Association Rule Based Hypergraph for Stroke data analysis using dominating set

S Pradeepa[1], Vijaya Kaushika A [2], Vindhya Guru Rao [3]
[1]Assistant professor- II, School of Computing, SASTRA Deemed to be University, Thanjavur, India
[2, 3]Student, School of Computing, SASTRA Deemed to be University, Thanjavur, India
[1]pradeepa.pradee@gmail.com, [2]avkaushika@gmail.com, [3]vindhyarao97@gmail.com

***Abstract*— Today people are posting their problems, queries and health issues on online health-discussion forums, in search of symptoms, cure and preventive measures. Especially for medical conditions like stroke, where the causes or the time at which it can occur is uncertain, it is highly essential to be aware of all possible causes and preventive measures that can be adapted in day to day activities. A clearer idea is obtained about this by considering the data taken from online health-blogs and posts - an easier source to obtain information from patients as well as experienced doctors. Also, statistical analysis on such data allows health-organizations to keep track of the diseases, their symptoms and the extent to which they are affecting people. The objective of this methodology is to identify the accurate symptoms, causes, treatments and other important information about stroke. A new methodology, called association rule-based, hypergraph where the domination set determines this from a hypergraph which is constructed by the frequently occurring words extracted from the apriori algorithm based on the reviews from various sources.**

Keywords*— *association rule, apriori algorithm, hypergraphs, minimal dominating set, stroke

I. INTRODUCTION

Social media and group discussions online have become very active places where people from across the world come to discuss various issues. People tend to post their health related queries and problems online, about the diseases that affects them or their families, in order to know more about it, or find a solution. Hence, these discussion forums have rich data about the various diseases and medical conditions. There are even exclusive support groups and blog posts on the internet to spread awareness about stroke. There is huge amount of information available on such sites and using the right data mining algorithms, analysis of the different aspects of a disease like symptoms, causes, medicines prescribed, and other suggestions given by doctors.

Stroke is a clinical syndrome which spans for about 24 hours, sometimes may lead to death, with a vascular cause. [29] It occurs due to blockage of supply of blood to the brain from the arteries (ischemic) or due to haemorrhage in the brain(haemorrhagic). It is caused due to a variety of factors like age, blood pressure, sex(greater in women), smoking, alcohol, diabetes, genetics. It can occur at any time and the effect it has an each individual differs. Sometimes it's just a minor attack which sometimes it proves fatal. There are various measures that can be adapted in day to day life like having a healthy and balanced diet, regular exercise and physical activity, maintaining good body mass index, avoiding alcohol or drinking habits.

In a dataset of posts and queries, multiple problems, symptoms and advices from doctors would be written. To infer all possible combinations of problems one might face, after pre-processing the dataset, association rule on text data has to be applied. ARM is used to identify the pattern or relationship between items within a transaction[17]. Here, the intention is to find relation between multiple symptoms for the same patient, or between symptoms and medicines doctor would have suggested. This is followed by construction of hypergraph using the association rules as the hyperedges [13]. A single association rule would thus contain symptoms, medicines or preventive measures mentioned by the patient. These would be the vertices of the hyperedges. To finally identify the essential aspects like symptoms, complications, and prescriptions that would exist in the dataset, the dominating set in the line-graph representation is found [10] of the hypergraph that was constructed.

II. RELATED WORKS

Mining based on Association Rule (Association Rule Mining) is a crucial part in process of Knowledge Data Discovery (KDD) in data mining [20]. It has been effectively used in Market Basket Analysis (MBA) [1][16][18]. It helps a company predict profitable products, and judge its customers' preferences. ARM has also found application in other fields like finance, banking through analysis of customer's credit details, or loan records [19]. In Social-Networks, ARM has been proposed for personal hobby mining or finding the influential users in the network [22] [23].

In the field of medicine - in bioinformatics, genomics [20] [21] - there are heated discussions regarding the prediction of protein function by analysing the techniques for pre-processing protein interaction networks. ARM has also been proposed on gene expression to obtain biologically relevant associations between different genes.

ARM has also been proposed for text mining [14] [16]. It has been applied for topic identification, following which classification techniques like Naive-Bayes have been applied to classify the topics so identified. Such classification techniques however require a predefined bag-of-words. To

automate the topic identification, moreover, identify the topics that affect mainly, the usage of dominating set concept is proposed, for which in turn, a graph-based approach for its calculation is introduced. On text data and hypergraphs so constructed, Minimal dominating set property [13] is used which produces all the important words and phrases among the data.

