An Approach of Mining Big Data from a Very Large Community Graph for Analyzing of Economic Standard of Communities using Distributed Mining Techniques

Bapuji Rao
Department of CSE & IT
V.I.T.A.M
Berhampur, Odisha, India

Anirban Mitra
Department of CSE & IT
V.I.T.A.M
Berhampur, Odisha, India

ABSTRACT
This paper gives an overview on the fundamental concepts of big data and its characteristics. We have discussed on the issues related to Graph Analytics for Big Data. Basic definitions are presented in order to describe the big data environments using the notation of Graph theory. Two cases, the first one includes the information and relation with in the film and movie industry and the second one is the web structure and relationship (crawling) between different web sites has been elaborated in this direction. The paper concludes with our observation on the proposed model followed by a case analysis on applications of big data in social media.

General Terms
Community Graph, States, Panchayats, Villages

Keywords
Data Mining, Big Data, Dataset, Community, Distributed Database**.**

1. INTRODUCTION
We live in a digital world now-a-days. In the digitization world, new methods are exploded for creation and storing of amount of structured and unstructured data [2]. Big Data refers to datasets whose sizes are beyond the ability of typical database software tools to capture, store, manage and analyse. Big data applies to information that can't be processed or analyzed using traditional processes or tools. There is no explicit definition of how big a dataset should be in order to be considered Big Data. The data is too big, moves too fast, or does not fit the structures of existing database architectures [2].

2. DEFINITIONS FOR BIG DATA
Big data is data that exceeds the processing capacity of conventional database systems [1]. The data is too big, moves too fast, or does not fit the structures of existing database architectures. To gain value from these data, there must be an alternative way to process it.

Big data is a collection of data sets of large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications [3].

Big data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze [5].

3. CHARACTERISTICS OF BIG DATA
The characteristic of big data consists of four simple terms i.e. Volume, Velocity, Variety [2, 13], and Veracity [3].

In "Big Data", the meaning of "big" means ***volume*** [2, 13]. So volume is a relative term. Depending on the size of organisations the data storage varies from gigabytes to terabytes. If the organisation is a big global organisation then the data storage may varies from petabytes to exabytes. Many of these organisations datasets are within the terabytes range today but, soon they could reach petabytes or even exabytes in the near future.

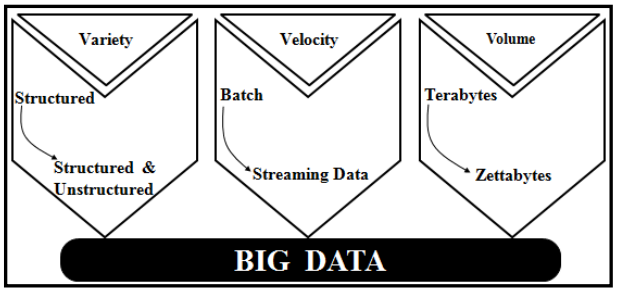

Fig 1: *3V's* in Big Data [3]

The ***velocity*** [2, 13] of data in terms of the frequency of its generation and delivery. Velocity in the context of Big Data means how quickly the data arrives and stores, and how quickly it can be retrieved. Velocity can also be applied to data in motion: the speed at which the data flows. The various information streams and the increase in sensor network deployment have led to a constant flow of data at a specific pace.

Data can come from a ***variety*** [2, 13] of sources (that is from internal and external source of organisation) and in a variety of types. With the inventions of sensors, smart devices which able to retrieve structured traditional relational data as well as also *semi-structured* and *unstructured* data.

3.1 Structured data
Data is grouped into a relational scheme or schema (e.g., rows and columns within a standard database).

3.2 Semi-structured data
It is a structured data that does not confirm to an explicit and fixed schema or scheme. The data is inherently self-describing and contains tags or other markers to enforce hierarchies of records and fields within the data. Examples include weblogs and social media feeds.

3.3 Unstructured data

Data consists of formats which cannot easily be indexed into relational tables for analysis or querying. Examples include images, audio and video files.

Veracity [3] is a kind of prediction from a collection of big data.

