International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 4, Issue 09, September-2018

A SURVEY OF ADAPTIVE TASK SCHEDULING FOR SOCIAL BIG DATA

[1]Hemant Chilhate, [2]Dr. Nishchol mishra

[1]*School of Information technology RGPV, Bhopal, India*
[2]*School of Information technology RGPV, Bhopal, India*

Abstract— Social media is a large source of Big Data. Its data is continuously increasing and changing therefore it requires new and innovative forms of information processing. Research areas like Data Mining, Machine learning and Social Networks finds its application in analyses of Big Data. Different graph and networks processing algorithms like calculating centrality, identifying clusters and identifying sources of information diffusion are applied on user graphs. Most of the data is in unstructured format of texts. This gives rise to perform different Text Analytics methodologies on social media data. Variety of tools, machine learning libraries and frameworks are developed for effective utilization of Data Mining Methods. Innovative methods of data management are also required as traditional storages are ineffective for unstructured data. These Big Data processing methods found their applications in wide areas like Marketing, Criminal activities and Fraud detection, Epidemic Intelligence etc. There are also a number of open challenges in these processing techniques. This paper gives an overview of methodologies and frameworks for processing social media big data.

Index Terms—Big Data, Social Media Data, Text Analytics, Network Analysis, Predictive Modeling, Information Diffusion, Information Fusion, Apache Hadoop, Apache Spark

I. INTRODUCTION

Social media has become a large pool of data in the recent years. Processing this data effectively is the need of the hour since it contains high potential of information insights. Data on social media is present in huge volume, is dynamically changing and is of different varieties. Because of these properties processing social media data with traditional methods of data processing is ineffective and innovative forms of processing are required. These processing methods fall under Big Data processing mechanisms. The paper covers recent frameworks and different algorithms which are used for analyzing social media data. Following is a conceptual model of Social Big Data as described in[1].

Fig 1: Conceptual Model of Social Big Data

Data Analytics, Social Media and Big Data are the different concepts. Social Media acts as the source of data. Parallel processing paradigms are covered in Big Data. Data Analysis encompassed the set of methods used to extract knowledge. Intersection represents mixing of concepts. The machine learning tools and frameworks comes in the intersection of Data

Analysis and Big Data. Current applications which are used by social media websites are depicted between Social Media and Data Analysis. Lastly, between Big Data and Social Media, applications for developing web based systems are shown. Some of them are MongoDB and Hadoop. The center is the goal i.e. knowledge and insights.

II. METHODOLOGIES ANDALGORITHMS

Relevant knowledge can be extracted from Social Big Data using various algorithms which spans in areas like Data Mining, Graph Mining, Information Retrieval and Networks Analytics.

A. NetworkAnalytics

A connected world of networks is formed on social media. These networks are used for collective information extraction. Connected networks are represented in the form of graphs, with users as nodes and Relationships as edges [1]. Various network analytics can be applied on these graphs. "Centrality measure" is one such measure which identifies the influence and importance in the network. As described in [2] centrality can be used to determine the manner in which information is flowing through the network. Measuring centrality requires high computational complexity. Various distributed graph processing systems are developed like Apache Giraph, Hama and GraphLab for effectively processing the network graphs. One such graph processing system is Pregel which was introduced by Google. In [3] the author describes the Google Pregel system as a graph processing system by Google based on Bulk Synchronous parallel (BSP) processing model. In [4] the author talks about this distributed graph processing paradigm. Single machine size graph processing problems cannot handle the large social media data. A single machine cannot handle the large data and may fail during execution. This has "think like a vertex" programming model.

