Extracting patterns from Global Terrorist Dataset (GTD) Using Co-Clustering approach

Muhammad Adnan[1], Muhammad Rafi[2]

[1]*Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST) Karachi, Pakistan*

[2]*National University of Computer & Emerging Sciences NU-FAST Karachi, Pakistan*

[1]adnansiddiq@outlook.com

[2]rafi.muhammad@gmail.com

Abstract — Global Terrorist Dataset (GTD) is a vast collection of terrorist activities reported around the globe. The terrorism database incorporates more than 27,000 terrorism incidents from 1968 to 2014. Every record has spatial data, a period stamp, and a few different fields (e.g. strategies, weapon sorts, targets and wounds). There were few earlier studies to find interesting patterns from this textual gamut of data. The author believes that GTD has numerous interesting patterns still hidden and the full potential of this resource is still to be divulged. In this Independent Study, the author tries to investigate the GTD through co-clustering method for pattern discovery. Author has extracted textual data from GTD as per motivation to cluster the data in space and time simultaneously, through co-clustering. Co-clustering has become an important and powerful tool for data mining. By using co-clustering, bilateral data can be analysed by describing the connections between two different entities. There are many applications in the real world that can extensively benefits from this approach of co-clustering, such as market basket analysis and recommendation system. In this study, the effectiveness of co-clustering model will be described by performing experiment on database of global terrorist events.

***Keywords*—Global Terrorism Dataset, GTD, Co-clustering, Bi-clustering, Two-Way-Clustering**

I. INTRODUCTION

The Global Terrorist Dataset (GTD) is one of the largest collections of recorded terrorist activities across the globe. The terrorism database includes more than twenty seven thousand terrorism incidents occurred between 1968 and 2006. Each record has some spatial information, a period stamp, and a couple of distinctive fields (e.g. methods, weapon sorts, targets and wounds). The author believes that Global Terrorist Dataset (GTD) has various fascinating patterns still abstracted and the maximum potential of this dataset is still to be unwrapped. This information can be extracted by means of various data mining techniques. Clustering can be the first choice to explore this dataset. Clustering is a fundamental apparatus in unsupervised machine learning that is utilized to gather comparable objects, and has functional significance in a wide variety of applications, for example, web-log and business crate information classification. Similarly, the information that emerges in these applications is organized as a co-occurrence table, for example, word report co-occurrence or site page client scanning information. Clustering methodology comprehend the patterns to implicitly cluster the objects. By definition clustering is the association between objects in such manner, so that collection object is collectively assembled on the premise of likeness between them. The clustering process finds the similar patterns among set of untagged data. The process of clustering involves the representation of information, counting the similarity measure and nature of the algorithm used for clustering.

In machine learning and applications of data mining, the data usually arises in matrix format [1]. Mostly, the data matrices that emerge in real world applications composed of large number of rows and columns. To understand the structure of such data, matrices are one of fundamental problems in data mining. The most algorithms for clustering aims on restricted clustering (one-way clustering), i.e. cluster one aspect of the table focused around similarities along the second aspect. [2]

For instance, documents may be clustered based on distributions of word or words may be clustered based on how they are distributed among a document. Similarly for clustering of document, the process gets the pattern of similar information and clusters them in accordance to calculated similarity measure.

It is mostly desirable to co-cluster or clustering both measures of a possibility table by exploiting the acceptable duality in the between rows and columns. Case in point, it is intriguing by discovering comparable documents and their interaction with word clusters. Similarly, one aspect of the possibility table is when managing inadequate and high-dimensional information, it ends up being advantageous to utilize co-clustering.

For this study, the author is using the approach of co clustering to find out the hidden parameters in Global Terrorist Dataset (GTD). The dataset have enormous information about almost every terrorist event that take place in the time period of 1972 to 2014. By using co clustering, it would be constructive to extract the patterns of terrorist activities that correlate with each other. This approach can

help in study and analyze the global terrorism more accurately.

A. Co clustering vs. One way clustering

Co-clustering is more advantageous compared to traditional "single-sided" clustering from various point of views:

The simultaneous grouping of column and row clusters is much more useful and informative. The approach of co-clustering gives more compressed representations that are interpretable while saving the majority of the data contained in the original information. This quality of co-clustering approach makes it more valuable applications of statistical data analysis [3].