To the best of our efforts, the use of ARM hasn't been made in the analysis of medical data available through media sources like health-blogs and patient-discussion forums and certainly not through a graph-based approach with the property of dominating set.

III. PROPOSED METHODOLOGY

The methodology proposed is an association rule based approach, to infer from the dataset and construct an hypergraph to establish relationships between items (aspects of a disease) and transactions (a post from the discussion forums). The steps involved are: 1. Dataset pre-processing, 2. Association rule generation, 3. Hypergraph construction and representing it as line graph, 4. Calculation of dominating set in this graph (minimal domination set)

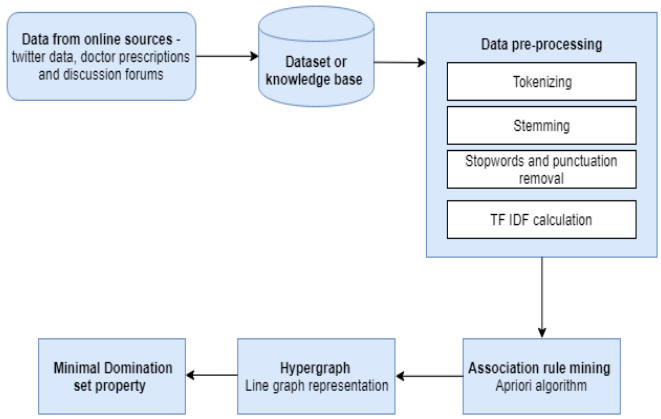

Figure 1: Schematic flow of ARM based approach – Architecture diagram

1. *Pre-processing the dataset*

Since this set of algorithms on data collected from media sources, it becomes very essential to pre-process this data highly so that, unnecessary words do not show up strictly in the final results. The words in the dataset are to be first stemmed after tokenizing, punctuations to be removed and then the stop-words. The resulting words from every post are to be processed with TF-IDF algorithm. The result of this comprises of only important and stemmed words.

a. Tokenizing is splitting a sentence into an array of its constituent words; this will include the punctuations involved in it. Stemming is reducing the words to their root forms. For instance, "paining" becomes "pain", also "painful" results into "pain". Essentially, it reduces redundancy while preserving the meaning.

b. Stop-words removal is an important step, wherein a post by a patient may contain a lot of commonly used words like "used to", "can", "couldn't", etc. so a lot of words apart from the words in the stop-words library of NLTK are also expected to be removed. Also, punctuation marks like '!', '?' etc., need to be removed.

c. Term Frequency - Inverse Document Frequency (TF-IDF) value calculation - The Term Frequency (TF) value of a word (or a term) 't' in a post/review (document) 'd' is given by the frequency 'f' of that word in the post divided by the number of words in that document (1). IDF value for a word refers to its importance within the whole dataset, considering its occurrence in every document (2). The TF-IDF value is simply the product of these values, as given by equation (3)

$$TF_{(t,d)} = \left(\frac{frequency\ of\ t\ in\ d}{number\ of\ words\ in\ d}\right) \quad (1)$$

$$IDF_{(t)} = ln\frac{n_{documents}}{n_{documents\ containing\ t}} \quad (2)$$

$$TF - IDF_{(word)} = TF_{(word)} * IDF_{(word)} \quad (3)$$

ALGORITHM 1: Data Pre-processing

Input: Data collected (written in an excel sheet)

1. While document in file
   1.1 tokenize the document
   1.2 stemming the tokenized words
2. Calculate, the TF value by passing the words as a blob-list using (1)
3. Calculate IDF value using equation (2)
4. Calculate the TF-IDF value, set a minimum threshold value
5. Consider the words with TF-IDF value greater than the threshold, write these words to a file, each row contains the words of one document

Output: List of words in a file after the TF-IDF calculation

2. *Association Rule generation – Apriori algorithm*

The data has been taken from patients and their posts on health discussion forums. The most important aspect of data mining is to integrate different data objects to get knowledge. This kind of an association between different data objects can be extracted using the a priori algorithm. For generating the association rules, one post or review is considered as a transaction and each word in the post as items of that transaction. So, the formal definitions of apriori and different measures of association rules like support, confidence and lift are as follows:

Association rule: Let IS = {IS_1, IS_2 ... IS_n} is the item-sets which consists of a set of attributes which are distinct (n attributes). Let D be the database consisting of all the uniquely

identified records, called tuples T and each record consists of an item-set, an association rule denotes implication like, $A => B$, where $A, B \in IS$, and $A \cap B = \emptyset$. The item-set A and B are known as the antecedent and consequent respectively. [1].