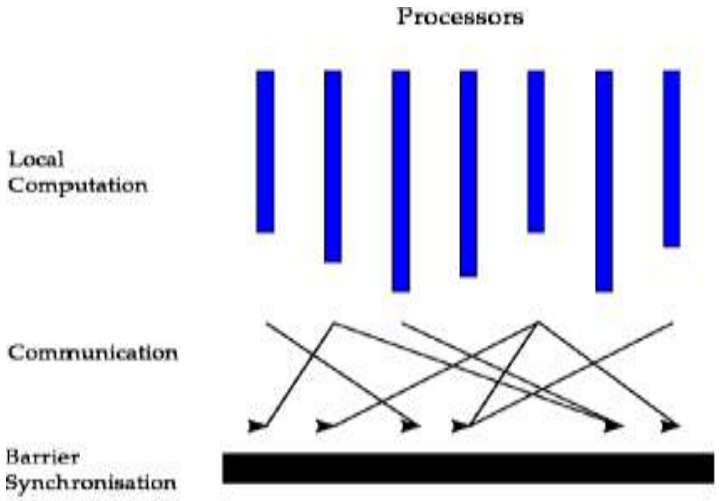

Fig 2: BSP Model

It has a sequence of global supersteps which are individually performed by different nodes in parallel. The outputs of these supersteps are communicated through message passing and a Barrier Synchronizer ensures that communications and processing is done in effective manner.

B. CommunityDetection

A cluster in a graph is depiction of a community which shares ideas. In a cluster multiple paths should be present to pairs of vertices. It should be connected. A cluster is good if it is dense and there are few connections to the rest of the graph [1]. It has maximal clique. Detecting communities require information about topologies ingraph.

"Edge Betweenness" measures the isolation of communities based on the basis of number of shortest path between vertices. It identifies edges that connects communities and removes them isolating the clusters of communities. It requires high computational complexity in large networks. In [5] the author proposes an algorithm which integrated accumulation technique with single source shortest path to compute the edge betweennessquickly.

"Modularity" is the measure of strength of division of networks into modules. More the modularity, the denser is the connection between the modules. Its computation is NP- complete problem but a number of algorithms are present which uses approximations to compute it in reasonable amount of time[1].

Some community detection techniques are:

1. Random Walks: Since the density of internal edges is high in a cluster, a random walker spends more time in cluster. Also the number of consecutive nodes to be followed is high. NetWalk is a random walk algorithm which calculates the proximity ofvertices.
2. Maximum Modularity: Maximize modularity method was proposed by Newman [6] to detect clusters. Here modularity is expressed in terms of Eigen vectors. A greedy approach is employed where vertex groups are joined to form

largercommunities.

3. Spectral Optimized Modularity: Greedy approaches have poor performance when compared to other approaches. In Spectral Optimized Modularity an improved version is proposed by author where Laplacian matrix is replaced by modularitymatrix.

A common scenario which is encountered is of a vertex belonging to more than one cluster. This depicts those users who belong to more than one community. Identifying these types of communities is also a challenge. These are identified by Fuzzy clustering and overlap approach. Techniques are SLPA, OSLOM, Game and COPRA [1].

C. TextAnalytics

Text is one of the most crucial formats of data shared on Social media. Text Analytics describes a set of linguistic, statistical, and machine learning techniques that model and structure the information content of textual sources for business intelligence, exploratory data analysis, research, or investigation [7]. The information on most sites is stored in form of text. Analysing this text has found some useful implementations in Business, Bio Sciences and Social Sciences [8]. Sentiment analysis is done on social media data and based on the results trading decisions are made. Techniques are developed to check spread of diseases and epidemic through social media posts and blogs. Moreover, methods are developed to monitor responses to social media posts, public announcements and speeches.

Text mining refers to extraction of data from text archives. Some examples of text mining tasks are classifying documents into specific topics, grouping documents related to a common topic and finding documents which specify some criteria [9]. Text processing starts with extraction of features which can capture the content of the document. Then different algorithms of supervised and unsupervised clustering methods to classify and cluster the documents are performed. Techniques to reduce the dimensionality of documents are also applied to get more meaningful representation of the features[9].

The features of social media text as listed in [7] are Time Sensitivity, Short Length, Unstructured Phrases and Abundant Information.

Real-time data is a common characteristic. New information is posted on social media sites almost daily and the data is highly time sensitive.