A column (or row) clustering can be considered dimensionality lessening along the column (alternately row) [4]. The simultaneous clustering along columns and rows diminishes dimensionality from both axes. Statistical issue can be prompted drastically with more modest number of parameters and thus, a substantially more compact representation for subsequent analysis can be obtained. Since co-clustering integrate column clustering data into row clustering and the other way around, one can consider it a "statistical regularization" method that can yield better quality clusters regardless that one is fundamentally interested in one-sided clustering. The impact of Statistical regularization is of a great degree imperatively managing large data matrices for instance, those emerging in text mining. A comparable instance can be drawn from subspace clustering techniques which utilizes the maximum potential of the co-clustering approach. [3]

B. Global Terrorists Dataset (GTD)

Open source databases containing information of terrorist activities have gained much attention and consideration in the recent decades owing to prominence of terrorism crime information collection. While the strategies of collection have varied but these databases depend on reports of terrorism from electronic and print media. The evolution of open source databases of terrorist events has taken into account more thorough analysis of terrorist activities. However, these databases have excluded the domestic terrorist attack.

Generally, an attack to foreign boundary includes local(s) from same country. Similarly, the domestic attacks involve a national or a group of nationals who attack their own homeland. Previously, some piece of the explanation behind excluding the domestic attacks from these databases was official. Many international agencies, like the US State-Department, have had a long history of focusing on universal terrorism. But beyond the traditional approach of dividing bureaucratic responsibilities in accordance with international and domestic, distinctions was considered as the significant challenge of collecting the data on very large number of terrorist incidents.

The primary mission of United States Department of Security Homeland (DHS) was to save the United States of America from any terrorist attack and to decrease the vulnerability of the United States of America from any future terrorist attack. With the help of United States Department of Security Homeland (DHS), a group of specialists at “United States consortium for study of terrorism and response to terrorism” has created a database intended to benefits examiners and experts and approach creators attain DHS's mission and additionally, this provides a construction for the next investigation of terrorist events. The purpose of Global Terrorism Database (GTD) is to provide every detail about the events of terrorist activities that happened in the world since 1970 (including local cases with-in United States). Like information about an American person attacking on some target with in united states or a French person attacking some target with in France and also international terrorist events (where the culprit attacked a target in a remote nation, as with the 9/11 attack) [5]. Organized by a group of social scientists, the information has been assembled to consider for the systematic analyses of patterns in terrorist activity, across places, over time and by the different classes of terrorist groups. While the world may confront extraordinary terrorist threats today, experts can explore a lot of things about today's dangers by investigating the practices of terrorist groups in the recent past and the effects that these terrorist actors, and the terrorist events they executed, have had. The GTD is a vital instrument to consider such investigation. [5-6]

The project of Global Terrorist Dataset (GTD) was started in 2001 when Laura Dugan and Gary Lafree at Maryland University, acquired a large amount of information initially gathered by Intelligence Services of Pinkerton Global (PGIS). In the time period starting from 1970 and ending at 1997, the Pinkerton trained some of the researchers to distinguish and record information about terrorism events from government reports, wire services and authentic international newspapers. In 2005, with of launching of START, funds were provided to check the validity and reliability of PGIS data, and in April 2006, partnership with the centre for Terrorism and Intelligence Studies (CETIS), collection of the data post-1997 events was started. In 2008, the newly collected data was purposely incorporated with the existing data to structure a solitary wellspring of data on terrorist assaults, from 1970 to 2014. Currently the Global Terrorist database (GTD) has in excess of sixty thousand incidents and one hundred and nineteen different dimensions. In excess of two thousand different groups of terrorists have been recorded in the database, associated with occasions on twenty seven years’ time duration of events from 1970 to 1997. [6]

II. LITERATURE REVIEW

An extensive amount of literature is devoted for the analysis and study of terrorism. The analysis and studies concentrate on introducing the after effects of their analyses by using qualitative methodologies or basic diagrams (for example, histograms or line charts) to show the patterns of variables. For the dataset of Global Terrorism Dataset (GTD) that holds more than a hundred types of different dimensions, the charts and descriptions of analysis did not sufficiently impart the complex connections between all variables and specifically their relationship with each other. In this circumstance, it is supposed to be challenging for the experts to distinguish and to identify all hidden patterns, or to form theories and to execute high level strategic hypothesis [7].

