Extracting Knowledge from Large Social Key Valued Data

Palvadi Srinivas Kumar[1] and Siva Ratna Sai[2]
[1,2]Research Scholar,Department of Computer Science & Engineering,
Sri Satyasai University of Technology and Medical Sciences,Sehore, Madhya Pradesh
[1]*srinivaskumarpalvadi@gmail.com*[2]*sivaratnasai@gmail.com*

ABSTRACT

With advances in computer and information technology, large amount of different types of valuable data are gathered and generate in the present time of huge information from a large range of sources of availability of information of various veracities at a high speed. Throughout late years, a couple of frameworks and applications have built up the utilization cloud, structure and organization enlisting to direct and analyze huge data with a specific end goal to help data science (e.g., identifying and extracting data). In this paper, we display an answer for social computing and social network analytics so as to provide services and support to big information mining of fascinating examples from huge interpersonal organizations that are stored in key-value databases.

DOI : https://doi.org/10.30991/IJMLNCE.2017v01i02.003

Kewords

Data Mining, Big Data, Social Computing And Social Network, Large-Scale Cloud Applications, Grid Computing, Key-Value Database, Key-Value Store

1. Introduction

Due to the rapid increase of technology and gadgets the usage of internet and social media is increased tremendously consider an example in internet in one minute many things are going on such as 15,000 GIF’s are transferred via Messenger, nine lakhs of logins are going on in face book,4.1 million videos are viewed in youtube, three lakhs forty two apps are being downloaded by users in play store and in I tunes, four lakhs fifty two thousand tweets are being transferring by users,156 million E-mails are sent,120 new linked accounts were being created,3.5 million search queries are processing by Google,$751522 dollars are being spent in online shopping[1]and many more are happening knowingly and unknowingly. Consider the enormous data being collected from hospitals from patient monitoring system, climatic information, cars, and airlines. Here we are surely living in a challenging and interesting era with cloud computing and big data. Big data and cloud computing playing a crucial role for organizations and government sectors for storing of large amount of information future estimation and analysis of data by the help of Present available data. Additionally service providers tracking the mobile numbers of the customers who has registered in there site, made purchases and those who are giving rating of their business by phone for growing their marketing efforts, estimating the future from past collected information, also, expanding consumer loyalty. The confuse between the requests of the huge information administration and the abilities that present DBMS has achieved more prominent request. The three Vs, for example, (volume, assortment, and speed) of huge information each suggests particular part of basic inadequacies of present DBMS.

Big data name itself says collection of large amount of data which can have ability to make operations on the data, manage the data Huge information can be characterize by its volume, assortment and speed. Volume is only the extent of the information which defines by volume .Variety means how many types of raw data is available in this content.

2. Background And Related Work

2.1 Background

In this context, we present some groundwork information about (A) big data and (B) social network analytics.

(1)Big Data

It is a combination of vast information which includes Text, Image, Audio, Video, Graphics and Animation. So the overall data is presented in form of structured and unstructured data. The structured data has processed and generated result will have meaningful information. The unstructured data is not properly defined about the type of data or category of data it belongs to. The combination of these types of data is combined and called Database. The collection of such databases form as Data Warehouse the data ware house includes patient medical data, weather reports, airlines information, Tele communication, life sciences, social media [2-6]are the different sources for generating data from this era Big data is formed. high volume of information is used for collecting storing and extracting data[7] many of the applications use the concept of Grid computing, clustered computing or Cloud Computing

Grid computing: It is formed by distributed or parallel heterogeneous loosely coupled computers which help to manage, collect and present data

Clustered Computing: Group of parallel or distributed computers combined by a high speed networks such as local area network

Cloud Computing: is a synonym of distributed network where users can share multiple resources over the internet cloud computing comprises of service models like software, platform and infrastructure

(2) Social network analysis

Various websites and social networking sites like facebook, twitter is increasing tremendously day by day so the lot of multimedia data is storing in the servers anomalously. So that the social network usage are increased more in 2016 the results are here:

i. 313 million monthly active Twitter users[10]
ii. 1.79 billion month to month dynamic Face book users [11]; and
iii. More than 467 million enlisted LinkedIn clients, which incorporate no less than 13 million clients in Canada [12].