Some sites restrict the length of strings. This leads to a reduced text length which poses a challenge for text analysis. Another limitation to this is that they don't use any external information for processing.

An important difference of social media data from traditional data is the unstructured nature of data. There is also difference in quality. Contents on social media websites could be posted by experts or by a lay man. Differentiating contents on the basis of quality becomes very difficult here. Also, most of the data is in the form of abbreviations and identifying semantic meaning become difficult.

Information is present in abundance and most of the content is noise. Filtering useful data becomes a challenge here. External information, apart from contents posted by users, also acts as a rich source of information. For example, links and tags are used to detect popular events.

The framework of Text data Analytics as given [7] is represented in fig 3.

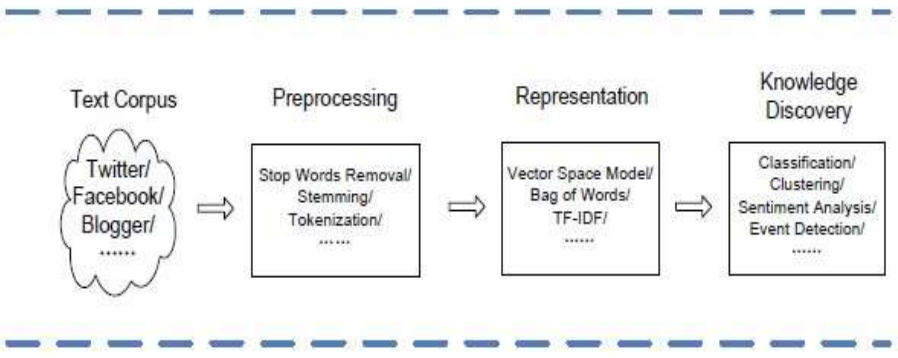

Fig 3: Text Analytics Framework

Text pre-processing is the first part of feature extraction. It makes the input document more consistent and easier for representation. Traditional text pre-processing methods include following phases [7]:

1. Stop word removal: Stop word removal eliminates words using a stop word list. The list contains words are considered more general andmeaningless.
2. Stemming: Reducing derived words to their root form is stemming. For example, "look", "looking", "looked" are represented as "look", so the words with variant forms can be regarded as same feature [7]. Stemming aims to reduce the number of unique words and is often applied in information retrieval [9]. Porter stemming is a popular stemming algorithm.

3. Term Frequency: Counting the frequency of terms is a popular technique to encode documents. Weights are assigned to terms depending on their importance. Inverse document frequency weighs terms on the basis of number of documents which contains it. The extension to this is term frequency inverse document frequency. The formula for which is[7]:

$W_{i,j} = tf_{i,j} * \log(N/df_i)$

Where, $W_{i,j}$ is weight of term i in document j, $tf_{i,j}$ is number of occurrences of term i in document j, N is total number of documents and df_i is number of documents containing term i.

Pre-processing methods vary depending on the type of application. For example, syntax containing words are to be retained in Opinion Mining and NLP, as they need to analyse message from syntactical point of view [7].

Representation of text is also a challenge as data is unstructured. "Bag of words" or "Vector space model" is a popular technique to represent documents. It does not require any form of natural language processing like part of speech tagging and considers the terms as variables. The corpus is represented as term-document matrix. Terms appear on rows and documents in column. Each element $x_{i,j}$ is the frequencyof term i in document j. This could also be weighted frequency [7]. This representation can be expanded to include words as pairs or triplets. This representation is called bigram proximity matrix. A biagram proximity matrix has rows and columns equal to number of words. Each element $x_{i,j}$ is 1 if term j appears after term i [9]. Another approach include representing substrings of documents as feature space and using kernel functions from Support Vector Machines as function for similarity[10].

Once the representation of text documents is done, different clustering, classification and machine learning algorithms can be applied on these documents for knowledge retrieval. Distance is used as a measure of similarity. The most common way to do so is by measuring the cosine of the angle between documents [7]. More the value, more the documents are similar.