Support: The support (determined in terms of probability) is defined as the number of transactions or records in the database D that contain both the item-sets, B and A. For $A \rightarrow B$ in apriori (association rule), the support value is given by References. [2][3][5]

$$Support\ (A \rightarrow B) = Support\ (A \cup B) = P(A \cup B) \quad (4)$$

Confidence: The confidence (determined in terms of conditional probability) is defined as the number of transactions or records in the entire set of records or database that contains item-set A that also has the item-set B. Confidence value can also be determined with support values. The equation for the confidence is given by References. [2][5]

$$Confidence\ (A \rightarrow B) = P(B|A) = \frac{Support(A \cup B)}{Support(A)} \quad (5)$$

Here, Support (A∪B) means that the probability of item-set A and item-set B occurring together in the database and Support (A) is the probability of item-set "A" occurring in the database. [3]

Lift: Lift/Interest value is a measure to identify the frequency of A and B together given that both the item-sets are independent statistically [6][7] The value of lift of rule A → B is defined as:

$$Lift(A \rightarrow B) = \frac{Confidence}{Expected\ Confidence} = \frac{Confidence(A \rightarrow B)}{Support(A)} \quad (6)$$

ALGORITHM 2: Association Rule Generation

Input: The transactions from the file (obtained after pre-processing) as an iterable object.

1. Define the following
   1.1 min_support
   1.2 min_confidence
   1.3 min_lift
   1.4 max_length
2. Generate support records using items in the transaction objects. Use formula in (4)
3. Generation order statistics, calculate the confidence and lift values for the transactions using results of step 3 in (5) and (6)
4. Yield results that satisfy the threshold set for min_support, min_confidence, min_lift and max_length. Generate RelationRecords.

Output: Store the records as JSON, for later retrieval using dump_as_json function from apyori package of Python.

3. Hypergraph construction – line graph representation

A hypergraph is a graph in which greater than two vertices are connected to form one edge [13].

Definition (Hypergraph): A hypergraph H is a graph which can be represented as a pair *(V, E)* of a finite set of vertices, $V = v_1, v_2 \ldots v_n$ and a set of edges *E*, nonempty vertex subset, *V*. [8][9].

Definition (Generalized hypergraph): The concept of a hypergraph can be improvised such that the hyperedges are the vertices, that is, a hyperedge E is made of vertices which in turn make hyperedges. [9]

Line graph: Let a hypergraph, *H = (V, E)*, such that *E* = ∅. Line-graph (also called the representative or an intersection graph) of H is the L(H) = *(V, E)* such that:

1. *V:= I* or *V := E*, **H** has no similar hyperedges

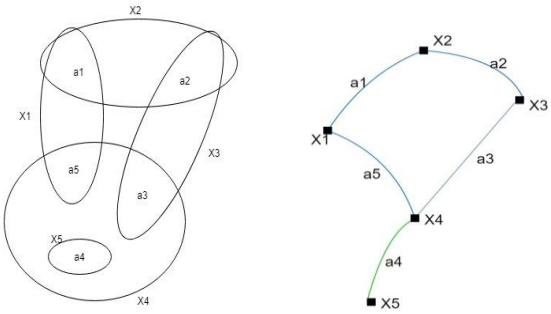

Figure 2 : Hypergraph and line graph representation

2. $\{i, j\} \in E\ (i \neq j)\ e_i \cap e_j = \emptyset$ [12] .

The hypergraph constructed here uses the association rules as hyperedges, where the vertices would be the set of words obtained after pre-processing. So, when the line graph is constructed, an edge connects the two nodes (association rules) only if there are one or more common words between them. The figures 2 and 3 demonstrate this effectively. The common elements (a1, a2...) represent the edges, while the nodes (X1, X2...) represent the association rules.