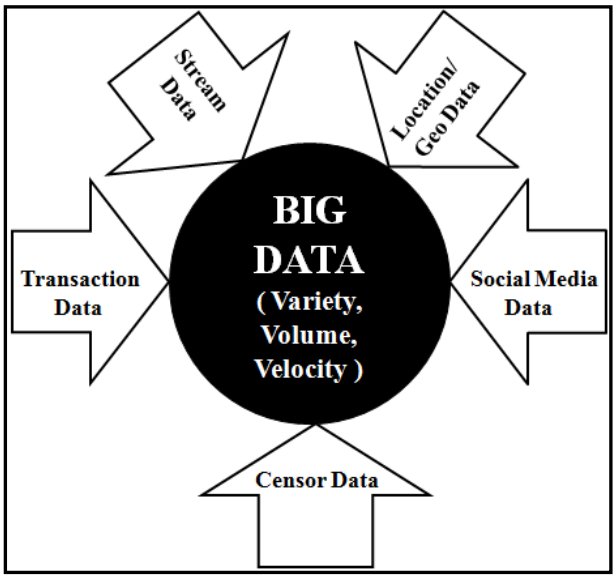

Fig 2: *3V's* in Big Data [2, 13]

4. GRAPH ANALYTICS FOR BIG DATA

Graph analytics is the study and analysis of data that can be transformed into a graph representation consisting of nodes and links. Graph analytics is good for solving problems that do not require the processing of all available data within a data set. A typical graph analytics problem requires the graph traversal technique. Graph traversal is a process of walking through the directly connected nodes. An example of a graph analytics problem is to find out how many ways two members of a social network are linked directly and indirectly. A more contemporary example of graph analytics relates to social networks.

4.1 On graph mining and its applications

The ability to mine data to extract useful knowledge has become one of the most important challenges in government, industry, and scientific communities. So far we have seen success in mining data which represents a set of independent entities and their attributes, for example, customer transactions. However, in most domains, there is interesting knowledge to be mined from the relationships between entities. This type of knowledge may take many forms from periodic patterns of transactions to complicated structural patterns of interrelated transactions. Extracting such knowledge requires the data to be represented in a form that not only captures the relational information but supports efficient and effective mining of this data and comprehensibility of the resulting knowledge.

4.2 Definition of graph mining

Mining graph data, sometimes called *graph-based data mining*, is the extraction of novel and useful knowledge from a graph representation of data [6].

In general, the data can take many forms from a single, time-varying real number to a complex interconnection of entities and relationships. While graphs can represent this entire spectrum of data, they are typically used only when relationships are crucial to the domain.

The most natural form of knowledge that can be extracted from graphs is also a graph. Therefore, the *knowledge*, sometimes referred to as *patterns*, mined from the data are typically expressed as graphs, which may be sub-graphs of the graphical data, or more abstract expressions of the trends reflected in the data.

5. SOME ISSUES ON GRAPH ANALYTICS

5.1 Single path analysis

The goal is to find a path through the graph, starting with a specific node. All the links and the corresponding vertices that can be reached immediately from the starting node are first evaluated. From the identified vertices, one is selected, based on a certain set of criteria and the first hop is made. After that, the process continues. The result will be a path consisting of a number of vertices and edges.

5.2 Optimal path analysis

This analysis finds the 'best' path between two vertices. The best path could be the shortest path, the cheapest path or the fastest path, depending on the properties of the vertices and the edges.

5.3 Vertex centrality analysis

This analysis identifies the centrality of a vertex based on several centrality assessment properties.

5.4 Degree centrality

This measure indicates how many edges a vertex has. The more edges there are, the higher the degree centrality.

5.5 Closeness centrality

This measure identifies the vertex that has the smallest number of hops to other vertices. The closeness centrality of the node refers to the proximity of the vertex in reference to other vertices. Higher the closeness centrality is the more number of vertices that require short paths to the other vertices.

5.6 Eigen vector centrality

This measure indicates the importance of a vertex in a graph. Scores are assigned to vertices, based on the principle that connections to high-scoring vertices contribute more to the score than equal connections to low-scoring vertices.