The document space consists of thousands of dimensions. In vector space representation, each term in corpus for a dimension. So, for a corpus containing n terms, the dimensionality will be n. Reducing this huge dimensionality is of interest as a lower dimensional space helps in better classification, clustering and visualization. It will also help in removing noise from data and identify the relationships betweendocuments.

Latent Semantic Indexing analysis is a popular technique which uses singular value decomposition for reduction of dimensionality. Another method for dimensionality reduction is Principle Component Analysis which uses eigenvectors of covariance matrix for dimensionality reduction [9].

Multidimensional method uses distances to find how close the documents are to one another [11]. This approach is used to find if the documents which were close in the higher dimensional space are also close in reduced space. Isometric Mapping (ISOMAP) is an extension of multidimensional modelling. It works on the principle that minimal distance along some unknown surface is a better indication of distance as compared to Euclidean distance. Numerical approximations of heat equations are used in the approach of Laplacianeigen maps to model the intrinsic geometry of the data. The strategies to design surface to encode datasets come under manifold learning. Determining number of dimensions to use is another researcharea.

For clustering, algorithms like k-means and agglomerative clustering are applied on represented data. Apart from this a graph based approach is also used for clustering [9]. A bipartite graph is created with documents on one side and terms on other. There is edge between only the document and terms and not between different documents or different terms. Then the clustering becomes a type of graph cut problem. Evaluating cluster quality is an open research problem. Once the documents are clustered evaluating whether the documents are genuinely related is done. Another approach called Suffix Tree Clustering is introduced in [12] which create clusters based on phrases shared between documents. It considers documents as group of strings rather than group of words. Based on proximity information the clusters are created. A combined approach of first finding frequent sets and then clustering is introduced in[13].

Classification is one of the most sought out technique for text processing. In [14] author gives importance of classification techniques in customer satisfaction of web sites. Classification is a type of supervised learning. Here the model is trained using previous knowledge and when a new document is to be classified, this model is used. Methods like Bayes classifier and Artificial Neural Networks can be applied forclassification.

Dealing with Synonyms and Polysemy is a challenge for classification [9]. Synonyms are word with same meaning and Polysemys are words with multiple meanings. Representation of large documents for classification is another challenge. Moreover, handling online streaming of data which is continuously changing is a required.

Applications of text mining as described in [7] are:

1. Event Detection: Event detection is monitoring a data source and capturing the occurrence of an event within the source. Although the data source can be anything from image, audio, video to texts. However, event detection from posts, blogs and new articles is most popular. In any event detection system, a classifier is used to classify tweets into positive and negative cases and then a probabilistic model identifies the target of the event. One application of event detection is identifying "Breaking News" from articles posted on websites. Information isspread over social links and tracking information diffusion patterns is important in event detection.

2. Collaborative Question Answering**:** Social Media also serves as a platform for users to ask questions and experts to give answers to these questions. Through this a question answer archive is created. Whenever a new question is asked by user, the question directory is searched for similar questions using classification techniques and similarity measures. On the

other hand, categorisation of answers is also done. A common parameter for this is expert answers or answers based on userinterests.

3. Social Tagging**:** Instead of sharing documents entirely, documents are tagged to be shared between users. By browsing the tags, relevant documents are located. Designing a tag ranking algorithm is a difficult task as they ignore semantic relationships between key words. Research topics in this field are improving tag recommendations and studying how to utilize social tagging to facilitate otherapplications.

4. Bridging the semantic gap: The semantic relationships between documents cannot be identified unless some external information is provided. Ontologies are used to bridge thisgap.

D. Predictive Modeling

In recent times Predictive Modeling has become a powerful tool. Big Data processing and online streaming of data has given rise to the need of new and innovating forms of predictive analysis. Predictive models are used to make right and quick decision with fewer expenses. The application of predictive modelling is significant in the field of marketing, customer retention, pricing optimisation and fraud prevention [17].