ALGORITHM 3: Hypergraph construction and line graph representation

1. Retrieve the transactions from the association rules generated.
2. Let H be the hypergraph, $H = (V_H; E_H)$. The vertices V_H are the set of all items. The hyperedges E_H, is the set of all transactions.
3. Initialize a graph $G = (V_L, E_L)$
4. Each transaction t is a node in its line graph representation, in other words add t to V_L, where $t \in E_H$
5. Add an edge e between two nodes t_1 and t_2, if there exists at least one item i in common to them, that is, if $\exists i, i \in t_1\ and\ i \in t_2$. Add e to E_L.
6. The line graph results as G.

4. Dominating set – Minimal dominating number

Let a hypergraph $H = (V, E)$ where V. E denoting the set of vertices and edges. A dominating set D, in the hypergraph $H = (V, E)$ can be defined as a subset $D \subseteq V$ of the set of vertices of the hypergraph, for each vertex in the graph $v \in V - D$ there is an edge $e \in E$ for which $v \in e$ and$e \cap D \neq \emptyset$. In other words, every vertex $v \in V - D$ is adjacent or has an edge to a vertex in the set D.

Domination number γ is defined as the minimum cardinality (number of elements in the set) of a dominating set in a hyperedge, H. [11]

This definition is with respect to hypergraphs; for an easier implementation, the dominating set on the line-graph constructed is found for the representation of the hypergraph.

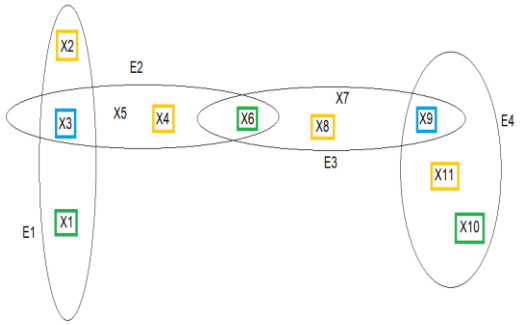

Figure 3: An Example of hypergraph with various dominating sets

The possible dominating sets for the Fig. 4 hypergraph would be {X3, X9}, {X1, X6, X10}, {X2, X4, X8, X11} - only the minimum cardinality dominating set is considered.

ALGORITHM 4: Domination set

Input: Hypergraph $G = (V,E)$ (line graph representation)

1. A vertex $v \in V$ is chosen from the vertex set V , let D be dominating set of $G = \{v\}$
2. If $V - (D \cup N(D)) \neq \emptyset$,
   2.1 A vertex w is chosen such that, $V - (D \cup N(D))$ and add it to the domination set $D \leftarrow D \cup \{w\}$
3. Find the dominating set with the minimal cardinality

Output: A minimal dominating set D of G

IV. Experimental Analysis

1. Data pre-processing

Around 2000 data was collected from Twitter, blog posts, social media, patient and health discussion forums and google forms through web scraping tools and Twitter API programs in Python. Words in the data are stemmed and tokenized using Porter Stemmer and sent_tokenize and word_tokenize using the NLTK (Natural Language Toolkit) library in Python. The stop-words were also removed by using the corpus library. The TF-IDF values for these processed words were found. The threshold value was set to 0.15 to eliminate unnecessary words.

These words are put into a CSV file such that, each row depicts one document and the cells of that row consists of the essential words obtained after TF-IDF calculation.

TF-IDF values for each document (post)

Top words in document 1
Word: pressur, TF-IDF: 0.22907
Word: blood, TF-IDF: 0.22713
Top words in document 2
Word: coconut, TF-IDF: 0.34112
Word: oil, TF-IDF: 0.21499
Top words in document 3
Word: diet, TF-IDF: 0.63995
Top words in document 4
Word: vascprotect, TF-IDF: 0.40456
Word: secondari, TF-IDF: 0.40456
Word: coronari, TF-IDF: 0.34578

2. Association Rule – Apriori Algorithm

According to apriori algorithm, each document is considered as a transaction and the words as items. Here, an effort is made to find the combination of words that would occur frequently in the entire dataset, create association rules for them and calculate their support value, lift value and confidence value. The threshold value (minimum) of values are considered as 0.003(support), 0.3(confidence) and 1(lift). The Python package apyori was used to obtain these rules, then these were saved as JSON objects to help for further processing.

