RESEARCH ARTICLE OPEN ACCESSOPEN

Frequent Item set Mining of Big Data for Social Media

Roshani Pardeshi,Prof. Madhu Nashipudimath
Pillai Institute of Engineering and Technology, New Panvel, India
Pillai Institute of Engineering and Technology, New Panvel, India

ABSTRACT
Big data is a term for massive data sets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results. Bigdata includes data from email, documents, pictures, audio, video files, and other sources that do not fit into a relational database. This unstructured data brings enormous challenges to Bigdata.The process of research into massive amounts of data to reveal hidden patterns and secret correlations named as big data analytics. Therefore, big data implementations need to be analyzed and executed as accurately as possible.
The proposed model structures the unstructured data from social media in a structured form so that data can be queried efficiently by using Hadoop MapReduce framework. The Bigdata mining is essential in order to extract value from massive amount of data. MapReduce is efficient method to deal with Big data than traditional techniques.The proposed Linguistic string matching Knuth-Morris-Pratt algorithm and K-Means clustering algorithm gives proper platform to extract value from massive amount of data and recommendation for user.Linguistic matching techniques such as Knuth–Morris–Pratt string matching algorithm are very useful in giving proper matching output to user query. The K-Means algorithm is one which works on clustering data using vector space model. It can be an appropriate method to produce recommendation for user.
Index Terms: Big data, K-Means Clustering, Frequent Itemset Mining, Linguistic Matching, Recommendation

I. INTRODUCTION

Social media is widely used communication media. Site such as Twitter, You tube, Facebook, LinkedIn are frequently used by people [1]. It includes unstructured data from email, documents, pictures, audio, video files, and other sources. To manage such massive unstructured data and construct an efficient, user-friendly structured data, both context and usage pattern of data items are used [2]. There has been an unprecedented increase in the quantity and variety of data generated worldwide. According to the IDC's Digital Universe study, the world's information is doubling every two years and is predicted to reach 40ZB by 2020. Such vast datasets are commonly referred to as "Big Data" [3]. Big data

characterized by 3Vs: the extreme volume of data, the wide variety of types of data and the velocity at which the data must be must processed. Big data doesn't refer to any specific quantity. Big data is a term used when speaking about petabytes and Exabyte's of data, which cannot be integrated easily. Hadoop framework is used to deal with such massive data. Hadoop works on MapReduce framework. MapReduce developed by Google along with Hadoop distributed file system is exploited to find out frequent itemset from Big Data on large clusters. MapReduce framework has two phases, Map phase and Reduce phase. Map and reduce functions are used for large parallel computations specified by users. Map function takes chunk of data from HDFS in (key, value) pair format and generates a set of (key', value') intermediate (key, value) pairs [1].

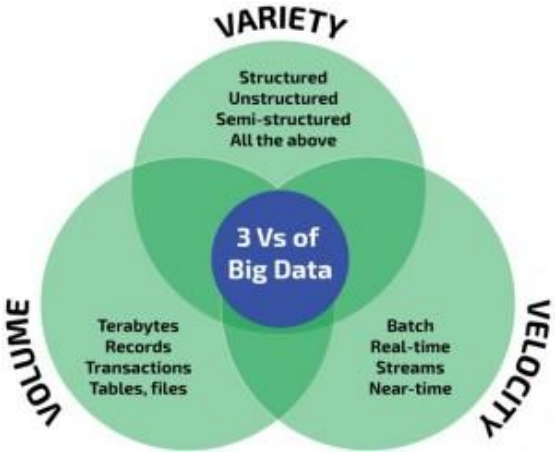

Figure 1 : 3 vs of Big Data

Data mining is an essential technique to extract information from large database. For competitive business world it is very necessary to search through this huge data to grasp useful information. To identify the customer behavior, their choices, to know there feedback for product etc. big data has to be mined by using appropriate mining methods. Frequent itemset mining is method to identify the itemset which appear frequently in a dataset [1].