6. APPLICATIONS OF GRAPH ANALYTICS

In the finance sector, graph analytics is useful for understanding the money transfer pathways. A money transfer between bank accounts may require several intermediate bank accounts and graph analytics can be applied to determine the different relationships between different account holders. Running the graph analytics algorithm on the huge financial transaction data sets will help to alert banks to possible cases of fraudulent transactions or money laundering.

The use of graph analytics in the logistics sector is not new. Optimal path analysis is the obvious form of graph analytics that can be used in logistics distribution and shipment environments. There are many examples of using graph analytics in this area and they include "the shortest route to deliver goods to various addresses" and the "most cost effective routes for goods delivery".

One of the most contemporary use cases of graph analytics is in the area of social media. It can be used not just to identify relationships in the social network, but to understand them. One outcome from using graph analytics on social media is to identify the "influential" figures from each social graph. Businesses can then spend more effort in engaging this specific group of people in their marketing campaigns or customer relationship management efforts.

6.1 An example on use of graph analytics

For analytical purposes, a social network is visualized as a *digraph* (in a graph if the relationship has no direction) [12]. So in a digraph, one unit may be an individual, a family, a household, a village, an organization in a village is called a *node* or *vertex*. A tie between two nodes indicates the presence of a relationship. No tie between two nodes is the absence of a relationship. A tie with a direction is called an *arc* and tie without direction is called an *edge*. The weight of a tie is the value or volume of flow. If the *arc* or *edge* is labeled with any weight then the graph is termed as *weighted graph*. In social networking we concentrate only the *presence (1)* or *absence (0)* of the relationship. We also assume that ties have directions.

Let us denote ***G*** is a digraph. The set of vertices of ***G*** can be denoted by ***V (G)*** and the set of arcs can be denoted by ***A (G)***. If *uv* is an arc, then diagrammatically it can be shown as an arrow from vertex *u* to vertex *v*. if both *uv* and *vu* are arcs, then we sometimes represent these two together by a line without arrow heads joining vertex *u* and vertex *v*.

The given "Fig 3" is a digraph ***G***. The vertex set is $\boldsymbol{V = \{v_1, v_2, \ldots\ldots, v_{21}\}}$. The different arcs are v_1v_2, v_2v_1, v_3v_1, v_4v_1, v_4v_5 etc., but v_1v_3 and v_2v_3 are not arcs in the graph ***G***.

6.2 Representing social network using properties of graph theory and matrices

The 1st network in "Fig 4" illustrates that everybody goes to everybody else. The 2nd network, the ties are reciprocated but the network is highly fragmented. The 3rd network is connected but highly centralized and shows concentration of power lies in only one node or vertex. The 4th network is connected cyclic i.e. everybody can go to everybody else through a large number of intermediaries. The 5th network illustrates a strong hierarchy, things flow only in one direction. The number of vertices and number of arcs constitute the most basic data in a social network.

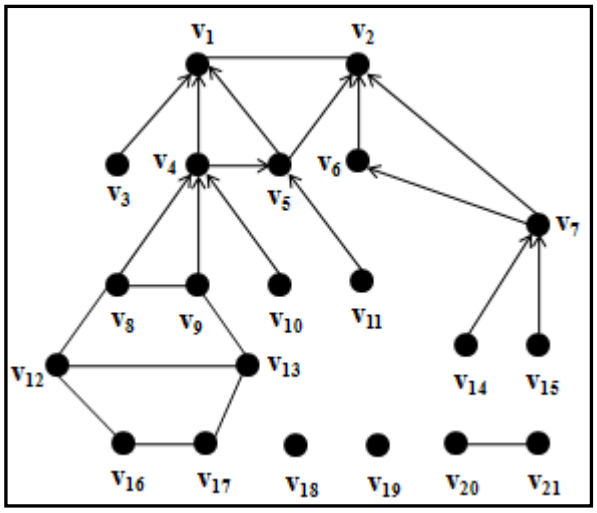

Fig 3: A digraph G

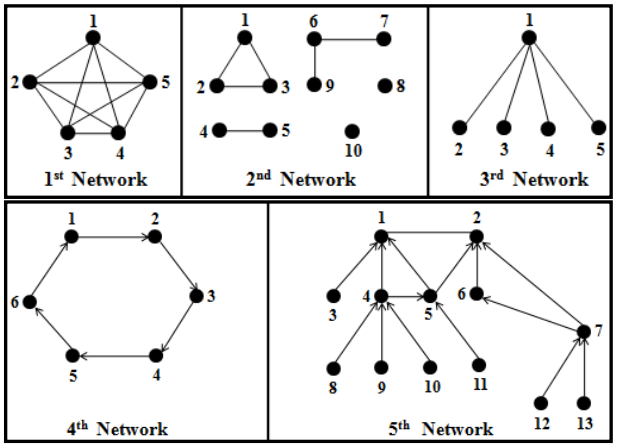

Fig 4: Five different networks

Let us consider an example for the 4th network of "Fig 4". Suppose there are six households in a neighborhood connected in a circular way. Here each of household goes to exactly one among the remaining five and only one of the remaining five comes to it. If we impose this condition further, there three other patterns possible besides that the 4th network of "Fig 4". These patterns are shown in "Fig 5".

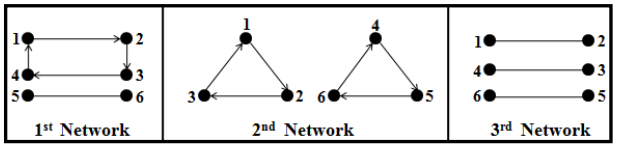

Fig 5: Three different patterns of 4th network

7. REPRESENTATION OF BIG DATA USING GRAPH THEORY CONCEPTS

7.1 Representation of movie database

Three domains that epitomize the tasks of mining graph data are the Internet Movie Database, the Mutagenesis dataset, and the World Wide Web. These databases may also serve as a benchmark set of problems for comparing and contrasting different graph-based data mining methods [6].

By representing movie information as a graph, relationships between movies, people, and attributes can be captured and

included in the analysis. "Fig 6(a)" shows one possible representation of information related to a single movie.

In the "Fig 6(a)" movie graph, we can report large fraction of sub-graphs. These sub-graphs may report discoveries such as movies receiving awards often come from the same small set of studios which is shown in "Fig 6(b)" or certain director/composer pairs work together frequently which is shown in "Fig 6(c)".

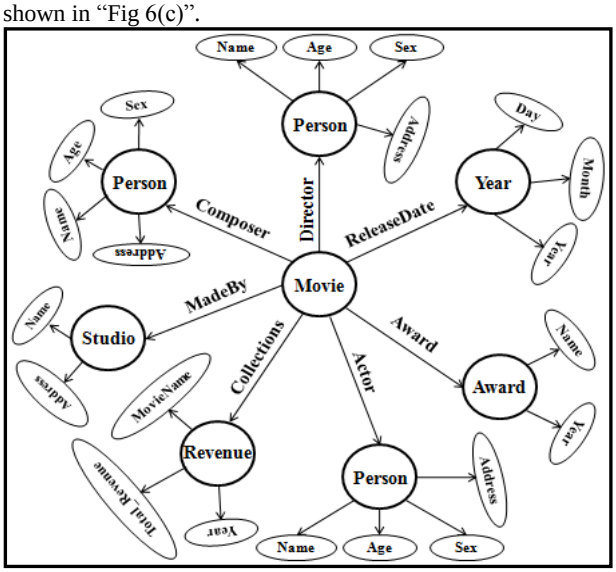

Fig 6: (a) Possible graph representation for information related to a single movie

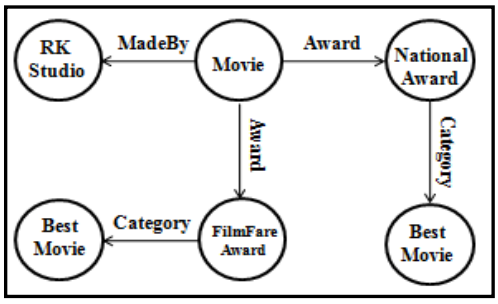

Fig 6: (b) One possible frequent sub-graph

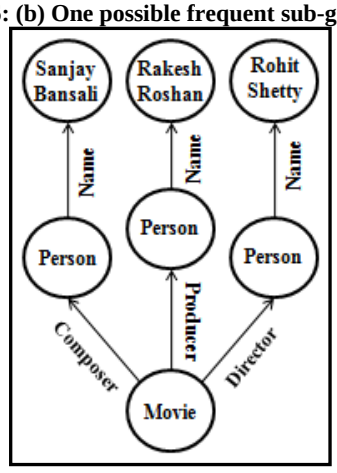

Fig 6: (c) Another possible frequent sub-graph

7.2 Representation of World Wide Web

World Wide Web is a valuable information resource that is complex, dynamically evolving, and rich in structure. Mining the Web is a research area that is almost as old as the Web itself. Etzioni coined the term "Web mining" [7] to refer to extracting information from Web documents and services. The types of information that can be extracted are so varied that this has been refined to three classes of mining tasks: Web Content Mining, Web Structure Mining, and Web Usage Mining [8].

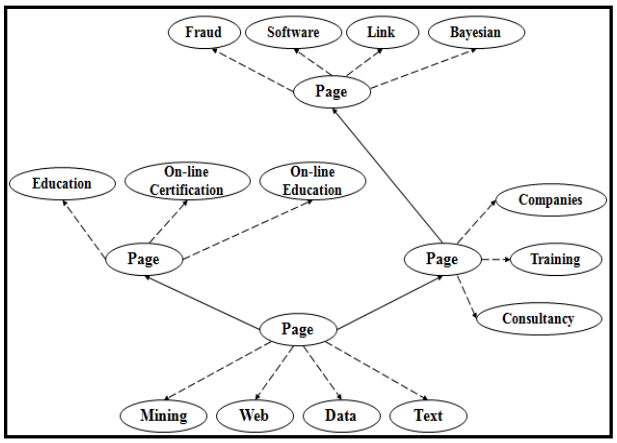

Fig 7: Graph representation for web text and structure data. Solid arrows represent edges labeled ''hyperlink'' and dashed arrows represent edges labeled ''keyword"

"Fig 7" shows a sample graph of this type for a collection of three Web pages. With the inclusion of this hypertext information, Web page classification can be performed based on structure alone [6] or together with Web content information [4]. Algorithms that analyze Web pages based on more than textual content can also potentially glean more complex patterns, considering the example of link between job pages and publication where exist the prevalence of data mining web pages with respective links.

8. OUR OBSERVATIONS AND PREDICTIONS

Considering a state is one large community graph. We combine India's 29 states together to form a very large community graph [9, 10, 11].

From this very large community graph how big data can be mined so that the standard of living of a particular community can be analyzed. For this we can apply distributed mining technique to discover the very big data which is in 4V. From it we try to select the desired item-set so that prediction can be taken place on those big data.

We have proposed a large community graph consisting of ***n***-states such as $\{S_1, S_2, S_3, \ldots\ldots, S_n\}$ which is shown in "Fig 8". Each state consists of various panchayats ***(panchayat is an Indian term for administration of villages)*** which is shown in "Fig 9". Each village consists of various communities which are shown in "Fig 10". Each village's community is considered as one database. The ***n***-states have connectivity and the connectivity is considered as edge. Similarly each state consists of various villages. Among these villages there is connectivity. Similarly each village consists of various communities. Among the communities there is connectivity. To keep this scenario in the mind, we can represent our Indian community network as a graph with ***29*** states. From this large graph database we can retrieve big data for analyzing to know about the economy standard of a particular community living in a village of a particular state. The community graph has

states $S = \{S_1, S_2, \ldots, S_n\}$, panchayats $P = \{P_1, P_2, \ldots, P_n\}$, villages $V = \{V_1, V_2, \ldots, V_n\}$, and communities $C = \{C_1, C_2, \ldots, C_n\}$ respectively.

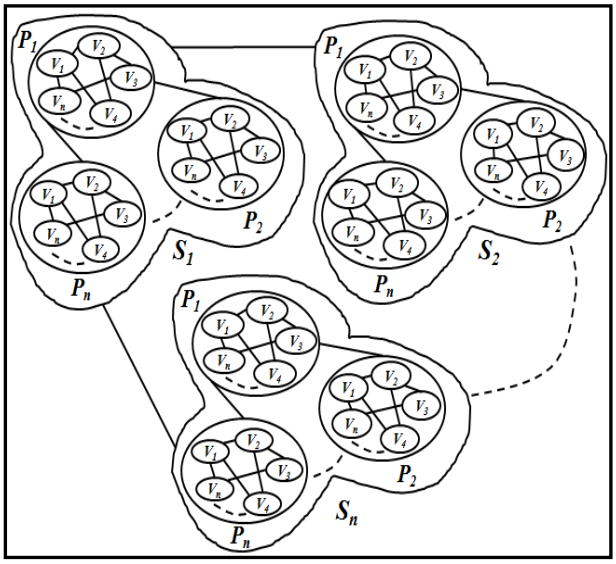

Fig 8: Proposed community graph with *n*-states

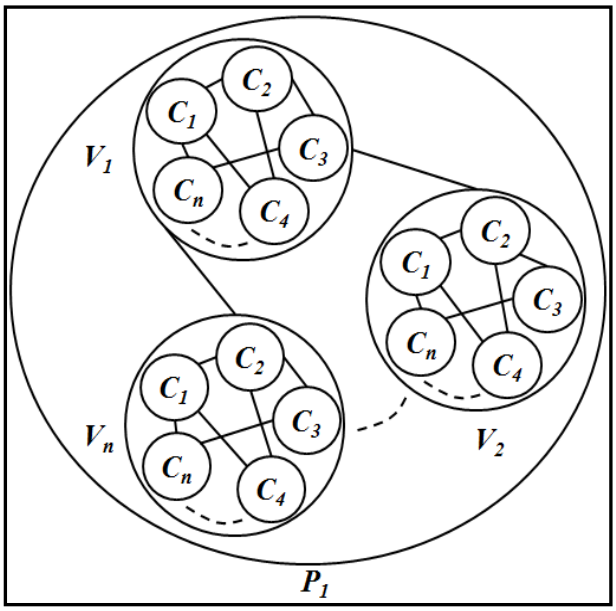

Fig 9: Panchayat P_1 with *n*-villages

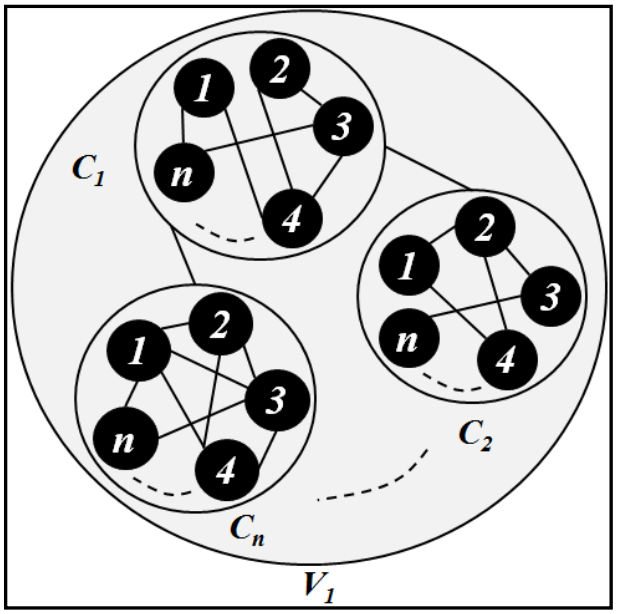

Fig 10: Village V_1 with *n*-communities

8.1 A brief case study on application of big data on social media

Social Media usages among its consumers are growing at exponential pace thus resulting in huge amount of data created at every minute. General users are no more restricted in using the Internet, rather usage of Smart phones', location based apps and other Internet of Things, has lead to generation of data in a much faster pace. Looking at the basic statistics on the social data growth, one can easily observe that more than 250 million tweets are generated in a day and it is increasing at a high speed. Further, at an average of 30 billion pieces of content are shared on Facebook per month. It is predicted that data will grow over 800% in the next 5 years and 80% of these data will be unstructured [2].