A predictive model is a mathematical algorithm that predicts a target variable from a number of factor variables. Predictive modelling is a name given to a collection of mathematical techniques having in common the goal of finding a mathematical relationship between a target, response, or "dependent" variable and various predictor or "independent" variables with the goal in mind of measuring future values of those predictors and inserting them into the mathematical relationship to predict future values of the target variable [18]. In [19] the author explains an example of CineMatch algorithm which is predictive modelling algorithm used by Netflix .It determines the movie which a user will enjoy based on Movies in a group, Customer rating and combined rating of users. The steps of algorithms described in [19]are:

1. Searches the CineMatch database for people who have rated the same movie - for example, "The Return of theJedi"

2. Determines which of those people have also rated a second movie, such as "TheMatrix"

3. Calculates the statistical likelihood that people who liked "Return of the Jedi" will also like "TheMatrix"

4. Continues this process to establish a pattern of correlations between subscribers' ratings of many differentfilms

This example gives an idea about the steps of predictive modelling which includes: analyse the historic data, identify and quantify relationships between predictive inputs and outcomes and then apply learning to predict outcomes of new cases.

Regression is one of the most famous techniques for predictive analysis. Regression analysis is a form of predictive modelling technique which investigates the relationship between dependent and independent variables. Here, a curve is fit so that the distance between data points is minimised. Other predictive modeling algorithms are generalized linear model, Neural Networks, Genetic Algorithms, Stochastic Gradient Boosted Trees and Support Vector Machines.

E. Information diffusion

Social media plays an important role in spreading information through social links. It is becoming difficult to analyze how the information is spread. The characteristics of diffusion model are [1]:

1. Topological Structure: Graphical analysis where a sub graph of users to whom information is send is analysed.

2. Temporal Dynamics: This analyse the evolution of number of users to whom information is spread over time.

The categories of diffusion model are [1]:

1. Explanatory Models: The aim here is to discover hidden cascade once the activation sequences are collected. These models can build a path that can help users to easily understand how the information has been diffused. NETINT method has applied sub modular functional based iterative model to discover spreadingcascade.

2. Predictive Models: These are based on learning sequences with the observed diffusion pattern. It forms Structure-based model and content-analysis- based model.

The cascade models are [15]:

1. Independent Information Cascade: Here information or decision taken by neighbors is present and based on that it is diffused. There are only two possible states active or inactive i.e. the neighboring node has either decided to be receive

(active) or not receive (in active) information. An example is:

2.

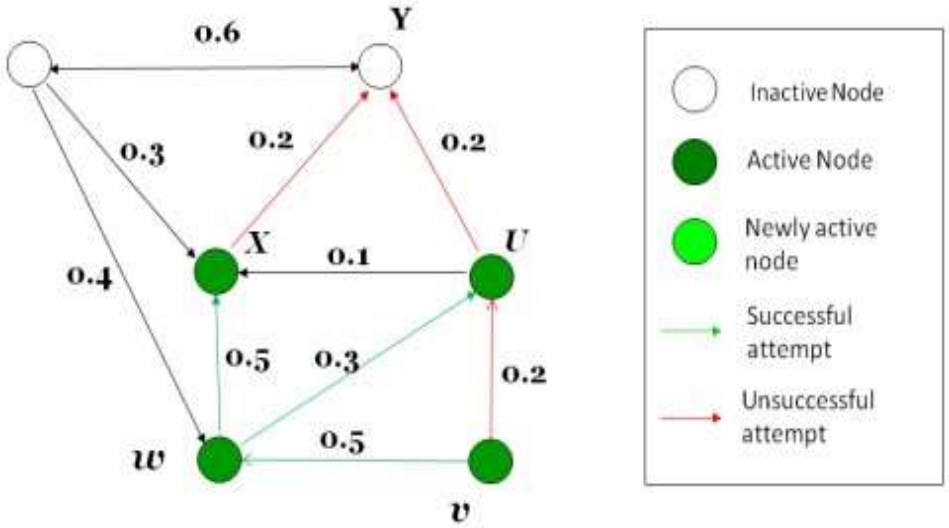

Fig 4: Independent Information Cascade

3. Linear Threshold Model: Here the neighboring node takes information based on some threshold value. This is based on the information about which of the adjacent nodes have received theinformation.