Frequent Itemset Mining Methods
There are different frequent itemset mining methods such as

- Association, correlation analysis,
- Sequential, Structural pattern.
- Classification.
- Clustering. etc.

Association, correlation analysis It is an important data mining model used to find the interesting relationship between the data in the database. It is mainly used for the Market basket analysis to help improve the business activity. Association rules are obtained using two main criteria the support and the confidence. The support indicates how frequently the items appear in the database. Association mining discovers the relation of items in same transaction [2].

Sequential, Structural Pattern It is to discover set of frequent subsequence in a transactional database. The sequential pattern mining is a very important concept of data mining. In a sequential pattern mining, events are linked with time. It discovers the correlation between the different transactions.

Classification Classification is the process to predict class of objects which do not have any class label. If finds the data classes and concepts. It is based on the analysis if sets of training data. Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome.

ClusteringClustering is method of grouping similar and related objects together. It is used to classify unstructured and semi-structured dataset. It is unsupervised learning technique. Clustering algorithm are broadly classified into hierarchical, partition and density based clustering. K-Means clustering algorithm is one of the techniques to provide structure to unstructured data.K-Means is traditional and widely used partitioned clustering algorithm. It has asymptotic running time with respect to any variable of the problem [6].

II. LITERATURE SURVEY

The traditional data is in structured format so it can be easily stored and queried. However nowadays data is not in typical structured format. The data uploaded on Facebook, Twitter, and YouTube etc. is completely unstructured. It is in format of images, video, audio, documents which are unstructured [3]. Traditional database cannot handle this unstructured data. The data generated in huge amount, with high velocity, and variety. Therefore it termed as Big data. There are different data mining technique's which are used to extract useful information. Linguistic matching techniques are used to give structure by identifying similarities between elements [2]. It can be applied to different data sources to avoid overlap.

Ashwin Kumar et.al [2] "Efficient structuring of data in Big Data" discusses the structuring of big data using linguistic matching. The data is then clustered using Markov's clustering algorithm. It uses techniques of string matching, tokenization matching stemming to find similarity in data items. Establishing the relationships or logical mapping among elements from different data sources is schema matching. Linguistic matching uses element name or description to find matching. Graph matching contains two algorithm such as fixed point computation on similarity and probabilistic constraint satisfaction algorithm. In constraint- based matching the properties of element like uniqueness, data types, value range etc. are considered. Proposes Markova's clustering algorithm to cluster data from different data sources. Data is analyzed for similarity between the clusters.

Frequent Itemset mining is very essential technique to extract useful information from Big data. There are many algorithms present which describes different methods to produce frequent itemset. But existing algorithm cannot efficiently deal with Big data [1]. Hadoop ecosystem is the solution for Big data. It has two basic components MapReduce and HDFS. The existing algorithms have challenges while dealing with Big Data. MapReduce has Map, Combine and Reduce functions which uses (key, value) pair. Distance between sample point and random centers are calculated for all points using map function. Intermediate output values from map function are combined using combiner function. All samples are assigned to closest cluster using reduce function. K means clustering algorithm extract frequent itemset very fast, reduce execution time.

Prajeshachalia, Anjan K Koundinya and Srinath N K [8] "MapReduce Design of K-Means Clustering Algorithm" Describes the K-Means algorithm for distributed environment using Hadoop. Mapper and reducer is designed to implement K-Means algorithm. Proposed method provides a system which group the data together with similar characteristics which reduces implementation cost. Method can be improve to handling outlier.

MadjidKhalilian, Norwati Mustapha, et al. [7] "A Novel K-Means Based Clustering Algorithm for High Dimensional Data Sets". Proposed method of divide and conquer technique which improves the performance of K-Means algorithm for high dimensional dataset. Experiment results demonstrate appropriate accuracy and speed up.