Some social networks user follow/un-follow of users where as some use like/dislike for voting purpose of users or posts here in every social sites their main intention is to get the data from the users which is more relevant and getting feedback from customers it is a type of business and target the customers.

3. Related Work

Versatility is at the center difficulties with monstrous information. The distributed computing innovation conveys the chief gives stage to value the required quantifiability with incontestable physical property and correspondence limits. Different remarkable makes an endeavor is started to exploit gigantic information handling structures. Map Reduce and its appropriated portrayal structure, GFS, address the fundamental undertakings done in the midst of this line. From the information mining perspective, mining colossal data has opened a couple of new troubles and openings. regardless of the possibility that monstrous learning bears bigger cost (i.e., concealed data and a considerable measure of significant experiences), it conveys enormous difficulties to remove these concealed data and bits of knowledge from huge information since the built up technique finding and information handling from run of the mill datasets wasn't intended to and can't function admirably with huge measure of information. The cons of present information preparing strategies once connected to vast learning are fixated on their lacking quantifiability and correspondence. By and large, In existing work the information preparing methods experience decent troubles after they are expected to deal with the exceptional non-consistency, volume, speed, protection, precision, and trust returning conjunction with more learning and extensive measure of information handling. after existing work by applying gigantic information preparing and dispersed capacity, arranging inventive mining procedures upheld new edge works with the possibility to with progress conquer the aforementioned difficulties would modification be able to and reshape the long keep running of the data mining innovation. huge volume needs similarly decent quantifiability correspondence that are on the far side the fitness of the present DBMSs; the considerable style of learning sorts of colossal adapting noticeably unfits the restriction of the close technique diagram of current information structures; the speed request of enormous data process demands coextensive period control that again is course on the far side wherever current DBMSs may reach. The constrained availability of current DBMSs slaughters the speed request of tremendous gaining from yet one more point. To beat this quantifiability trial of tremendous data, many makes and undertaking are made on manhandling enormous data planning models. The basic such endeavor was made by Google. Google made a programming model named Map Reduce [13] that was and also the Google grouping framework , a disseminated order framework wherever the data will be basically parceled off more than a large number of hubs in an exceedingly bunch. Afterward, Yahoo relates degreed elective monstrous firms made an Apache ASCII content document form of Google's Map Reduce system, alluded to as Hadoop Map Reduce. It utilizes the Hadoop Distributed characterization framework which relates degree open supply form of the Google's GFS. The Map Reduce structure grants clients to plot 2 capacities, guide and scale back, to strategy sizable sum learning sections in parallel a ton of particularly, in Map Reduce, the info is part into a curiously large arrangement of key-esteem matches first; at that point the guide perform is named and forked into a few occurrences in the meantime procedure on the enormous key-esteem sets. regardless of everything information passages are handled, a fresh out of the plastic new arrangement of key-esteem sets are made, at that point the scale back perform is named to gathering/consolidate the made esteems upheld normal keys. in order to coordinate/bolster the Map Reduce registering model, Google built up the Big Table – a conveyed stockpiling framework intended for overseeing organized learning. Big Table will scale well to a truly goliath measure: petabytes of information crosswise over a huge number of exchange products disjoins . inside a similar soul, Amazon made generator , that is furthermore a key-esteem consolidate capacity framework. The Apache ASCII content document.