Thus, each association rule obtained was of the form:

{"items": ["blood", "pressur"], "support": 0.1555, "ordered_statistics": [{"items_base": ["blood"], "items_add": ["pressur"], "confidence": 0.9873015873015872, "lift": 6.349206349206349}, {"items_base": ["pressur"], "items_add": ["blood"], "confidence": 1.0, "lift": 6.349206349206349}]}

Itemsets	Support	Confidence	Lift
["genet", "nitric", "oxid", "predisposit"]	0.0145	1.0	66.666
["coronari", "secondari", "vascprotect"]	0.017	1.0	58.823
["atherosclerot", "coronari", "disea"]	0.0075	0.288	133.333
["nitric", "oxid"]	0.0145	0.275	68.965
["exercis", "physicalact"]	0.0055	0.687	125.0
["death", "stroke"]	0.0805	0.670	8.333
["cholesterol", "obes"]	0.003	0.352	30.690

Table 1: Association rule – Apriori algorithm results

3. Hypergraph – line graph representation

The items of each item-set, the apriori algorithm rules are taken and they are made as the nodes of the graph. If there is an intersection between the two nodes, that is, if there are two or more common words between the two “items” of the rules, an edge is established. The NetworkX library in Python was used to implement line-graph representation of the hypergraph.

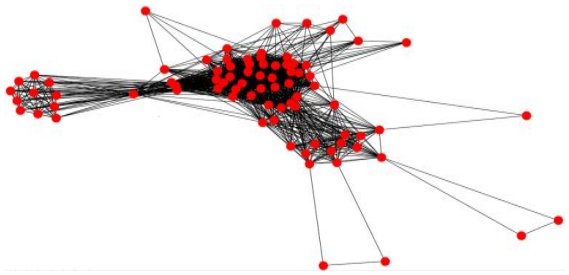

Figure 4: Resulting line graph (of the hypergraph)

4. *Minimal Dominating Set property*

Dominating set is the set of nodes from the graph which have the potential to represent the entire graph. This dominating set obtains the most essential nodes of the graph which can almost represent all the nodes of the graph. This concept of graphs or hypergraphs can be implemented here to extract the most primary combination of words that would almost represent the entire graph. For our purpose, the dominating set of this graph provides the most essential of all the discussions and tells us which symptoms, medicines or preventive measures are being talked about the most throughout the dataset. The NetworkX library of Python provides the function that can calculate the dominating set of any graph constructed through the use of the library. The domination number is 20.

The minimal domination set is

["['macular', 'vitamin']", "['pressur']", "['coconut', 'oil']", "['diab', 'stroke']", "['nitric', 'oxid']", "['cardioembolic', 'periodont']", "['heart']", "['cancer', 'chronic', 'lung']", "['reproduct']", "['physicalact']", "['blood']", "['malaria', 'tuberculosi']", "['genet', 'gucy1a3', 'predisposit']", "['death', 'respirato']", "['cholesterol', 'obes']", "['atherosclerot', 'disea']", "['calcium']", "['coronari']", "['exercis']", "['secondari', 'vascprotect']"]

V. RESULTS

The lift value uses the all the values present in the associated words. Hence, this value can be used to calculate the percentage with respect to the other lift values to obtain the most important associated words of the domination set.

Lift percentage is calculated by considering the lift values of each associated words and divided by the sum of lift values of all the words.

$$Lift\ percentage = \frac{lift\ (word)}{sum\ of\ lift\ values} * 100$$

This metric indicates which associated minimal domination set words have a greater occurrence.

The support, confidence and lift values of the associated words are as in table 2

Words	**Support**	**Confidence**	**Lift**	**Lift percentage**
['macular', 'vitamin']	0.0035	0.7	199.999	20.447
['pressur']	0.1555	0.1555	1.0	0.102
['coconut', 'oil']	0.0325	0.280	8.620	0.881
['diab', 'stroke']	0.0805	1.0	12.422	1.269
['nitric', 'oxid']	0.0145	1.0	68.965	7.050
['cardioembolic', 'periodont']	0.008	0.571	71.428	7.302
['heart']	0.0045	0.004	1.0	0.102
['cancer', 'chronic', 'lung']	0.004	0.16	13.333	1.363
['reproduct']	0.0105	0.0105	1.0	0.102
['physicalact']	0.0055	0.0055	1.0	0.102
['blood']	0.1575	0.1575	1.0	0.102
['malaria', 'tuberculosi']	0.003	1.0	333.333	34.077
['genet', 'gucy1a3', 'predisposit']	0.0145	1.0	68.965	7.050
['death', 'respirato']	0.0045	0.037	8.333	0.851
['cholesterol', 'obes ']	0.003	0.352	30.690	3.137
['atherosclerot', 'disea']	0.0075	0.714	95.238	9.736
['calcium']	0.009	0.009	1.0	0.102
['coronari']	0.031	0.031	1.0	0.102
['exercis']	0.008	0.008	1.0	0.102
['secondari', 'vascprotect']	0.017	1.0	58.823	6.013

