Keyword Based Association Analysis For The Identification Gene Disease Association

K. Santhosh Kumar[1], Dr. P. Sudhakar[2]
[1, 2] Department of Computer Science & Engineering
[1, 2] Annamalai University, Annamalai Nagar, India.

Abstract- *In the field of bio-informatics the determination of gene disease association becomes very essential for understanding the relation between the gene disease mechanisms. This understanding will help physicians to better cure the disease. The gene disease associations are mined from the various medial documents using various association mining algorithms and methods. In this paper keyword based association analysis is carried out. By this approach, keywords or the frequently occurring terms is determined first and then association analysis is performed. For performing association analysis as a preprocessing step the following operations such as parsing, stemming and the removal of stop words are implemented. In a document database each document is viewed as a transaction. The set of keywords may form a phrase and the association mining process helps us to determine the compound association between the keywords. The keywords here represent the names of the genes, disease and the association between them.*

Keywords- Bio-informatics, association analysis, stemming, compound association

I. INTRODUCTION

In this digital world due to the development of internet technologies the growth of data available in the data warehouses, structured data base management systems and social media is rapid. As the availability of the data is high, discovering pattern and meaning information becomes an essential task [1]. Thus mining the data using the knowledge discovery algorithms and methods has now become a challenging area for researchers. The knowledge discovery process utilizes the techniques from the following field such as machine learning, NLP, document retrieval, information management and data mining. The main objective of all these processes is to extract information from the large text documents. Depending upon the type and usage of the application the final output of the mining process varies. Association rule mining is an important process in various text mining applications. This highlights the similarity/ association/ correlation between the keywords in text [2]. The mined association rules can be interpreted easily by an analyst. This paper focuses on identifying the association between the extracted keywords (gene and disease names) from the medical documents and texts. The main motivation behind this work is that the digitized medical documents if analyzed using these mining techniques will help the physicians to better understand the nature, cause and cure of the disease. Rather than using the conventional approaches for the text mining new and novel techniques are required for knowledge acquisition.

In the post genome era the analysis about the role of genes in the human diseases is a hot area of research. As the biological data is growing day-by-day, the knowledge extraction about the gene-disease association has to be expanded to a greater rate. Identifying the genes that causes a particular disease consumes a huge manual work when the identification is carried out using conventional methods. Therefore, an automatic method to detect the association of a genome sequence with a particular disease is very essential [3]. The Online Mendelian Inheritance in Man (OMIM) [4] is a database collected manually from various medical literature. Certain common diseases in human occurs because of the presence of specific genetic characteristics. The problem faced by the existing conventional methods for text mining is the documents are not well structured as like the relational or transactional databases. The unstructured nature of the text documents makes it very complex to process them. As a text can be expressed in many ways in a language, it is hard to represent them in an abstract form.

This paper introduces an association rule mining method to find the association between the gene and disease using keyword detection algorithms and other text mining techniques. The work is concentrated towards the detection of the keywords in the medical documents initially. The next the correlation between the identified keywords are identified using association mining methods. The keywords give an abstract and compact representation of a digital text document. Keyword detection process is used to improve the functionality of the association mining system [5]. Keyword detection is carried out in this work using the concept of co-occurrence. The algorithm has a high computational efficiency than other keyword extraction algorithms like TextRank.

II. OVERVIEW OF KEYWORD DETECTION METHOD

To begin with the frequently occurring terms are extracted and the co-occurrence of a term and the frequently occurring terms are counted. If a word or a phrase occurs frequently with a set of terms, it is sure that the phrase has an important meaning and information. Using the χ2-measure the co-occurrence distribution's degree of bias is measured. To identify the importance of a word or a phrase the degree of bias is used in this work. But if the frequency of the word is small then this approach is not reliable. If a word appears *'M'* times and co-occurs only with a term 'N' times and seems to be reliably biased. For evaluating the statistical significance of the biases, the X^2 test is used.