Gathering acted quickly once more, made HBase – relate degree open source type of Google's BigTable composed on high of HDFS and prophetess – relate degree ASCII content record variation of Amazon's generator. Apache Hive is relate degree open supply data distribution center framework outlined on high of Hadoop for questioning and examining records keep in HDFS utilizing a simple order language referred to as HiveQL. Hadoop isn't the only one; it's option candidate stages. of these stages do not have a few amenities existing in DBMSs. some of the contenders enhanced existing stages (for the most part on Hadoop), et al. thought of a current framework style. Nonetheless, the vast majority of those stages are still in their early stages. as a case, BDAS, the Berkeley information Analytics Stack [15], is relate degree ASCII content record learning investigation stack created at UC Berkeley AMP Lab for figuring and breaking down which incorporates the ensuing primary parts: Spark, Shark,, and Mesos. Start might be a rapid group framework that performs computations in memory and might surpass Hadoop by up to 100x. Shark may be an immense scale data examination structure for Spark that gives a united engine running SQL request, great with Apache Hive. Shark will answer SQL request up to 100x speedier than Hive, and run unvarying machine learning estimations up to 100x snappier than Hadoop, and might get over missing the mark mid-inquiries at between times seconds [14]. Mesos may be a bundle executive that may run Hadoop, Spark and choice structures on a capably shared pool of figure center points. ASTERIX [15] is learning concentrated limit and handling stage. Some conspicuous drawbacks of Hadoop and alternative relative stages, e.g., single system execution, inconveniences of future help, unskillfulness in impulse learning up to request and besides the accidentalness of record limits, are genuinely overcome in ASTERIX by researching runtime models charges by parallel information structure execution engines. In ASTERIX, the open programming group stack is stratified in an exceedingly completely sudden strategy that it sets the information records at astoundingly terrible layer, empowering a more lifted sum vernacular API and no more vital. despite the fact that most of the colossal data organization and process stages are (or are being) created to fulfill business needs, SciDB is relate degree open supply learning organization and examination (DMAS) PC code for data genuine intelligent applications like cosmology, earth remote recognizing and surroundings discernment and showing. The capability among SciDB and choice stages is that SciDB is inferred reinforced the beginning of group programming (i.e., course of action data) wherever colossal data is layout as assortments of articles in one-dimensional or multidimensional zones. SciDB is expected to help fuse with unusual state essential vernaculars, figuring’s, and to an extraordinary degree goliath sizes of learning

Data Mining Process

It is the process of extracting the user relevant information from large sets of data which can be helpful for summarizing the results. It helps to analyze the data from different categories, fields based on the given query and search operation in relational database management system will be performed in following manner:

This procedure comprises of following strides as appeared in figure 3.1:

i.Expelling, change, stacking the trade data onto the data appropriation focus structure.

ii. Securing and managing the data in a multidimensional database structure.

iii. Giving data access to business examiners and information development specialists.

iv. Dismember the data by application programming.

v. Show the data in a significant association, for instance, a graphical or table depiction

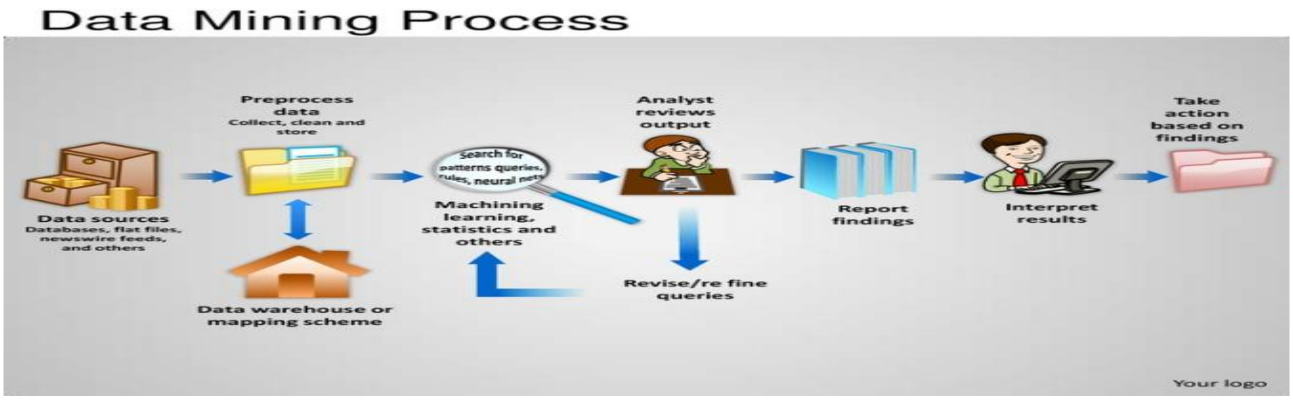

Data Mining Process
Preprocess data
Collect, clean and store
Analyst reviews output
Take action based on findings
Search for patterns queries, rules, neural nets
Data sources
Databases, flat files, newswire feeds, and others
Machining learning, statistics and others
Report findings
Interpret results
Revise/re fine queries
Data warehouse or mapping scheme
Your logo