Table 2: Results – association words in the minimal dominating set with apriori values and lift percentage

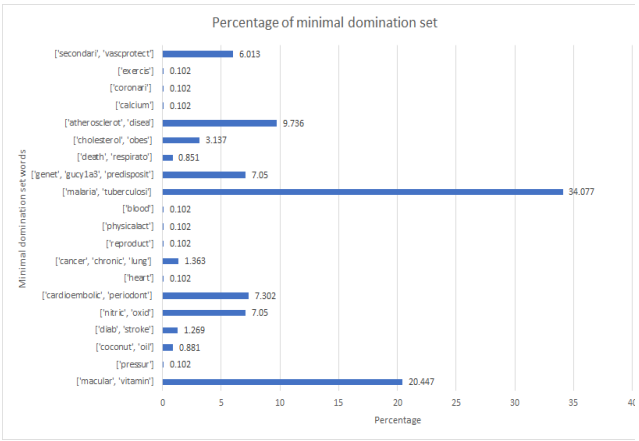

Figure 5 : Lift percentage of various associated words in the minimal dominating set

It can be inferred that there is maximum discussion on risks of vertebrobasilar stroke associated with malaria(cerebral) and tuberculosis that affect the central nervous system followed by the combination macular, vitamin indicating that Age-related macular degeneration (AMD) is associated with cardiovascular risk and decreased survival and due to lack of vitamin A and E. Another important result is that if a person suffers from periodontal disease, he/she also is at the risk of having a cardioembolic stroke. Nitric oxide plays a significant and positive role in reducing stroke attack. GUCY1A3 is a type of gene that is associated with vessel stroke with a predisposition to diabetes. Atherosclerosis is one of the major causes of stroke. People suffering from lung cancer, blood pressure, reproductive syndromes and having high levels of cholesterol have a high possibility of a stroke. Exercises, diet

and doing some kind of physical activity everyday are ways by which the risk of stroke can be reduced which is correctly indicated in the results.

VI. CONCLUSION AND FUTURE WORK

The results obtained indicate the mostly discussed causes, symptoms and preventive measures taken by people across the globe for stroke with a percentage metric to indicate the importance of the associated word combinations.

It is a digital world and people share their problems and put their queries forward in a hope to find solutions, preventive measures, even try to identify causes, by posting them online. With so much data about diseases and their symptoms, and everything related to them, there is a need for implementation of data mining algorithms to discover proper knowledge about it and derive statistics as to which problem affects one the most. The hybrid-set of algorithms that have been proposed has successfully identified major symptoms, eating habits and all stroke related problems that patients have talked of online. These algorithms can be applied for various other diseases, about which people or even doctors have talked about, and derive conclusions about health conditions of people all over the world at once. Medical field is where this seems well appropriate, but perhaps it can be extended to other fields as well.

REFERENCES

[1] M. H. Dunham, Y. Xiao, L. Gruenwald, and Z. Hossain, “A Survey of Association Rules,” Technical Report, Southern Methodist University, Department of Computer Science, Technical Report TR 00-CSE-8, 2000.

[2] Jiawei Han, Micheline Kamber, and Jian Pei. 2011. Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[3] Dinesh J. Prajapati, Sanjay Garg, N.C. Chauhan, Interesting association rule mining with consistent and inconsistent rule detection from big sales data in distributed environment, Future Computing and Informatics Journal, Volume 2, Issue 1, 2017, Pages 19-30, ISSN 2314-7288

[4] Hotho, A., Nürnberger, A. & Paaß, G. (2005). A Brief Survey of Text Mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, 20, 19-62.

[5] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining association rules between sets of items in large databases. SIGMOD Rec. 22, 2 (June 1993), 207-216.

[6] T. Brijs, K. Vanhoof, and G. Wets. Defining interestingness for association rules. International journal of information theories and applications, 10(4):370-376,2003.

[7] Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur. 1997. Dynamic itemset counting and implication rules for market basket data. In Proceedings of the 1997 ACM SIGMOD international conference on Management of data (SIGMOD '97), Joan M. Peckman, Sudha Ram, and Michael Franklin (Eds.). ACM, New York, NY, USA, 255-264.