The unconditional probability of a frequent term $g \in G$(G set of frequent terms) is denoted as the expected probability p_g and the total number of co-occurrences of term *w* and frequent terms *G* as n_w. Frequency of co-occurrence of term w and term *g* is written as *freq(w, g).* The statistical value of χ2 is defined as

$$X^2(w) = \sum_{g \in G} \frac{\left(freq(w,g) - n_w p_g\right)^2}{n_w p_g} \quad (2.1)$$

The term $n_w p_g$ represents the expected frequency of co-occurrence; and $(freq(w, g) - n_w p_g)$ represents the difference between observed and expected frequencies. Therefore, large χ2(w) indicates that co-occurrence of term *w* shows strong bias. In general words with large χ2 are relatively important in the document; words with small χ2 are relatively trivial.

A co-occurrence matrix is calculated by counting frequencies of pairwise term co-occurrences. If the total number of terms in the document is 'N' the co-occurrence matrix is of size N × N.In which columns corresponding to frequent terms are extracted for calculation. The remaining columns related to low frequent terms are removed. Because it is difficult to estimate precise probability of occurrence for low frequency terms. To improve extracted keyword quality, it is very important to select the proper set of columns from a co-occurrence matrix. The set of columns is preferably orthogonal; assuming that terms g_1 and g_2 appear together very often, co-occurrence of terms *w* and g_1 might imply the co-occurrence of *w* and g_2. Thus, term w will have a high χ2 value; this is very problematic.

It is straightforward to extract an orthogonal set of columns, however, to prevent the matrix from becoming too sparse, the terms are clustered. Two major approaches for clustering are Similarity-based clustering - If terms w_1 and w_2 have similar distribution of co-occurrence with other terms, w1 and w2 are considered to be in the same cluster. [6]
Pairwise clustering - If terms w_1 and w_2 co-occur frequently, w_1 and w_2 are considered to be in the same cluster. [7]
In the used keyword detectionmethodboth the clustering techniques are used. Initially the similarity terms are clustered using the Jensen-Shannon divergence [8] and then the pairwise clustering is applied using mutual information. When the frequent terms are clustered properly, it results in an appropriate χ2 value for each term. To make the algorithm as simple the size of the cluster is not considered.If the clusters are balanced it might improve the performance of the algorithm.

2.1 Block Diagram of Keyword Detection algorithm

The algorithm starts with a preprocessing step in which a stemming algorithm is used to stem words [9]. In text processing and information retrieval, stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form. Porter algorithm is used to stem words and phrases are extracted using the APriori algorithm [10]. The phrasesof up to 4 words with frequency more than 3 times are extracted and stop words are excluded. Next, select the top frequent terms up to 30% of the number of running terms, N_{total}. Then cluster a pair of terms whose Jensen-Shannon divergence is above the threshold (0.95 × log 2). Cluster a pair of terms whose mutual information is above the threshold (log(2.0)). Mutual information between terms *w1* and *w2* is defined as

$$M(w_1, w_2) = \log \frac{P(w_1, w_2)}{P(w_1)\,P(w_2)} \quad (2.2)$$

Two terms are in the same cluster if they are clustered by either of the two clustering algorithms. The obtained clusters are denoted as *C*. After that the expected probabilityis calculate by finding thecountof the number of terms co-occurring with $c \in C$, denoted as n_c, to yield the expected probability $p_c = n_c/N_{total}$.

Finally the χ2 value is estimated for each term *w,* count co-occurrence frequency with $c \in C$, denoted as *freq(w, c).* Then count the total number of terms in the sentences including *w,* denoted as n_w. Calculate the χ2 value following

$$X^2(w) = \sum_{g \in G} \frac{\left(freq\ (w,c) - n_w p_g\right)^2}{n_w p_g} - \max_{g \in G} \left\{ \frac{(freq\ (w,c) - n_w p_c)^2}{n_w p_c} \right\} \quad (2.3)$$

The set of words having the largest χ2 value is considered as the keywords. The block diagram of the proposed approach is given in the Fig. 1 below.

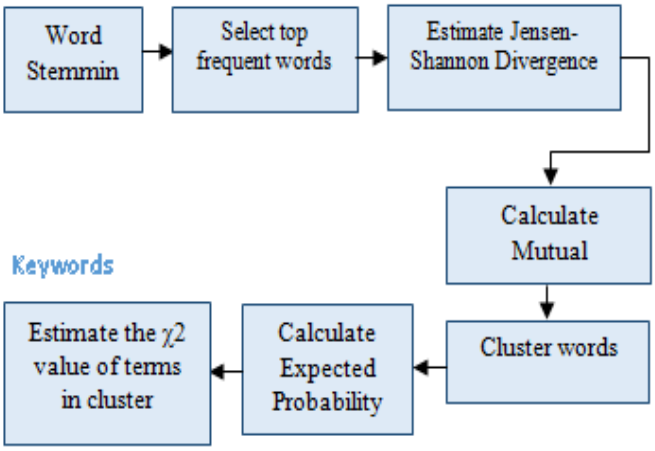

Fig. 1 Block Diagram of Keyword Detection Algorithm

III. GENE DISEASE ASSOCIATION ANALYSIS

After finding the keywords from the medical documents, each sentence containing the keyword in the document is marked as 'Key' and other sentences are marked as 'None'. In order to extract the association between the gene and disease phrase structure parsing is used. It generates parse tree of a sentence which can be analyzed for finding the relation between the gene and disease. The parse tree is analyzed using the depth first search method. Each path between the keywords node and the root node were collected. The paths which contains the disease of interest are used to capture the association. The associations are represented in to form of an adjacency matrix. In the matrix the presence of edge is represented as '1' and if there is no edge between $node_i$and $node_j$ then the A(i,j) is marked as 0. To detect the most related gene disease association, a centrality approach is used. It is used for calculating the correlation of the corresponding gene and disease,represented as node in the parse tree, based on the weights connection node edges in the parse tree.

3.1 Degree centrality

Degree centrality represents central tendency ofeach node in the network, the more directconnects it has, the more power it has in thenetwork and so the more important it is. Thedegree of centrality $C_D(v)$of $node_v$ is calculatedas follows.

$$CD(v) = \sum_{j=1}^{n} A_{ij} \quad (3.1)$$

3.2 Closeness centrality

The closeness centrality of a node calculates the centrality based on how closethe node corresponding node is with other nodes in the tree structure. If the distance of a node to other nodes is smaller, then the closeness centrality will be high. If a node has multiple path to the root, then the shortest path is chosen and the distance is find by calculating the shortest path available in the list.

3.3 Betweenness centrality

The betweenness centrality of a node is estimated from the count of the shortest paths between intermediate nodes that flows through the concerned node. For a node x, this measure is computed by taking the sum of the number of shortest paths between pairs of nodes that pass through node x divided by the total number of shortest paths between pairs of nodes. It characterizes how a particular node can control the flow of information in the network. An intermediate node is considered as a central node if it lies in many paths which connects a pair of node.

Table 1. Highest ranking candidate autism risk genes (rankedaccording to SFARI Gene).

High confidence genes	Strong candidate genes	
ADNP	ANKRD11	KATNAL2
ANK2	BCKDK	KDM5B
ARID1B	BCL11A	KMT2A
ASH1L	CACNA1H	KMT2C
ASXL3	CACNA2D3	MAGEL2
CHD8	CHD2	MED13L
DYRK1A	CNTN4	MET
GRIN2B	CNTNAP2	MSNP1AS
POGZ	CTNND2	MYT1L
PTEN	CUL3	NLGN3
SCN2A	DEAF1	NRXN1
SETD5	DSCAM	PTCHD1
SHANK3	ERBB2IP	RANBP17
SUV420H1	FOXP1	RELN
SYNGAP1	GRIP1	SHANK2
TBR1	ILF2	SLC6A1
	INTS6	SPAST
	IRF2BPL	USP7
	KAT2B	WAC

IV. RESULTS AND DISCUSSION

In our experiments the SFARI gene database [11] is used for analyzing the performance of the proposed algorithm. This database is an comprehensive source of information related to the association between human genes and the autism disease. A set of medical documents related neural development disorders are collected and analyzed by the proposed method. To ascertain the quality of the results produced, they are validated against the information available in the SFARI database. Using each centrality measures a detailed analysis for the top 20 genes found having association with the autism disease is done using the SFARI database i.e. to verify the newly found (inferred) genes, the SFARI database is used. If a gene is not marked by SFARI as being related to Autism disease, we manually searched for articles indexed in PubMed that state that the gene is related to autism and also checked whether the gene appears in the pathway of the parse tree constructed using the prior knowledge on the gene disease association. It is a manually drawn pathway map of the currently known gene interaction and reaction network for Autism disease.

Using the degree centrality among its top 20 ranking genes, 5 genes of the original 15 seed genes are found (ADNP ANK2 ARID1B ASH1L ASXL3). The remaining 15 genes (75% of the top 20 genes) are inferred genes in which we were able to confirm the association of 14 genes (93.33% of the inferred genes) to autism, except for 1 gene: DEAF1. For this exceptional gene, we did not find negative nor positive evidence, which implies that the gene may still potentially be an autism disease causing gene. Using closeness centrality, we found 2 seed genes (ADNP and CHD8) and inferred 18 new genes. A total of 13 of the inferred genes (72.22% of the inferred genes) have evidence, which indicate that they are related to autismand 5 inferred genes (PTEN, SCN2A, SETD5, POGZ and DYRK1A) do not have such affirmative evidence. Betweenness centrality found the most seed genes among the four centrality methods. In its result, we have 7 seed genes and 13 inferred genes, of which 8 inferred genes (61.54% of the inferred genes) are verified to have relation to the disease.

V. CONCLUSION

In this paper a new approach to predict the association between the genes and diseases based on keyword detection and association analysis is presented. Initially from the known genes and the neural disease disease-gene interaction network is constructed manually from the knowledge gained from the medical literature. Then using the centrality measures the genes in the network are ranked according to the relevance to the disease. Then the developed approach is tested with the autism disease and showed that the degree centrality achieved higher rate in the detection of gene-disease association. For testing the approach medical journals related to autism disease are extracted by keywords search and the key sentences are merged with the already constructed parse tree based on the relevance of the genes with the autism disease. By this approach it was able to find the genes associated with the autism disease which are marked in the SFARI gene database. The genes which have a high association with the autism disease is identified by the centrality approach are verified with the SFARI database.

REFERENCES

[1] Ananiadou, Sophia, and John McNaught. Text mining for biology and biomedicine. London: Artech House, 2006.

[2] Moore, Jason H., Folkert W. Asselbergs, and Scott M. Williams. "Bioinformatics challenges for genome-wide association studies." Bioinformatics 26.4 (2010): 445-455.

[3] Risch, Neil. "Implications of multilocus inheritance for gene–disease association studies." Theoretical population biology 60.3 (2001): 215-220.

[4] Hamosh, Ada, et al. "Online Mendelian inheritance in man (OMIM)." Human mutation 15.1 (2000): 57-61.

[5] Hulth, Anette, et al. "Automatic keyword extraction using domain knowledge." International Conference on Intelligent Text Processing and Computational Linguistics. Springer Berlin Heidelberg, 2001.

[6] Zhang, Min, et al. "Discovering relations between named entities from a large raw corpus using tree similarity-based clustering." International Conference on Natural Language Processing. Springer Berlin Heidelberg, 2005.

[7] Fischer, Bernd, Thomas Zöller, and Joachim M. Buhmann. "Path based pairwise data clustering with application to texture segmentation." International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition. Springer Berlin Heidelberg, 2001.

[8] Abnizova, I., et al. "A statistical approach to distinguish between different DNA functional parts." WSEAS

Transactions on Computational Methods 2.4 (2003): 1188-1196.

[9] Ramasubramanian, C., and R. Ramya. "Effective pre-processing activities in text mining using improved porter's stemming algorithm." International Journal of Advanced Research in Computer and Communication Engineering 2.12 (2013): 2278-1021.

[10] Lazcorreta, Enrique, Federico Botella, and Antonio Fernández-Caballero. "Towards personalized recommendation by two-step modified Apriori data mining algorithm." Expert Systems with Applications 35.3 (2008): 1422-1429.

[11] Banerjee-Basu, Sharmila, and Alan Packer. "SFARI Gene: an evolving database for the autism research community." Disease Models and Mechanisms 3.3-4 (2010): 133-135.