Figure 1: Data Extraction Procedure

The fundamental motivation of information mining is to get the demand from client and process it by extricating in light of the client question and discover the relations between client query item with RDBMS Comparing with the outcomes got from mining the customary datasets, disclosing the tremendous volume of interconnected heterogeneous enormous information can possibly boost our insight and bits of knowledge in the objective area**.**

4. Existing And Proposed Work

Social networking sites like twitter has attracted million of users by giving updated information to the users every second, Here they are performing Social Network Analysis(SNA) on the data which is available on the server with respect to the Active Popular Users(APU).Here by using Active Popular

Users by taking the active users we are considering the name entity recognition. Recently SNA has gained more popularity due to large number of users attacked due to micro blogging on twitter which provides users to tweet publicly tweets with a 140 characters of message i.e.;Tweets .A user wants to follow APU's for receiving tweets because of their popularity.

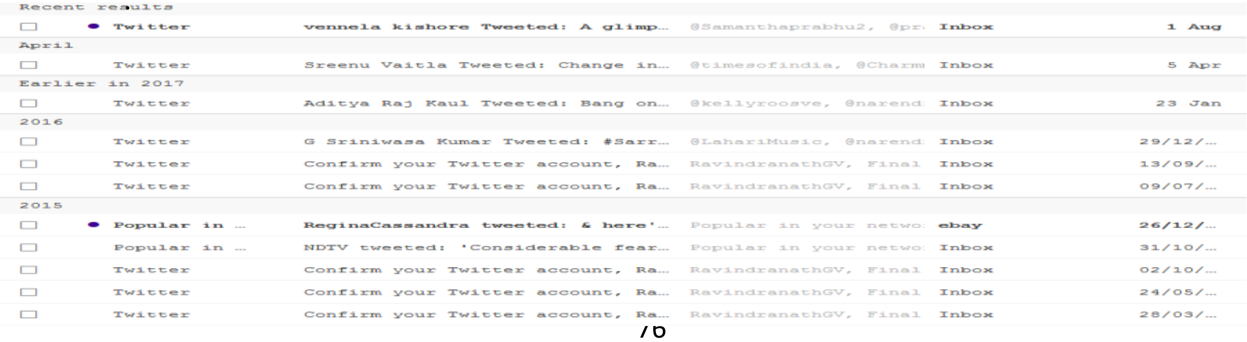
Several algorithms are proposed for SNA since past few years but outlier Detection (OD) placed a major role but this algorithm suffered from memory exceptions which leads to inaccurate analysis so by considering the categories like user reviews, active users we have proposed APu which highlights on

i. Social network analysis on follow/Subscribe relationship and
ii. Social network analysis on mutual friendships

By these we find the popularity of the users by selecting Top-N- revalent users from a big user pool and filter outs less popular ones these process of collecting is done from collecting historical data/information .So to overcome this problem here in this paper[16] they propose a Map Reduce-based SNA method that comprises of Key Value Pair model analysis that is used to learn APU's, and it is optimized in a layer wise fashion to handle big data using Map Reducer's. They have concentrated on applying affiliation lead or successive example mining methods to break down and mine huge interpersonal organization information for interdependencies or associations among social elements in a major informal organization.

The Map Reduce-based SNA demonstrate is a pile of processors, which is an acclaimed as a profound learning model. It utilizes mappers and reducers as building pieces to make a profound information analyzers. The client tweet connections are innately considered in this displaying. Layer astute preparing of Tweets Data regarding User's, Follower's, Associations offers much better comprehension of User Activities.In expansion, it shows that the proposed technique for APU finding has better execution looked at than little information strategies like Clusters or Outlier Detections. A reasonable utilization of such preparing of Big Social Network Data yields the likelihood of separating more prominent clients, for example, the accompanying

:



Recent results						
□	●	Twitter	vennela kishore Tweeted: A glimp...	@Samanthaprabhu2, @pr	Inbox	1 Aug
April						
□		Twitter	Sreenu Vaitla Tweeted: Change in...	@timesofindia, @Charm	Inbox	5 Apr
Earlier in 2017						
□		Twitter	Aditya Raj Kaul Tweeted: Bang on...	@kellyroosve, @narend	Inbox	23 Jan
2016						
□		Twitter	G Sriniwasa Kumar Tweeted: #Sarr...	@LahariMusic, @narend	Inbox	29/12/...
□		Twitter	Confirm your Twitter account, Ra...	RavindranathGV, Final	Inbox	13/09/...
□		Twitter	Confirm your Twitter account, Ra...	RavindranathGV, Final	Inbox	09/07/...
2015						
□	●	Popular in ...	ReginaCassandra tweeted: & here'...	Popular in your netwo	ebay	26/12/...
□		Popular in ...	NDTV tweeted: 'Considerable fear...	Popular in your netwo	Inbox	31/10/...
□		Twitter	Confirm your Twitter account, Ra...	RavindranathGV, Final	Inbox	02/10/...
□		Twitter	Confirm your Twitter account, Ra...	RavindranathGV, Final	Inbox	24/05/...
□		Twitter	Confirm your Twitter account, Ra...	RavindranathGV, Final	Inbox	28/03/...