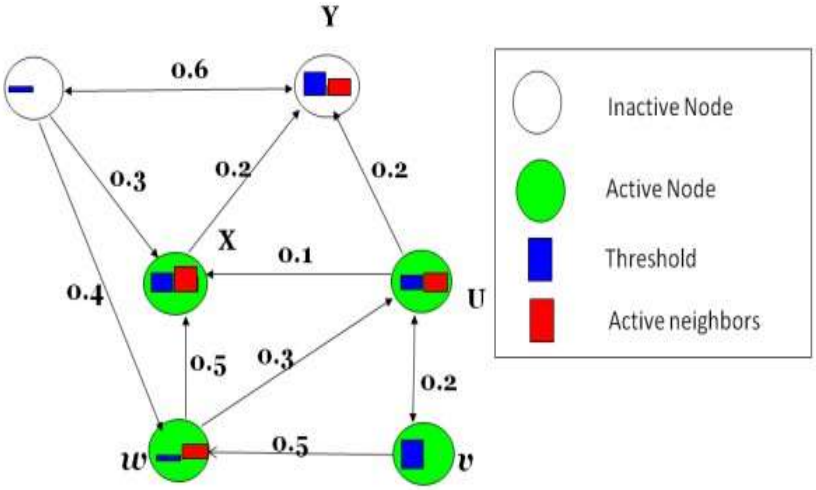

Fig 5: Linear Threshold Model

F. Information Fusion

Social data from various sources is fused. The methods used are [1]:

1. Ontology based fusion: Heterogeneous data, present in different forms is represented in the form of Ontologies. Then, these Ontologies are fused together. It inks open data based on RDF is used as a unified data model for combining, aggregating and transforming data from heterogeneous data resources to build linked datamash-ups.

2. Social Network Integration: It searches the social identity and then finds the best matches between socialidentities.

III. RECENT TOOLS AND FRAMEWORKS

A. Map Reduce:

Map Reduce is one of the most effective methods for parallel processing of big data. It contains two main functions: Map and Reduce. Map Function generates a set of intermediate key/value pairs. Reduce function merges all the intermediate values associated with the same key. The key aspect is that if every map and reduce is independent of each other, then they can run in parallel on different keys and set of data.

Following is an example of Word Count Program in Map Reduce [1]:

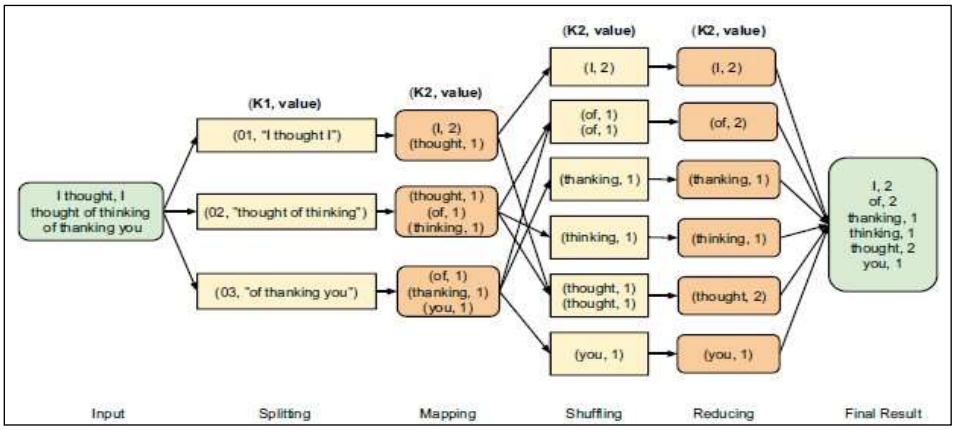

Fig 6: Map Reduce for word count

The steps are:

1. preparing the data :designates the mappers, assigns input value on which the mapper will work on and provides thedata

2. Map(): apply map function to local data and writes the output to temporary storagespace.

3. Shuffle(): provides the K2 key value each processor should work on and provides all data relating to a particular node ondata.

