of IEEE Workshop on Content-Based Access of Image and Video Libraries, in conjunction with CVPR'98, pages 45–49, Santa Barbara,
[13] Yangqiu Song, Shixia Liu, Xueqing Liu and Haixun Wang, Automatic Taxonomy Construction from Keywords via Scalable Bayesian Rose Trees, IEEE 2015.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY