ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

Video Retrieval Using Spectral Clustering and B-Tree Indexing

R.G.Sakthivelan, Dr.P. Rajendran

Assistant Professor, Department of CSE, Mahendra Engineering College, Namakkal, India

Professor, Department of CSE, Knowledge Institute of Technology, Kakapalayam, India

ABSTRACT: Initially linear video dataset is formed for efficient retrieval of video data from a huge video database. Before undergoing the process of knowledge discovery feature reduction process is implemented. This reduces the dimensionality and increases the space of video video data storage. Hence the map reduce is processed for the next step in knowledge discovering process to remove unwanted and irrelevant video data from the video database. The Support Vector Machine is one of the classifications technique is used. This overcome the problem of k Means disadvantage, it does not support effectively for both linear and nonlinear format of video data. Map Reduce method to add privacy to a huge video database can be obtained by adding dual authentication technique which ensures the privacy of the user without over heading the process. This overcomes the overlapping issue caused by the k means algorithm and it also reduces the issue of finding the distance between the record and cluster.

KEYWORDS: Mapreduce, Group Decision, Support Vector Machine, B-Tree

I.INTRODUCTION

In internet, amount of video data increases day by day which leads to difficult analysis by video data mining techniques. The sources of video data can be a video database, video data warehouse, the web, other information repositories or video data which are retrieved and stored in the system dynamically [1]. It causes inefficient and scalability problem. But when video dataset are humongous in size, a wide distribution of video data is needed and complexity arises which leads to the development of parallel and distributed video data-intensive mining algorithms [2]. Big Video data Analytics is the process of computing such large video dataset in parallel using MapReduce environment [3].

A. *Opinion Mining*

Opinion Mining also refers to sentiment analysis is the process of analyzing the text in the document and provides the suggestions to the people by extracting opinion through online [4]. Users post their opinion about the services or products in the blogs, shopping sites, or review site Reviews about hotel, automobiles, movies, restaurants are available on the websites [18], [19], [20], [21]respectively.Text analysis in opinion mining is the process of getting high quality information from the text. Approximately, 90% of the world’s video data is available in unstructured format.By parsing this unstructured video data, the patterns involved in it are identified and recommendations are provided.

B. *Recommendation and Collaborative Filtering*

Traditional system provides recommendation to particular application based upon the ranking given by the personalized user[5]. Now-a-days many application uses recommendation system which includes CDs, books, webpage, hotel reservation system and various[6], [7], [8]. In hotel reservation system, if one user is concerned about particular services and another user is looking for different services in the same hotel. But the same recommendation service is provided for both the user. It is not the good recommendation and the people will not satisfy to the recommendation. Moreover, in hotel reservation system the ratings of services and service recommendation list to the users are same does not consider user preferences [9].

Recommendation system is classified as content based, collaborative based and hybrid based recommendation system. Content based recommendation provides recommendation by taking the user preference from the previous user

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

reviews. Collaborative Filtering (CF) recommends service based on the reviews of the previous user, by checking the similarity with the current user. Hybrid recommendation system combines recommendation of both content and CF.

C. *Big Video data Framework*

Cloud computing is an effective platform to facilitate parallel computing in a collaborative way to tackle large-scale video data. Big Video data Analytics provides solution to these problems. Big video data explains term of video data sets which is large or complex so that traditional techniques failed to perform task [22].

The main characteristics of Big Video data are volume, variety, veracity and velocity. In Big Video data, the large video dataset are partitioned into small video data. Each video data is further processed in parallel, by searching the patterns. The parallel process may interact with one another. The patterns from each partition are eventually merged and produce the result. Cloud computing tools are Hadoop, Mahout, MapReduce [10]. Hadoop is the open source tool for MapReduce and Google File System [11] which supports MapReduce programming framework written in Java, originally developed by Yahoo. Nowadays everything acts as a service, so creating and recommending the service using big video data analytics in the social networking will be more efficient and accurate. The File System used for storing large video data is Hadoop Distributed File System (HDFS) and simply adding the number of servers can achieve growth in storage capacity and computing power [12].

II. RELATED WORK

Recommendation is based on the people having similar preferences and interest (i.e. stable one) from past reviews [7]. It provides similarity computation using k-nearest neighbors. It uses user history profile as rows, their reviews as column and forms rating matrix. Cosine similarity used for calculating weight of rank matrix, which gives number of interaction between rows and columns. Finally, calculate the item rating from weighted average rating of the neighbor user. It implements in MapReduce framework for overcoming scalability. It takes large computational time when dealing with huge amount of input video data. So improvement must be done on Hadoop platform to reduce the computation time when dealing with these algorithms.

The system with most predicted rating by same user for similar items [6]. User-item matrix is formed by finding relationship between different items and to provide recommendation to the user. Consider reviews of similar item and identify similarity computation for item-item based approach. It computes using cosine based similarity, correlation based similarity and adjusted cosine based similarity. Finally, predicted rating for the target user is provided. Some other method is used in order to overcome the scalability issue.

Keyword based service recommendation system [13] which takes the preferences from the previous user keyword set and finds the similarity with the active user keyword set. Using CF, personalized rating for each service is considered and lists the top recommended services. Drawbacks of this system, it does not consider the positive and negative preferences. In order to make more accurate the bigrams of words is taken.

III. PROPOSED SYSTEM

The proposed system uses previous user reviews to find similarity with the active user and provide recommendation of service based on the active user needs [14]. First step is to form candidate service list for the application along with domain thesaurus i.e. semantic words [15]. Then collect the previous user post in the form of reviews, which includes their opinion about the application. After the collection of reviews, a review sentence is given to video data preprocessing. Video data preprocessing consist of stop word removal and Part-Of-Speech (POS) tagging. The keywords obtained are taken as keyword set of previous user. Meanwhile active user needs to provide the service as keywords. The system extracts opinion words in reviews and classified as n-level orientation scale [16]. Opinion orientation is an intended interpretation of the user satisfaction in terms of numerical values. It used to identify the number of positive and negative opinions of each keyword by using supervised learning algorithm [17]. Next, the similarity between the active and previous user's preference keyword set is calculated. The similarity computation is done by jaccord and cosine similarity method [13]. Finally, personalized rating for each service of the active user is calculated as shown in Fig. 1 and recommend top-k rating is provided to the active user [5].

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

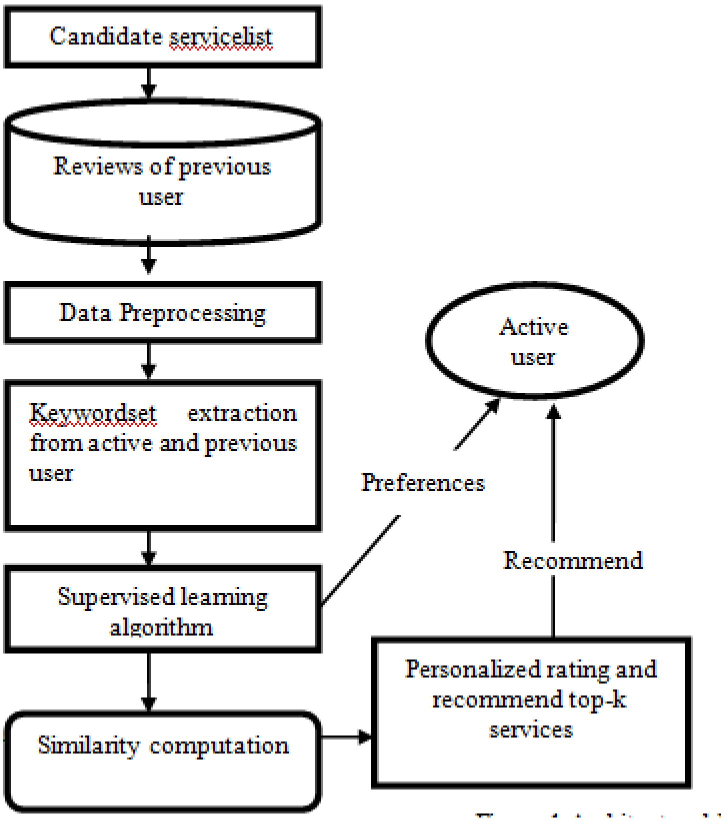

Figure 1.Architectural Diagram

The main steps of semantic based service recommendation system are described as follows**:**

A. Video data Preprocessing

Stop word removal involves removing of unwanted and low priority words in each review sentence. Reviews are stored in HDFS which is given as input to stop word removal. Then each word is tagged using POS tagger.