4. Reduce(): each node performs thereduce function once on each K2 value.

The Limitations of Map Reduce are [1]:

1. Maps and Reduces should not depend on any data generated in the same MapReducejob
2. The order in which maps and reductions run is notdefined
3. Ineffective when similar searches are run repeatedly
4. Reduce operations do not take place until all maps have beencompleted
5. Large data generates smaller finalvalues

B. ApacheHadoop

It is an Open Source software framework for distributed storage and processing [1]. The Storage area is Hadoop Distributed File System or HDFS. It spreads multiple copies of data across different machines. If a machine is busy, another machine can use the copy of data. The processing is done by MapReduce. JobTracker is a job scheduler to track which map reduce jobs are executing, schedules individual maps, reduces individual reduce operations to specific machines and monitors success and failures of thesemachines.

C. ApacheSpark

It is an Open Source cluster computing framework containing Machine Learning tools. Although Spark is similar to Hadoop, it has some useful differences from it. It was designed for specific kind of workloads in cluster computing, those that reuse a working set of data across parallel operations [1]. In-memory cluster computing is used where datasets are caches in-memory to reduce the latency of data. Resilient Distributed Database is used where a manager tracks the data as lineage and uses it to recover the datalost.

D. Distributed ProgrammingSystems

1. Apache Pig: It is abstraction over MapReduce. It is a tool or platform which is used to analyse large set of data representing them as data flows. Pig Latin provides various operators which programmers can use to read write or process the data. It is SQL like language and is friendly to programmers familiar with SQL. Programmers write the scripts in Pig Latin and Pig Engine converts these scripts to map reduce jobs.

2. Apache Storm: It is a real time big data processing system, designed to process vast amount of data in fault tolerant and horizontally scalable method [16]. All kind of operations on real time data can be performed. Online Analytics, real time machine learning, continuous computation and distributed RPC. Master worker paradigm is used here and it is claimed to have the highest data ingestionrate.

3. ApacheHDFS
4. Stratosphere: It is a general purpose cluster computing framework which has more transformations than MapReduce. It analyse data using data flow graphs. YARN can be used as cluster manager.

E. Distributed DataModel

1. Graph Model: Data can be represented in the form of graph. Apache Cassandra and Apache Giraph are datamodels.
2. Document Model: MongoDB represent data in the form of documents. It is an open source document oriented database system. It provides high performance, high availability and automatic scaling. It stores data in JSON-like documents containing field and value pairs. Index system supports faster query and includes keys from embedded documents and arrays. It also allows users to distribute data across cluster ofmachines.

IV. SOCIAL BASEDAPPLICATIONS

Social Big Data analysis finds applications in various fields like:

1. Marketing: Advertisements on social Media can increase number of visitors and profit. Opinions about a brand can be extracted from vast number of users and it can be used for marketing. Viral Marketing can be done. Heat diffusion patterns can be applied for spreading the information about a brand.

2. Criminal Activities: Criminals have repetitive pattern of behavior. Filtering reports and identifying patterns can help in providing useful information to analyze crime trends. Geospatial distribution - identify the graphs storing data for a particular region. In Hotspot mapping data from past can be used to extract information about future. Each crime event is represented as a spot and point mapping techniques are used to identify the patterns. Fraud detection can be done using classification techniques like Bayesian beliefnetwork.

3. Epidemic Intelligence: Information posted on twitter about an illness, their contacts are available, by GPS the location can be known. Text mining techniques are used like named entity recognition, text classification and terminology detection and extraction. Some words can mean different things and one disease can have different symptoms. Ontologies can help to implement human understandings to achieve certain level ofaccuracy.

V. CONCLUSION, CHALLENGES AND RESEARCH AREAS

With the large number and rapid growth of socialmedia systems and applications, social big data has become an important topic in a broad array of research areas. The aim of this research was to provide an overview of the methods used in the processing of Big Social Data. It covers some frameworks and algorithms in a nut shell. Big data requires innovative form of processing and a survey of the techniques used for it is provided in this paper. Introduction of few frameworks is also provided.