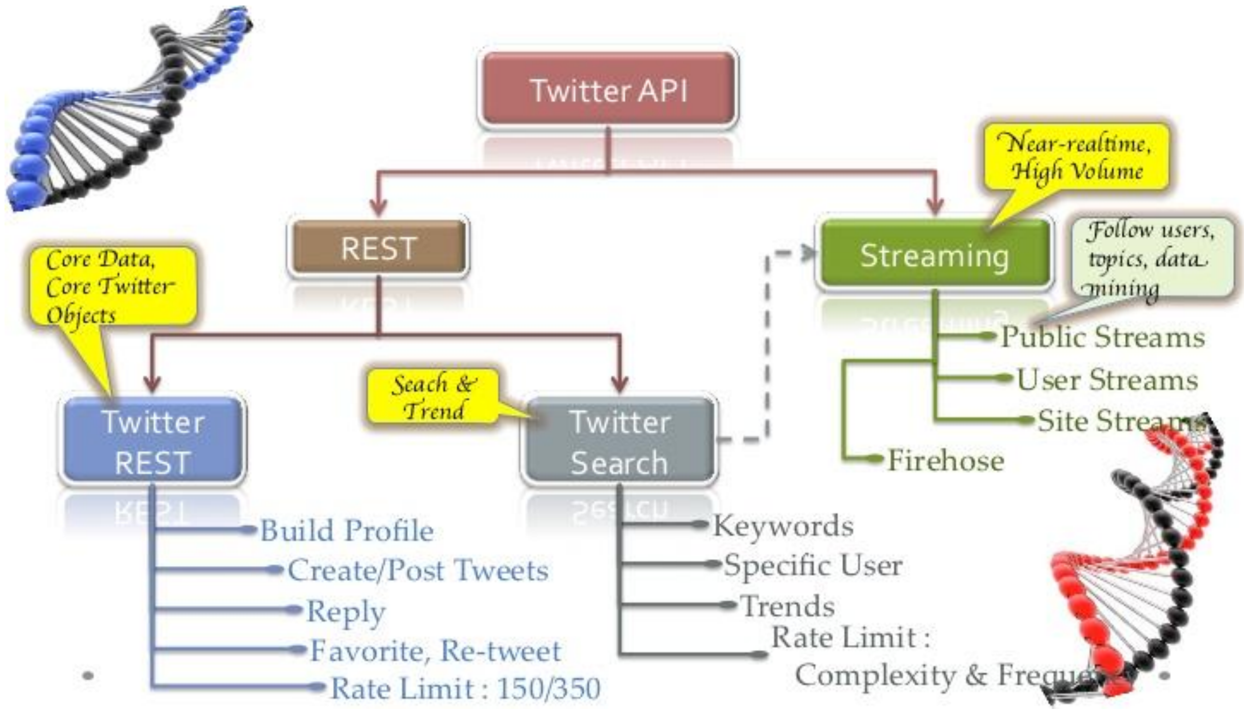


Figure 2: Extracting Data Analysis from Twitter Stream Analysis

Based on the work in the paper [16] they have previously we utilize the different forms of relations among users which improve the accuracy of finding APU's on a given big tweets dataset. Identifying opinion from the different servers from different applications based on name entity is also a greater task so it leads to opinion entity and named entity recognition. For instance most of the users show favor on the tweets based on cosmetics, dicor every user uses many tweets but these are more popular. Expertise are finding information by various users by the help of query factor and finding the similarity of queries in various enterprises. Prior approaches used authenticity of APU's based on tweets & followers only, where as we consider other factors such as the tweets itself and rumor monger aspect. We consider the likeness between clients' distributed tweets and the given question; and furthermore the skill scores of clients on a given subject in Twitter.

Using the information sort we assess the probability of each customer being a master on a given topic, for that we propose an imaginative Semi-Supervised Graph-based Ranking strategy, called SSGR, to figure the overall master of customers on a given point, by utilizing particular sorts of relations in Twitter Lists and follower outlines.

SSGR phases include the following steps:

Utilizing the data sort we gauge the likelihood of every client being a specialist on a given subject, for that we propose an innovative Semi-Supervised Graph-based Ranking approach, called SSGR, to enlist the overall master of customers on a given point, by using particular sorts of relations in Twitter Lists and disciple outlines.

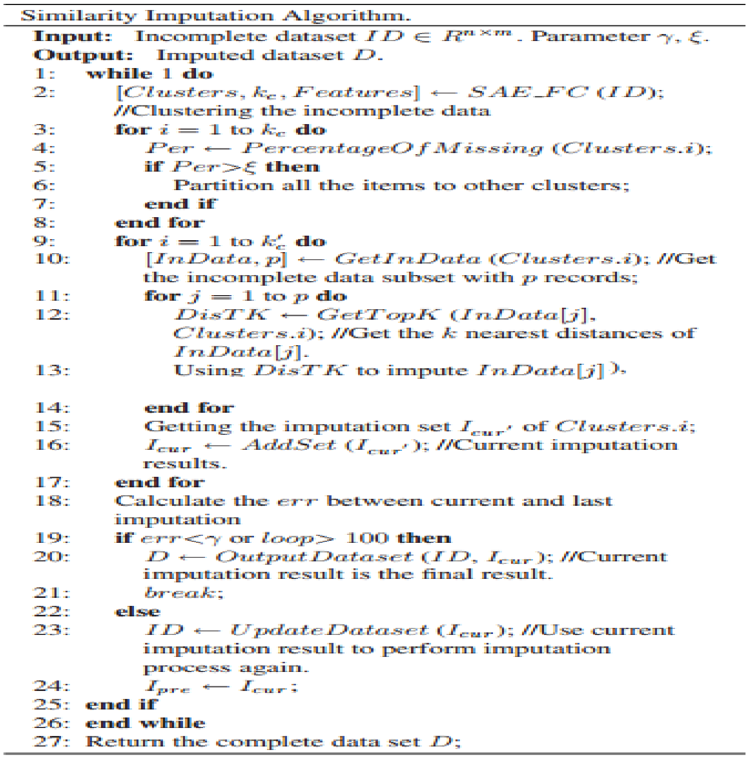
i. A standardized Laplacian regularize term to smooth the positioning of clients and records on three distinctive subject particular diagrams; and

ii. A misfortune term to guarantee the worldwide specialist of clients is as per the astuteness of Twitter swarms.

iii. Based on the processed positioning scores got by above calculations, we select the best N important clients for any given point (specialists) and sift through talk mongers to acquire a productive master discoverer arrangement in smaller scale online journals.