Objective of proposed framework is combining relational definition of clustering space

and divide and conquer method has been describe to overcome difficulties. Based on Proposed method data stream processing can be improved. Data type is another direction to examine proposed method. K-Means has been used for second phase whereas it can be improve with clustering algorithms.

Sisirkumar and anjan amehanta [4] "An alternative technique of selecting the initial cluster centers in the K-Means algorithm for better clustering".The problem with using randomly selected initial centroids,is overcome Furthest-First technique with K-Means. It produces better quality clustering by lowering sum of squared errors.

Mugdha Jain and ChkradharVerma [6] "Adapting K-Means for clustering in Big data". Proposed techniques describe an approximate algorithm based on K-Means. It is a novel method for big data analysis which is very fast, scalable and has high accuracy. It overcomes the drawback of K-Means of uncertain number of iterations by fixing the number of iterations, without losing the precision.

R. C. Saritha and Dr. M. Usha Rani [15] "Map Reduce Text Clustering Using Vector Space Model" describes map reduce approach for clustering the documents using vector space model. proposed is efficient with the increase of text corpus.

III. PROPOSED SYSTEM

Proposed model provides approach for structuring and frequent itemset mining for big data. There are different methods of frequent itemset mining.Association, correlation analysis, Sequential, Structural pattern, Classification, Clustering, Regression analysis etc. Here we are using clustering method for frequent itemset mining. Hadoop Environment is used for structure the data by the help of MapReduce framework. KMP string matching algorithm and K-Means clustering algorithm together can give better recommendation for user query.The proposed model is given below figure 3.1

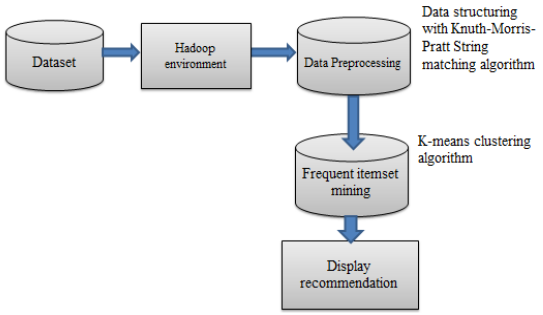

Figure 2: Proposed Architecture

Stages in the system

The proposed system consists of following**.**

A. Data collection
B. Hadoop Environment
C. Data Pre-processing
D. Analyzing context: Linguistic matching(KMP string matching algorithm)
E. Recommendation : Frequent Itemset Mining (K-Means Clustering)

A. Dataset

Data is collected from social media. Dataset will be Movie Dataset of IMDB. IMDB Dataset is in .txt file format. The dataset with 1682 movie data containing total 24 attributes like Movie id, Movie_name, Movie _release year, Movie_release date, Movie_URl, and Other 19 are genre variable represents presence of genre as 1 and absences of genre as 0.Movies in the IMDB dataset are categorized in genres as follows.

1.Unknow	6.Comedy	11.Film Noir	16. Sci-Fi
2.Action	7.Crime	12. Horror	17. Thriller
3.Adventure	8.Documentary	13.Musical	18. War
4. Animation	9. Drama	14.Mystry	19. Western
5. Children	10.Fantacy	15.Romance	

Table 1: List of Genre

Example 1|ToyStory (1995)| 01-Jan-1995||http://us.imdb.com/M/title-exact?Toy%20Story%20(1995)
|0|0|0|1|1|1|0|0|0|0|0|0|0|0|0|0|0|0|0
Movie_id is 1
Movie_name is Toy Story
Movie_release Date is 01 –Jan-1995
Toy Story |0|0|0|1|1|1|0|0|0|0|0|0|0|0|0|0|0|0|0key provides genre information about the movie "Toy Story" is categorized as Animation, Comedy, and Children's.

A. Hadoop Environment

Hadoopis an open-source software framework for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware.The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part called MapReduce. Hadoop splits files into large blocks and distributes them across nodes in a cluster [13].