While handling such huge volumes of data is big challenge as well as it also provides numerous opportunities and competitive edge for the enterprises to acquire, store and manage the data for information and knowledge extraction. Thus, the need for a robust platform is highly necessary, which can efficiently handle hardware challenges that are common in case of Big Data.

A high performance scalable architecture is essential component of the infrastructure for processing big data. Segregating or storing the unstructured data from different sources within minimum response time for query processing with accuracy remains a software challenge for such an environment. Providing security and safety and thus extracting pattern, meaning or sense from the huge stored data is the ultimate challenge for the big data environment.

The analysis of Big Data will help in combining social media data streams with enterprise applications in a powerful way to derive meaningful insights on the social conversations. Progress can be made in analysis the opinion and sentiments employee based on their interaction with social media.

Further, the developed analysis agent can be built so as it can process huge volume of data at a higher speed and can identify sentiment, buzzword, predictive, correlation and influencer data.

The knowledge extracted from analysis of big data of social media has many important advantages. The organization can easily detect and respond to a social outburst before negative sentiments go viral in the social world. Decisions can be easily taken after analyzing the customer's sentiments and purchase pattern. Identifying and engaging key influencers who are impacting the increase or decrease of sales is now easy, as the decision is based on the information extracted from huge data.

9. CONCLUSIONS

Fundamental concepts of big data and its characteristics have been discussed. This work has focused on issues of Graph Analytics and its applications in Big Data. Two cases, the first one includes the information and relation with in the film and movie industry and the second one is the web structure and relationship (crawling) between different web sites has been elaborated in this direction. The paper concludes with our observation on the proposed model followed by a case analysis on applications of big data in social media.

10. REFERENCES

[1] Edd Dumbill. What is big data?[Online] Available from: http://radar.oreilly.com/2012/01/what-is-big-data.html.

[2] Infosys – Connect Architecture, Big Data Spectrum, published by Infosys Ltd(India), 2013, pp. 1-61.

[3] ieee.bigdata.tutorial.1.1, published by ieee society, 2014.

[4] J. Gonzalez, L. B. Holder, and D. J. Cook. Graph-based relational concept learning. In Proceedings of the International Machine Learning Conference, 2002.

[5] James Manyika, et al. Big data: The next frontier for innovation, competition, and productivity.http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation.

[6] *Mining Graph Data*, Edited by Diane J. Cook and Lawrence B. Holder Electrical Engineering and Computer Science, Washington State University, Pullman, Washington, Copyright ©2007 John Wiley & Sons, Inc.

[7] O. Etzioni. The World wide web: Quagmire or gold mine? *Communications of the ACM* 39(11):65–68, 1996.

[8] P. Kolari and A. Joshi. Web mining: Research and practice. *IEEE Computing in Science and Engineering* 6(4):49–53, 2004.

[9] Rao, Bapuji and Mitra, A. An approach to Merging of two community graph using Graph Mining Techniques. 2014 IEEE International Conference on Computational Intelligence & Computer Research (IEEE-ICCIC-2014), pp. 460-466, India, Dec 18-20, 2014.

[10] Rao, Bapuji and Mitra, A. A New Approach for Detection of Common Communities in a Social Network using Graph Mining Techniques. 2014 IEEE International Conference on High Performance Computing & Application (IEEE-ICHPCA-2014), Bhubaneswar, India, Dec 22-24, 2014.(Available at http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=7032933)

[11] Rao, Bapuji, Mitra, A and Narayana, U. An approach to study properties and behavior of Social Network using Graph Mining Techniques. In proceedings of DIGNATE::ETEECT 2014, New Delhi, India, Oct, 2014.

[12] Social Network Analysis by Prof. Suraj Bandyopadhyay, Prof. Bikas K Sinha and Late Prof. A.R.Rao, Indian Statistical Institute, Calcutta, India; Publishers: Sage Publications, Inc.

[13] Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data by Chris Eaton, Dirk Deroos, Tom Deutsch, George Lapis, and Paul C. Zikopoulus, Publishers: McGraw-Hill.