Due of the complexity and huge size of dataset, identifying the patterns of terrorist activities and practices is demanding. To help investigators in better understanding of activities of terrorist events, Wang *et al* propose a visual analytical system that focuses on characterizing one of the basic concepts of investigation analysis, the five W's (what, where, who, why, and when). [8] The description in their system was extensively correlated and each description express one of five W's (what, where, who, why, and when). By using this approach, investigators can now investigate activities of terrorists more efficiently and effectively and find reasons of the attack by distinguishing patterns transiently between various group of terrorists (who), geo-spatially (where) and from different modes and methods of attack (what). By combining global point of view with the points of interest gathered by investigating these questions, system permits investigators to think of both strategically and deliberately. [8]

The Visualization and investigation of social network is a settled territory in both the visualization community [9] and sociology [10] yet very few number systems are connected to understand large amount of activities of terrorists. Shen et al. [9] created Ontovis which uses a cosmology chart to picture huge heterogeneous systems and connect it to delineate connections between different terrorist groups.

Perer *et al* [11] analysed and investigated the global terrorist dataset (GTD) in their social action system and show connections between group of terrorist and countries. Then again, despite the fact that geo-transient visualization is additionally a well-established area for new researches, especially in the field of geological information system, there exists an application of this approach for terrorism data. It has been discovered that few comprehensive systems have incorporated both the geo-temporal and social parts the of terror activities. A system proposed by Zhu *et al* [5] exhibited a structure for consequently recognizing what, when, who, and where in a notable story in regards to terrorist event. While their proposed approach focuses on the extraction of data from ambiguous and unstructured context.

Additionally, Lee [12] presented the Global Terrorist Dataset (GTD) Explorer, an electronic intelligent visual exploratory instrument. It checks the quantity of terrorist events assembled over a certain criteria along with stack of diagrams on top of one another to see both aggregated and individual patterns over time. The web based tool gives first hand deep understanding of patterns to the experts by making the information more readable and descriptive. To make this visualization light-weight and available was one of the fundamental concerns of making tool. [12]

For this study, the author is using the technique of co-clustering on Global Terrorist Dataset (GTD) and will co-cluster the dataset on basis of time and location of terrorist activity. The author tries to extract the hidden patterns from Global Terrorist Dataset. The technique of co-clustering has already been used in many different areas of biological study likewise gene classification and extraction of hidden patterns from gene microarray data.

The theory of co-clustering was initially presented by Hartigan [2] in the year of 1972. He specified first algorithm for co-clustering approach [13]. Similarly, the term of co-clustering was initially presented by Mirkin later. There was no generalized algorithm for co-clustering until Cheng and Church [13] presented their first generalized algorithm for co-clustering and they applied it to gene data study. This research is still considered as most important writing for co-clustering. Further contributions to the study of co-clustering include two algorithms that were presented by I.S. Dhillon in 2002 and 2003. One of the co-clustering algorithms presented by I.S.Dhillon was focused on bipartite spectral graph partitioning [14] while other was based on information theorem [15]. These two algorithms become bases for file co-clustering algorithms presented recently.

Similarly, Yizong Chengzx and George M. Churchz have introduced the concept behind co-cluster for gene study. They have introduced the significance of co-clustering compared to a subset of conditions and to subset of genes with a high likeness score. Comparability is not considered as a function of sets of conditions or sets of genes. Rather, it is measured as a coherence of the conditions in co-cluster and genes.

Co-clustering is additionally specified in the writing as direct clustering and bi-clustering among other names, the approach of co-clustering has also been utilized as a part of fields, like data mining and data recovery. In a survey, Madeira *et al* [16] investigate a substantial number of existing methodologies to co-clustering and arrange them in accordance with types of co-cluster they can discover, the types of co-clusters that are found, the approaches that are used to perform the research, the target applications, and the methodologies used to evaluate solutions.[16]

III. RESEARCH OBJECTIVE

The objectives of this research work are mentioned below:

- The primary objective of this research work is to use co-clustering approach on Global Terrorist Dataset (GTD) to explain the terrorist dataset in context of time (when) and location of terrorist event (where).
- The co-clustering approach is mostly used for the study of gene classification in biological studies In order to utilize the enriched features of co clustering for data analysis; the tests are conducted on the dataset of GTD.
- After applying the mentioned co-clustering algorithm on the dataset, results were displayed using graphical diagrams to easily understand the outcomes of the process.

IV. RESEARCH METHODOLOGY

The below mentioned methodologies are followed for this research.

A. Experimental Research

The author has used experimental research design for this study. . In this type of research, trials are carried out to achieve a result. The main advantage of this research design is that it gives opportunity to author to explain the cause and effect relation of the problem. The tests were conducted on the dataset of Global Terrorist Dataset (GTD) to produce co-clusters based on the location and time of the terrorist activity and finally results were displayed using graphical representations.

B. Quantitative Research

For this study, the author has implemented the approach of using co-clustering to identify the relation between different terrorist activities. This research is quantitative in nature in such a way that the quantitative clusters are produced on the co-related data and similarly the results are produced.

V. EXPERIMENT

A. Experimental Setup

For this study, several experiments were performed on the dataset of global terrorist dataset (GTD). The author has created a setup for experiments. The experiment was conducted on HP Pro-Book 450 machine. The machine processing power was composed of 8GB ram, with processor of Intel core i5 third generation and a storage of 450 gigabyte of hard drive. R (software for statistical computing) was used for generating the co-clusters from the pre-processed dataset. Some pre-processing was already done on the dataset to achieve the required representation of data for clustering. The pre-processing of dataset was performed using Microsoft excel.

B. Data Set

This study is based on justifying the use of co - clustering approach for the purpose of extraction of information from the dataset of GTD and therefore several experiments were performed on the dataset. The dataset of Global Terrorist Dataset (GTD) is very much descriptive. It has enough information related to almost every terrorism event that has taken place on the earth from 1970 until today. A sub dataset was used from the dataset of GTD containing information of terrorist events from January 2010 to December 2013.

C. Data Processing

A subset of global terrorist dataset was selected. The subset of global terrorist dataset that was selected for co-clustering contains the information of terrorist activities from 2010 to 2013. Initially, the dataset was pre-processed to achieve the required representation of clustering. The required representation is the matrix of data composed of x and y axis of different attributes of information. The terrorist data set of GTD composed of data containing thousands of rows, each row correspond to a single terrorist activity. Similarly, each record has multiple columns, and each column describe certain property related to that terrorist event. So, the first part clustering process was to achieve a representation of data where a property on x axis and a different property related to same terrorist activity the y axis of matrix can be obtained. After generation of the required representation of data, it was processed on R software where the library of block-cluster for co-clustering data was used. The library of block cluster read the data and produces the required number of cluster by shuffling the rows and columns in the optimal position. The generated co - clusters are represented by using visual maps of data using plot library of R software.

D. R Software

R is computer programming language and statistical computing environment which is mostly used by computer scientist for data processing. R computing environment comes bundled with many important libraries for data processing; however, there are many other open source libraries which are available for any requirement. The best feature of R is the generation of publication quality graphs. R handles mathematical formulae's and symbol with care and

can generate required outputs with little effort. R worked on the principle of ease of use. It provides the user all the controls if it is required to control certain parameters or else, R handles much of the complexities of computing processes itself.

E. Block-cluster Library for R

Block cluster library of R was used in this study for co-clustering. This library comes with much different functionalities for co-clustering depending upon the type and nature of data. The library is capable of performing co-clustering for continuous, binary and contingency data.

VI. RESULTS

For this study, the results were produced by co-clustering two dimensional data of a subset of global terrorist dataset. The experiments were performed and results were extracted for two use cases.

- Co-clusters for number of month wise terrorist attacks in different countries
- Co-clusters for number of month wise attacks by different terrorist groups

A. Co-clusters for number of month wise terrorist attacks in different countries

The data was pre-processed in format of month wise sum of terrorist activities for different countries and then generated co-clusters using block-cluster library in R. The pre-clustered formation of dataset is shown below in figure 1.

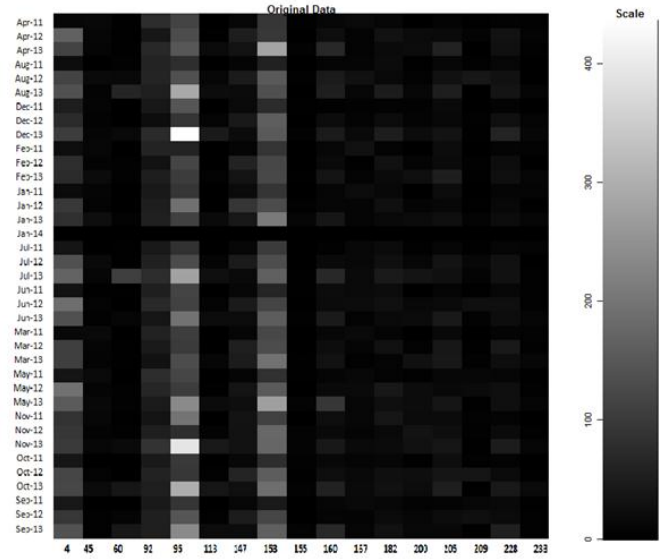

Fig. (1). Graphical Representation of pre-clustered dataset for experiment: 1

Where on X axis we have country codes

4	=	Afghanistan
45	=	Colombia
60	=	Egypt
92	=	India
95	=	Iraq
113	=	Libya
147	=	Nigeria
153	=	Pakistan
155	=	West Bank and Gaza Strip
160	=	Philippines
167	=	Russia
182	=	Somalia
200	=	Syria
205	=	Thailand
209	=	Turkey
228	=	Yemen
233	=	Northern Ireland

The data is shuffled before clustering, as it can be seen that the data is hard to analyse and no visible patterns present, therefore, it is required to have common patterns to analyse the data and get the meaningful information from this state of dataset, so after performing co - clustering on this dataset, this output was produced. The essence of co-clustering is finding the similarity among a dataset on the basis of dual properties. Like, in this case, it is required to co-cluster data that can provide meaningful information related to the countries having similar number of attacks. By this way, co – clustering of data can provide the clusters of countries with similar number of attacks with respect to month of years.

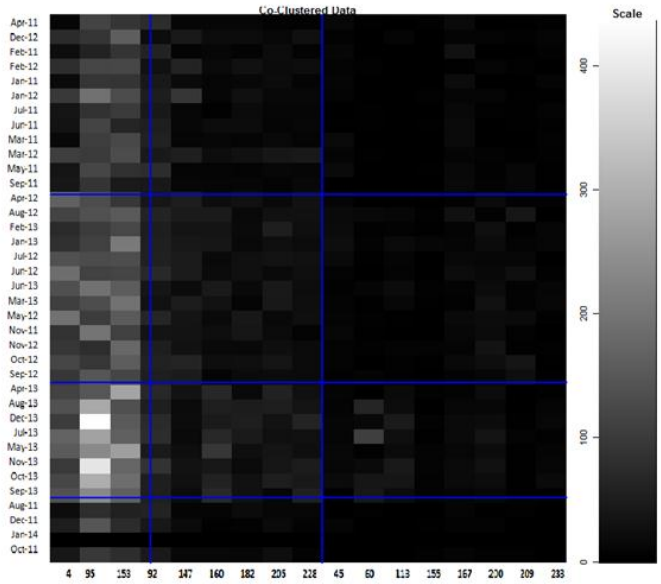

Fig. (2). Graphical Representation of co-clustered dataset for experiment: 1

Where on X axis we have country codes

4	=	Afghanistan
45	=	Colombia
60	=	Egypt
92	=	India
95	=	Iraq
113	=	Libya
147	=	Nigeria
153	=	Pakistan
155	=	West Bank and Gaza Strip
160	=	Philippines
167	=	Russia
182	=	Somalia
200	=	Syria
205	=	Thailand
209	=	Turkey
228	=	Yemen
233	=	Northern Ireland

As it can be seen in figure 2, the blocks in white are the months with highest number of attacks for specific country. For example, the second down cluster from left, it can be seen that countries namely Pakistan, Afghanistan and Iraq have higher number of attacks in corresponding months. Similarly, there are other co-clusters that are displaying more meaning full information for country and month wise attacks.

B. Co- Clustering for month wise number of attacks by terrorist groups

Similarly, the co-clusters for number of attacks by specific month and terrorist group was also generated. The results produced several co-clusters with meaningful information. After generation of required representation of data by month on Y axis and terrorist group number of X axis, get this original state of data in R in figure 3 was acquired.

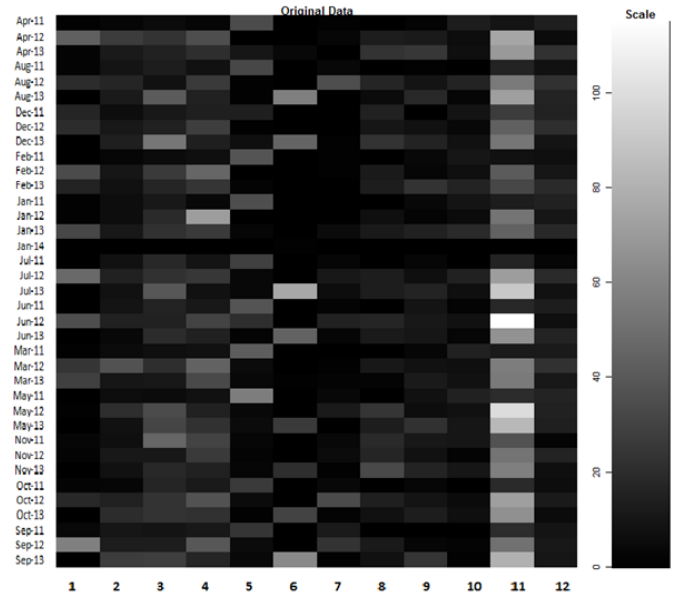

Fig. (3). Graphical Representation of pre-clustered dataset for experiment: 2

Where on X axis we have terrorist group codes

1 = Al-Qa'ida in Iraq
2 = Al-Qa'ida in the Arabian Peninsula (AQAP)
3 = Al-Shabaab
4 = Boko Haram
5 = Communist Party of India - Maoist (CPI-Maoist)
6 = Islamic State of Iraq and the Levant
7 = Kurdistan Workers' Party (PKK)
8 = Maoists
9 = New People's Army (NPA)
10 = Revolutionary Armed Forces of Colombia (FARC)
11 = Taliban
12 = Tehrik-e-Taliban Pakistan (TTP)

For this experiment, the number of terrorist activities by different terrorist groups and months were selected. The goal was to find similarity between the behaviors of terrorist in different months. This approach can be beneficial in finding the similar nature of terrorist groups with respect to time period. As it can be seen, the dataset of GTD was preprocessed and the required representation of data was acquired where terrorist groups were laid on x axis and on y axis months of years were set. After performing co-clustering on this dataset, several co-clusters were obtained that define meaningful information related to common factor between months and terrorist groups in figure 4.

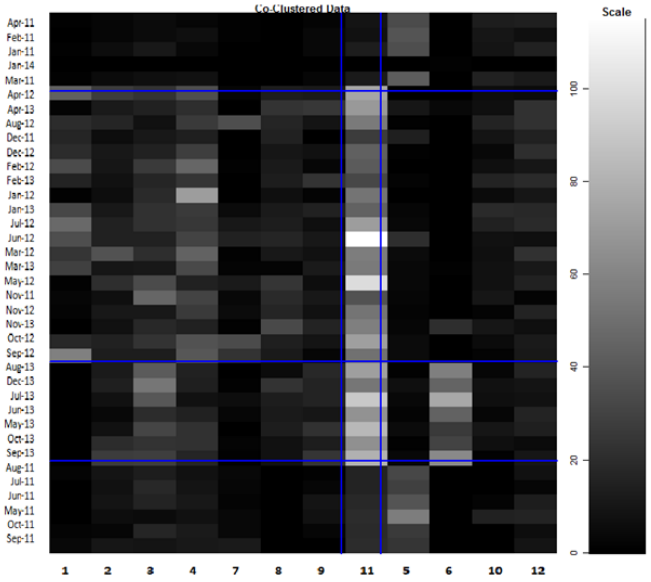

Fig. (4). Graphical Representation of co-clustered dataset for experiment: 2

Where on X axis we have terrorist group codes

1 = Al-Qa'ida in Iraq
2 = Al-Qa'ida in the Arabian Peninsula (AQAP)
3 = Al-Shabaab
4 = Boko Haram
5 = Communist Party of India - Maoist (CPI-Maoist)
6 = Islamic State of Iraq and the Levant
7 = Kurdistan Workers' Party (PKK)
8 = Maoists
9 = New People's Army (NPA)
10 = Revolutionary Armed Forces of Colombia (FARC)
11 = Taliban
12 = Tehrik-e-Taliban Pakistan (TTP)

As per results generated through block cluster library, several clusters had been generated, depending on the similarity between type terrorist groups and months of year. These co-clusters describe the behaviour of different terrorist groups with time.

VII. CONCLUSION

The purpose of co-clustering is to extract the hidden information from the dataset depending upon the similarity on x and y axis of the dataset. This approach is widely used in biological sciences for gene classification but in other applications, it is not widely used. By extracting the patterns and similarity between x and y axis of the dataset of GTD, it has been explained the importance of co-clustering approach in real world applications. This approach has never been used before on Global Terrorist Dataset. Several experiments were performed to prove the importance of using this approach to extract meaningful information from the dataset.

VIII. FUTURE WORK

In future the co-clustering can be used on Global Terrorist Dataset to extract more meaningful information. This approach will help the experts to analyse the patterns of terrorist events and will be helpful to control the terrorism.

REFERENCES

[1] D. Agarwal and S. Merugu. "Predictive discrete latent factor models for large scale dyadic data". In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '07)*, 2007, pp: 26–35,

[2] J. A. Hartigan. "Direct clustering of a data matrix" *Journal of the American Statistical Association*, vol. 67, no. 337, 1972.

[3] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu and D. S. Modha. "A Generalized Maximum Entropy Approach to Bregman Co-clustering and Matrix Approximation" *Journal of Machine Learning Research*, vol. 8, pp: 1919-1986, 2007.

[4] I. S. Dhillon, S. Mallela and D. S. Modha. "Information Theoretic Coclustering" In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp: 89-98, 2003.

[5] G. LaFree. "The Global Terrorism Database: Accomplishments and Challenges", *Perspectives on Terrorism*, vol. 4, no. 1, 2010.

[6] G. Lafree and L. Dugan. "Introducing the Global Terrorism Database" Terrorism *and Political Violence*, vol. 19, no. 2, pp: 181-204, 2007.

[7] D. Guo, K. Liao and M. Morgar. "Visualizing patterns in a global terrorism incident database" *Environment and Planning B: Planning and Design*, vol. 34, pp: 767-784, 2007.

[8] X. Wang, E. Miller, K. Smarick, W. Ribarsky and R. Chang. "Investigative Visual Analysis of Global Terrorism" In *Proceedings of the 10th Joint Eurographics / IEEE - VGTC conference on Visualization*, vol. 27, no. 3, pp: 919-926, 2008.

[9] Z. Shen , K. L. MA and T. Eliassi-Rad . "Visual analysis of large heterogeneous social networks by semantic and structural abstraction". *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 6, pp: 1427-1439, 2006

[10] L. C. Freeman. "Visualizing social networks". *Journal of Social Structure*, vol. 1, 2000.

[11] A. Perer and B. Shneiderman. "Balancing systematic and flexible exploration of social networks". *Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp: 693-700 2006.

[12] J. lee. "Exploring Global Terrorism Data: A Web-based Visualization of Temporal Data" Crossroads, vol. 15, no. 2, pp: 7-14, 2008.

[13] Y. Cheng and G. M. Church. "Biclustering of expression data". In *Proceedings of International Conference on Intelligent Systems for Molecular Biology (ISMB)*, 2000, pp: 93–103.

[14] I. S. Dhillon. "Co-clustering documents and words using bipartite spectral graph partitioning". In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001)*, pp: 269–274, 2001.

[15] I. S. Dhillon, J. Fan, and Y. Guan. "Efficient clustering of very large document collections". In *Data Mining for Scientific and Engineering Applications*, R. Grossman, C. Kamath, P. Kegelmeyer, V. Kumar, and R. Namburu, eds, Kluwer Academic Publishers, 2001, pp: 357–381.

[16] S. C. Madeira and A. L. Oilveiral. "Biclustering Algorithms for Biological Data Analysis: A Survey" *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 1, no. 1, pp: 24-45, 2004.