HDFS is the primary distributed storage used by Hadoop applications. Data is loaded into HDFS System. A HDFS cluster primarily consists of a NameNode that manages the file system metadata and DataNodes that store the actual data.

B. Data Pre-processing

Data is further structured by using Map Reduce framework. Input text file is tokenized in different word tokens. After tokenizing unstructured data, we create two modules which interfaced with Hadoop.First module process the data for structuring by using KMP algorithm. Second module check for frequent itemset in dataset. Output from these two modules is combined [2].

Read input string form text file and tokenize it.

Example

1413 | Street Fighter (1994) | 01-Jan-1994
1408 | Gordy (1995) | 01-Jan-1995
1406 | When Night Is Falling (1995) | 01-Jan-1995

Above lines can be tokenize as follows :

1413, Street Fighter, (1994) ,01-Jan-1994
1408, Gordy, (1995) ,01-Jan-1995
1406, When, Night, Is, Falling,(1995) ,01-Jan-1995

MapReduce Framework

MapReduce framework proposed by Google is a processing and execution model for distributed environment that runs on large clusters [1]. MapReduce with Hadoop distributed file system is used to find out frequent itemset from Big Data on large clusters.MapReduce framework has two phases [1] [8].

- Map phase: Map phase takes data from HDFS and generate (Key, Value) pair. Map function collects all intermediate key - value pair and passes it to reduce function.
- Reduce phase: Reduce function takes these intermediate values through iterator and merge it to produce (Key', Value'). By Map Reduce framework too large values can also fit in to the memory.

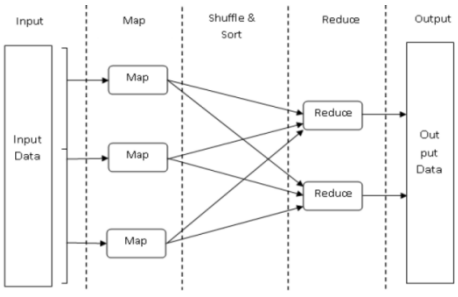

Figure 3: MapReduce Framework [1]

Example

Map 1

1408, 1
Gordy, 1
(1995) , 1
01-Jan-1995 1

Map 2

1406, 1
When, 1
Night, 1
Is, 1
Falling , 1
(1995) , 1
01-Jan-1995 1

Reduce :: (key', list (value')) → (key'', value'')

Reducer

1408, 1
Gordy, 1
(1995) 2
01-Jan-1995 2
1406, 1
When, 1
Night, 1
Is, 1
Falling , 1

*D. Analyzing context: Linguistic matching (KMP string matching algorithm***)**

To give recommendation for user on the basis of user query string matching algorithm is used. We use Knuth–Morris–Pratt algorithms to find out the similarity of data items contextually.

The Knuth–Morris–Pratt algorithm searches for matching pattern in dataset. It takes the pattern entered by user, checks for sub-pattern matching in pattern first and then it matches with the dataset. By this KMP algorithm avoids unnecessary search and avoid backtracking.

The worst case complexity of Knuth–Morris–Pratt algorithm is O (n).

For proposed model the Knuth–Morris–Pratt string matching algorithm produces recommendations by matching movie names. If user searches for movie "Star War" then recommendation should be movies such as "Star War II".

Knuth–Morris–Pratt string matching algorithm

- Detects String Similarities between elements from different data sources.
- To get better recommendation for user.
- To find out the similarity of data items

Components of KMP

The prefix function, Π

Prefix function matches pattern against shift of itself. The function avoids useless shifts of pattern 'p'and avoids backtracking on string 'S'.

- Compute-Prefix-Function (Π)

```
1  m ← length[p]   //'p' pattern to be matched
2  Π[1] ← 0
3   k ← 0
4   for q ← 2 to m
5   do while k > 0 and p[k+1] != p[q]
6   do k ←Π[k]
7   If p[k+1] = p[q]
8   then k ← k +1
9   Π[q] ← k
10  ReturnΠ
```

The KMP Matcher

The occurrence of pattern with prefix function in string is calculated.