Big Data processing contains few open challenges like Privacy, Streaming Online Algorithms and Fusion of data. Risk of data leaks increases with Advanced Analytics. Analyzing massive amount of online data is a challenge. It requires high scalability in memory and time consumption. Fusing multiple features of multimedia objects is a challenge. Some of the areas of research in social network analysis as given in [20] are Linkage and Structural Analysis, Content based Analysis, Statistical Analysis of Social Networks, Random Walks and their application in social networks, Community Detection in Social Networks, Node Classification of Social Networks, Evolution in Dynamic Social Networks, Social Influence Analysis, Expert Discovery in Network, Link Prediction in Social Network, Privacy in Social Network, Visualizing Social Networks, Data Mining in Social Media, Text Mining in Social Networks, Integrating Sensors and Social Networks, Multimedia Information Network, Analysis in Social Media and SocialTagging.

REFERENCES

[1] G. Bello-Orgaz et al., Social big data: Recent achievements and new challenges, Information Fusion (2015),http://dx.doi.org/10.1016/j.inffus.2015.08.005
[2] Stephen P. Borgatti, Centrality and Network flow. Presented at Sunbelt International Social Networks Conference 2002, in NewOrleans.
[3] GrzegorzMalewicz et al., Pregel: A system for Large-Scale Graph Processing, SIGMOD'10, June 6–11, 2010, Indianapolis, Indiana,USA. Copyright 2010 ACM 978-1-4503-0032-2/10/06
[4] Bulk Synchronous Paper, Available at: http://www.cse.unt.edu/~tarau/teaching/parpro/papers/Bulk%20synchro nous%20parallel.pdf
[5] UlrikBrandes, A Faster Algorithm for Betweenness Centrality, Journal of Mathematical Sociology 25(2):163-177,(2001).

[6] M.E.JNewman,ModularityandCommunitystructureinnetworks, Phys. Rev. E 69 (6) (2004) 066133+ , doi:10.1103/physreve.69.066133
[7] Xia Hu, Huan Liu, Text Analyticsin Social Media, Mining Text Data, DOI 10.1007/978-1-4614-3223-4_12
[8] BogdanBatrica et al., Social media analytics: a survey of techniques, tools and platforms. Available at: https://link.springer.com/article/10.1007/s00146-014-0549-4
[9] Jeffrey L. Solka, Text Data Mining, Theory and Methods. Available at: https://arxiv.org/pdf/0807.2569.pdf
[10] HumaLodhi et al., Text Classificaiton using String Kernels, Journal of Machine Learning Research 2 (2002)419-444
[11] Chapman and Hall, Multidimensional Scalling. Cox, T. and Cox, M. (2000).
[12] Oren Zamir and Oren Etzioni, Web Document Clustering: A Feasibility Demonstration. Available at:https://dl.acm.org/citation.cfm?id=29095
[13] Rajendra Kumar Roul et al., An effective Web Document Clustering for Information Retrieval. Available at: https://arxiv.org/ftp/arxiv/papers/1211/1211.1107.pdf
[14] ChoudurLakshminarayan et al., Improving Customer Experiences via Text Mining. Availableat: https://link.springer.com/chapter/10.1007/978-3-540-31970-2_23
[15] Dr. Ding-Zhu Du. Available at:www.utdallas.edu/~dzdu/cs6301/unit2- 2.ppt
[16] Apache Storm official Website:http://storm.apache.org/
[17] Predictive Analytics, White Paper by CGIGroup.
[18] David A. Dickey, N. Carolina State U., Raleigh, NC; Introduction to Predictive Modeling with Examples, SAS Global Forum 2012, Paper 337-2012
[19] Julie Chambers, Predictive Modeling. Available at http://www.crconline.ca/2012_presentations/12-06%20Predictive%20Modeling.pdf
[20] Charu C. Aggarwal, Social Network Data Analytics, IBM T.J Watson researchCentre