[8] Molnár, B.: Applications of hypergraphs in informatics: a survey and opportunities for research. Ann. Univ. Sci. Budapest. Sect. Comput. 42, 261–282 (2014)

[9] Bretto A. (2013) Applications of Hypergraph Theory: A Brief Overview. In: Hypergraph Theory. Mathematical Engineering. Springer, Heidelberg

[10] Bretto A. (2013) Hypergraphs: First Properties. In: Hypergraph Theory. Mathematical Engineering. Springer, Heidelberg

[11] Csilla BujtÃ¡s, Michael A. Henning, Zsolt Tuza, Transversals and domination in uniform hypergraphs, European Journal of Combinatorics, Volume 33, Issue 1, 2012, Pages 62-71, ISSN 0195-6698

[12] Berge, C. (1989). *Hypergraphs: Combinatorics of Finite Sets*. North-Holland.

[13] Purnima Gupta, Rajesh Singh, S Arumugam. (2016). Characterising minimal point set domination sets : AKCE International Journal of Graphs and Combinatorics.

[14] Chowdhury Mofizur Rahman, Ferdous Ahmed Sohel, Parvez Naushad, Kamruzzaman S. M, “Text Classification Using the Concept of Association Rule of Data Mining,” In Proceedings of International Conference on Information Technology, Kathmandu, Nepal, pp 234-241, May 23-26, 2003.

[15] A.A. Lopes, R. Pinho, F.V. Paulovich, R. Minghim, Visual text mining using association rules, Computers & Graphics, Volume 31, Issue 3, 2007, Pages 316-326, ISSN 0097-8493.

[16] Manpreet Kaur, Shivani Kang, Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining, Procedia Computer Science, Volume 85, 2016, Pages 78-85, ISSN 1877-0509

[17] J. Manimaran and T. Velmurugan, "A survey of association rule mining in text applications," *2013 IEEE International Conference on Computational Intelligence and Computing Research*, Enathi, 2013, pp. 1-5. doi: 10.1109/ICCIC.2013.6724258

[18] Kaur Paramjit, Attwal Kanwalpreet S. Data Mining: Review, International Journal of Computer Science and Information Technologies, 5 (5) (2014), pp. 6225-6228.

[19] X. Wu, V. Kumar, J.R. Quilan, J. Ghosh, Q. Yang, H. Motoda Top 10 Algorithms in Data Mining. Springer-Verlay London Limited, 14 (2007), pp. 1-37.

[20] Atluri G., Gupta R., Fang G., Pandey G., Steinbach M., Kumar V. (2009) Association Analysis Techniques for Bioinformatics Problems. In: Rajasekaran S. (eds) Bioinformatics and Computational Biology. Lecture Notes in Computer Science, vol 5462. Springer, Berlin, Heidelberg

[21] M. Anandhavalli, M. K. Ghose, and M. Gauthaman. Association Rule Mining in Genomics. International Journal of Computer Theory and Engineering, 2(2):17938201, April 2010.

[22] X. Yu, H. Liu, J. Shi, J. N. Hwang, W. Wan and J. Lu, "Association Rule Mining of Personal Hobbies in Social Networks," *2014 IEEE International Congress on Big Data*, Anchorage, AK, 2014, pp. 310-314. doi: 10.1109/BigData.Congress.2014.52

[23] F. Erlandsson, P. Bródka, A. Borg, H. Johnson, "Finding influential users in social media using association rule learning", *Entropy*, vol. 18, May 2016.

[24] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: Analysis and an algorithm. Advances in neural information processing systems, 2:849–856, 2002.

[25] Ulrike Von Luxburg. A tutorial on spectral clustering. Statistics and computing, 17(4):395–416, 2007

[26] Jianbo Shi and J. Malik, "Normalized cuts and image segmentation," Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, 1997, pp. 731-737. doi: 10.1109/CVPR.1997.609407

[27] Lowell W. Beineke and Robin J Wilson – “Topics in Structural Graph Theory” – Chapter 12 - Abdol–Hossein Esfahanian, “Connectivity Algorithms”

[28] Hugh Markus, “Stroke: causes and clinical features” - Medicine, Elseiver - Volume 44, Issue 9, September 2016, Page no - 515 - 5