With string 'S', pattern 'p' and prefix function 'Π' as inputs, finds the occurrence of 'p' in 'S'.

KMP-Matcher(S,p)

```
1 n ← length[S]
2 m ← length[p]
3 Π← Compute-Prefix-Function(p)
4 q ← 0                                  //number
    of characters matched
5 for i ← 1 to n                         //scan S
    from left to right
6    do while  q> 0 and p[q+1] != S[i]
7    do  q ←Π[q]                         //next character
    does not match
8    if p[q+1] = S[i]
9    then q ← q + 1                      //next character
    matches
10  if q = m                             //is all of p
    matched?
11  then print "Pattern occurs with shift" i – m
12   q ←Π[ q]                            // look for the next
    match
```

Example

Input Pattern: velvet

- Computing Prefix

	1	**2**	**3**	**4**	**5**	**6**
P	V	E	L	V	E	T
Π	0	0	0	1	2	0

K	0	0	0	1	2	0
Q		2	3	4	5	6

Output: Blue **Velvet** (1986
Velvet Goldmine (1998)
Terminal **Vel**ocity (1994)

E. *Recommendation: Frequent Itemset Mining (K-Means Clustering)*

Cluster is collection of data having similar characteristics. For unstructured data, there is need to analyzed to derive meaningful information. K-Means clustering is technique which provides structure to unstructured data [8]. Only structuring the data is not efficient for big data, as volume of data is too big. Therefor meaningful information extraction is very important [2]. Various data mining techniques are available for big data [1]. Clustering is technique we are using here to mining frequent itemset.

The frequent itemset mining uses K-Means algorithm which works to find similarity between data items which are similar in some characteristics [2]. K-Means is one of the most common and simple, famous partition clustering algorithm [8][6].

K-MEANSCLUSTERING

K-Means is a common and well-known clustering algorithm. K-Means works on MapReduce routine. The input is given as <key,value> pair where Key is the centroid of cluster and value is data object. BasicallyK-Means clustering works as follow.

K-Means Clustering Algorithm

- Take k random values as cluster centers.
- Let each item belong to the cluster whose cluster center it is closest to.
- For each of the k clusters, find a new cluster center by taking the mean of its items.
- Repeat steps 2 and 3 until the cluster centers are stable.

In Hadoop, K-Means runs in mapper and reducer. Centroid file with data items are the input to mapper. The distance between centroid and data object is evaluated using Euclidian distance measure.

Distance(i,j)

$$=\sqrt{(x_{i1}-x_{ji})^2+(x_{i2}-x_{j2})^2+(x_{i3}-x_{j3})^2} \quad ....(1)$$

The data object with minimum distance will go in the cluster of the centroid. Each remaining objects are assigned to centroid by distance similarity score. New mean is calculated for each cluster. Iterate through the previous steps to find new cluster till no change occur in cluster centroid.

For 'n' objects to be assigned into 'k' clusters, the algorithm will have to perform a total of 'nk' distance computations. While the distance calculation between any object and a cluster center can be performed in parallel, each iteration will have to be performed serially as the center changes will have to be computed each time.