5. Algorithm and Result analysis

Here we have used Similarity imputation algorithm for finding out the similarity coincidence between the tweets and focused on most users searched tweet based on user entered search keyword in query.



```
Similarity Imputation Algorithm.
Input: Incomplete dataset $ID \in R^{n\times m}$. Parameter $\gamma$, $\xi$.
Output: Imputed dataset $D$.
while 1 do
    $[Clusters, k_c, Features] \leftarrow SAE\_FC\ (ID)$;
    //Clustering the incomplete data
    for $i = 1$ to $k_c$ do
        $Per \leftarrow PercentageOfMissing\ (Clusters.i)$;
        if $Per > \xi$ then
            Partition all the items to other clusters;
        end if
    end for
    for $i = 1$ to $k_c'$ do
        $[InData, p] \leftarrow GetInData\ (Clusters.i)$; //Get
        the incomplete data subset with $p$ records;
        for $j = 1$ to $p$ do
            $DisTK \leftarrow GetTopK\ (InData[j],$
            $Clusters.i)$; //Get the $k$ nearest distances of
            $InData[j]$.
            Using $DisTK$ to impute $InData[j]$ ).

        end for
        Getting the imputation set $I_{cur'}$ of $Clusters.i$;
        $I_{cur} \leftarrow AddSet\ (I_{cur'})$; //Current imputation
        results.
    end for
    Calculate the $err$ between current and last
    imputation
    if $err < \gamma$ or $loop > 100$ then
        $D \leftarrow OutputDataset\ (ID, I_{cur})$; //Current
        imputation result is the final result.
        $break$;
    else
        $ID \leftarrow UpdateDataset\ (I_{cur})$; //Use current
        imputation result to perform imputation
        process again.
        $I_{pre} \leftarrow I_{cur}$;
end if
end while
Return the complete data set $D$;
```

Figure3: Similarity Algorithm

6. Conclusion

The task of sentiment analysis, particularly within the domain of micro-blogging, continues to be within the developing stage and much from complete. Thus we tend to propose a handful of concepts that we tend to feel are value exploring within the future and will lead to any improved performance. Right now we've got worked with solely the terribly simplest uni-gram models; we will improve those models by adding additional data like closeness of the word with a negation word. we tend to may specify a window before the word (a window may for instance be of two or three words) into account and therefore the impact of negation is also incorporated into the model.

7. References:

[1] What happens in 1 minute on the internet 2017, http://www.visualcapitalist.com/happens-internet-minute-2017/

[2] P. Braun, J. J. Cameron, A. Cuzzocrea, F. Jiang, and C. K.Leung, “Effectively and efficiently mining frequent patterns from dense graph streams on disk,” in Proc. KES 2014, 338–347

[3] A. Cuzzocrea, Z. Han, F. Jiang, C. K. Leung, andZhang, “Edge-based mining of frequent subgraphs frgraphstreams,” in Proc. KES 2015, pp. 573–582

[4] F.Jiang, C. K. Leung, and A. G. M. Pazdor, “Web pagerecommendation based on bitwise frequent pattern mining,”in Proc. IEEE/WIC/ACM WI 2016, pp. 632–635.

[5] Agrawal D., Bernstein P., Bertino E., “Challenges and Opportunities With big data A Community White Paper Developed by Leading,Researchers Across the United States”, 2012.

[6] R. K. MacKinnon and C. K. Leung, “Stock price prediction inundirected graphs using a structural support vector machine,”in Proc. IEEE/WIC/ACM WI-IAT 2015, vol. 1, pp. 548–555.

[7] S. Madden, “From databases to big data,” IEEE InternetComputing, 16(3), 2012, pp. 4–6.

[8] M. Chetty and R. Buyya, “Weaving computational grids:how analogous are they with electrical grids?” Computingin Science & Engineering, 4(4), 2002, pp. 61–71.

[9] A. Ros`a, L. Y. Chen, and W. Binder, “Predicting and mitigatingjobs failures in big data clusters,” in Proc. IEEE/ACMCCGrid 2015, pp. 221–230.

[10] https://about.twitter.com/company

[11] http://newsroom.fb.com/company-info/

[12] https://press.linkedin.com/about-linkedin

[13] NewVantage Partners: Big Data Executive Survey(2013), http://newvantage.com/wpcontent /uploads /2013/02/ NVP-Big- Data-Survey-2013-Summary-Report.pdf

[14] Xin R.S., Rosen J., Zaharia M., Franklin M., Shenker S., Stoica I., “Shark: SQL and Rich Analytics at Scale”, ACM SIGMOD Conference,2013.

[15] Knowledge Discovery from Big Social Key-Value Data ,2016 IEEE International Conference on Computer and Information Technology by Carson K. Leung