Algorithm for K-Means Mapper

Input: A set of objects X = {x1, x2,…..,xn},

A Set of initial Centroids C = {c1, c2, ,ck}
Output: A output list which contains pairs of (Ci, xj)
Where 1 ≤ i ≤ k and 1 ≤j ≤ n
Steps:

1. M1←{x1, x2,…….,xm}
2. current_centroids←C
3. distance (p, q) =$\sum_{t_i \in C_s} \|pi - q_i\|^2$
   (where pi, qi are is the coordinate of p and q in dimension i)
   for all xi ∈ M1 such that 1≤ i ≤m do
4. bestCentroid←null
   minDist←∞
5. for all c ∈ current_centroids do
   dist← distance (xi, c)
   if (bestCentroid = null || dist<minDist)
   then
   minDist←dist
   bestCentroid ← c
   end if
   end for
6. emit (bestCentroid, xi)
   i+=1
   end for
7. return Outputlist

Algorithm for Reducer
Input: (Key, Value), where key = bestCentroid and Value= Objects assigned to the centroid by themapper
Output: (Key, Value), where key = oldCentroid and value= newBestCentroid which is the new centroid value calculated for that bestCentroid
Procedure

1. outputlist←outputlist from mappers
   $\vartheta \leftarrow \{\}$
2. newCentroidList ← null
3. for all β outputlist do
   centroid ←β.key
   object ←β.value
   [centroid] ← object
   end for
4. for all centroid ∈ϑ do
   newCentroid, sumofObjects,
   sumofObjects← null
   for all object ∈ϑ [centroid] do
   sumofObjects += object
   numofObjects += 1
   end for
5. newCentroid ← (sumofObjects +numofObjects)
   emit (centroid, newCentroid)
   end for
   end

For K-Means clustering algorithm, Input given as Key, Value pair and number of cluster want to be form as a four then four cluster are created at first step and randomly it select four Movies from input an place each in separate cluster and remaining are assign to cluster on the basis of similarity and calculate centroid of each cluster and redistribute dataset on the basis of similarity to centroid this process perform recursively till no change is occur.

Example
K-Means Mapper
Input: A set of objects X = {x1, x2,…..,xn},
A Set of initial Centroids C = {c1, c2, ,ck}

Let Centroid is Toy Story: (0,0,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0)
Distance between Points centroid and data item is evaluated by Euclidian distance as follow.
Toy Story: (0,0,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0)
Batman Forever: (0,1,1,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0)
As given in equation (1) Distance is calculated.

Distance=$\sqrt{(0-1)^2 + (0-1)^2 + (1-0)^2 + (1-0)^2 + (0-1)^2 \ldots}$
= 2.2360
Home Alone (0, 0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0)

Distance =$\sqrt{(0-0)^2 \ldots + (1-0)^2 + (1-1)^2 + (1-1)^2 \ldots .}$
= 1
True Romance (0,1,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0)

Distance=$\sqrt{(0-1)^2 + (1-0)^2 + (1-0)^2 + (1-0)^2 + (0-1)^2 + (0-1)^2 \ldots}$

= 2.44
As result shows Distance of Centroid with the movies are as follows

Movie	Distance
Batman Forever	2.2360
Home Alone	1
True Romance	2.44

Table 2: Distance Measure

Distance of Movie Home Alone is minimum from centroid therefore Home Alone will be placed in cluster with centroid Toy Story.
bestCentroid for Home Alone is Toy Story.
minDist =1
For all centroid and all Objects it is iterated and depending upon minimum distance the movie will get placed in appropriate cluster.

Output will be (bestCentroid ,object)
K-Means Reducer
Input: (Key, Value), where key = bestCentroid and Value= Objects assigned to the centroid by themapper
Output: (Key, Value), where key = oldCentroid and value= newBestCentroid which is the new centroid value calculated for that bestCentroid
Recommendation:
If user searches for movie with genre like animation , comedy as recommendation output should be matching genre top five movies.

IV. CONCLUSION AND FUTURE WORK

Conclusion

The proposed model handles data context analysis and frequent itemset mining to generate an efficient structure for unstructured data in Big data. Proposed method can minimize the difficulties faced by the user of Big data. User would require to invest a lot of time in studying the unstructured data. By using Hadoop Map Reduce Framework, proposed method tried to minimize the time required by user of Big data to investigate the unstructured data. Proposed approach is able to structure data which can make the data more usable by the user and the recommendation system gives appropriate recommendation to user using K-Means algorithm. The proposed model can find matching movies set using KMP string matching algorithm. K-Means clustering algorithm on Map Reduce framework is used for clustering the data, which will help user to get better recommendation.

Future Work

As future work our approach can be expand to meet other demands of big data such as velocity. Various other attributes of movies can be used to produce more accurate recommendation. Frequent item set mining algorithm and map reduce framework on stream of data which can be real time insight in big data.

REFERENCES

[1]. SheelaGole and Bharat Tidke.”Frequent itemset mining for Big Data in social media using ClustBigFIM algorithm”. IEEE International conference on Pervasive Computing (ICPC) 2015.
[2]. Ashwin Kumar T K, Hong Liu, and Johnson P Thomas. ”Efficient structuring of data in Big Data”. IEEE International Conference on Data Science and Engineering (ICDSE) 2014.
[3]. Han hu, Yonggang wen, (senior member, IEEE), Tat-sengchua, And Xuelong li, (Fellow, IEEE) “Toward Scalable Systems for Big Data Analytics: A Technology Tutorial”. VOLUME 2, 2014
[4]. Sisir Kumar Rajbongshi and AnjanaKakotiMahanta.” An Alternative Technique of Selecting the Initial Cluster Centers in the K-Means Algorithm for Better Clustering”.International Journal of Computer Applications April 2013.
[5]. Xingjian Li” An Algorithm for Mining Frequent Itemsets from Library Big Data”.Journal Of Software, Vol. 9, No. 9, September 2014.
[6]. Mugdha Jain ChakradharVerma“ AdaptingK-Means for Clustering in Big Data” International Journal of Computer Applications (0975 – 8887) Volume 101– No.1, September 2014.
[7]. DhamdhereJyoti L., Prof. DeshpandeKiran B. “A Novel Methodology of Frequent Itemset Mining on Hadoop”.International Journal of Emerging Technology and Advanced Engineering ,JULY 2014.
[8]. Prajesh P Anchalia, Anjan K Koundinya, Srinath N K “MapReduce Design of K-Means Clustering Algorithm”IEEE 2013.
[9]. Ronghu,wanchundou,jianxunliu.”ClubCF: A clustering –based Collaborative Filtering Approach for Big Data Application” IEEE SEPTEMBER 2014.
[10]. NawsherKhan,IbrarYaqoob,IbrahimAbakerTargioHashem,ZakiraInayat,WaleedKamaleldinMahmoudAli,MuhammadAlam,MuhammadShiraz,and Abdullah Gani “Big Data: Survey, Technologies, Opportunities, and Challenges”.
[11]. Victoria L´opez, Sara delR´ıo, Jos´e Manuel Ben´ıtez and Francisco Herrera “On the use of MapReduce to build Linguistic Fuzzy Rule Based Classification Systems for Big Data”.2014 IEEE International Conference on Fuzzy Systems.
[12]. Ms. Vibhavari Chavan, Prof. Rajesh. N. Phursule “Survey Paper on Big Data” International Journal of Computer Science and Information Technologies, Vol. 5 (6) 2014.
[13]. Hadoop tutorial viewed [Online] Available at: http://www.tutorialspoint.com/hadoop/hadoop_big_data_overview.htm(Last accessed on 20 August, 2016 at 9 PM)
[14]. Hadoop command guide [Online] Available on:http://hadoop.apache.org/docs/current/hadoopprojectdist/hadoop-common/CommandsManual.html(Last accessed on 21 August, 2016 at 11AM)
[15]. R.C. Saritha and Dr. M.Usha Rani “Map Reduce Text Clustering Using Vector Space Model”September 2014 International journal

of emerging technology and advanced engineering, Volume 4,Issue 9.

[16]. Kyung-Rog Kim, Ju-Ho Lee, Jae-Hee Byeon "Recommender System Using the Movie Genre similarity in Mobile service" IEEE 2010.