B. Keyword Extraction

Active user gives their preferences of service as keywords by selecting from the candidate service list. From the active user preference services, keyword set is formed as Active Preference Keyword (APK). Then correspondingly previous reviews will be transformed as Previous Preference Keyword (PPK) set along with semantic words. Keywords tagged as noun by tagger is extracted [16] and check for the most frequent keyword using Apriori algorithm with minimum support count. The algorithm for keyword extraction is shown as follows:

```
keyword extraction (pos tagged input reviews)
        if word is in noun then
             extract (word)
        endif
count numbers of each word
set a minimum support count
        if count is  greater than minimum support count
             display (word)
       else
             remove (word)
       endif
```

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

Bayes theorem calculates probability using supervised term counting based approach. It is used to identify keyword orientation by determining whether a given review is a positive, negative or neutral using opinion word [17]. In this algorithm, the probabilities of the labels are found according to the words. Steps are as follows:

- The positive and negative opinion words and review sentences are stored in text file.
- Split the sentence into the combination of words. It means first combination of two words and then single words.
- First compare the combination of two words, if matched then delete that combination from the opinion. Again start comparing for the single words.
- Initially, the probabilities of all the labels are zero [positive=0, negative=0]. Based on opinion, the probabilities of positive and negative labels get incremented

Similarly, the negation rule algorithm is applied for opinion orientation is as follows:

```
if opinion _word is near a negation word then
    orientation←Apply Negation Rules(orientation)
end if
    return orientation
```

Negation rules have a negation word or phrase which usually reverses the opinion expressed in a sentence. Three rules must be applied:
Negation Negative->Positive e.g., “no problem”
Negation Positive ->Negative e.g., “not clean” and
Negation Neutral-> Negative e.g.,” does not suite”, where “suite” is a neutral verb.

D. Jaccard Similarity Method

Jaccord similarity is an approximation method used for finding similarity between APK and PPK. It is does not consider the repetition of keywords in the keyword set. It takes the extracted keyword set of different previous users and compares the similarity with the preference keyword set of active user. The jaccord similarity method is given in algorithm as follows:

To calculate the similarity between APK and PPK,

$$sim\,(APK,PPK) = \frac{|APK \bigcap PPK|}{|APK \bigcup PPK|}$$

return sim (APK,PPK)

E. Cosine Similarity Method

It is an exact similaritymethod to find the most similarity between active preference keyword set and previous preference keyword set. The number of times the particular keywords is repeated in the APK and PPK is taken as weight of the keyword. If the keyword is not available in the preference keyword set, then the weight of the keyword will be taken as zero (i.e. $w_{ij} = 0$). Cosine similarity is also known as vector space model, in which the weight of the keywords in keyword set will be transformed as vector. Then the Term Frequency and Inverse Document Frequency (TF-IDF) is used for finding the number of times the particular term occurs in the document i.e. the frequency of the keywords. The frequency of the keyword will be taken as weight of the keyword in the keyword set. TF-IDF is calculated for both active preference keyword set and previous preference keyword set [5], [13].

TF-IDF in which Term Frequency(TF) takes the distinct keywords and number of times the particular keywords appears in the reviews and in the active keyword set in the following function:

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

$$TF = \frac{N_{pk_i}}{\sum_g N_{pk_i}} \qquad (2)$$

where, N_{pki} number of times particular keyword appears in the keyword set, g is the number of keywords in the preference keyword set. The Inverse Document Frequency (IDF) is computed by number of documents containing the keywords divided by the number of keywords present in that document. It is given by following function:

$$IDF = 1 + \log_e\left(\frac{N}{n_i}\right) (3)$$

where, N is the total number of reviews posted by the user, n_i is the number of occurrence of the keywords in all reviews. TF-IDF scores for each keywords is calculated as weight by the function:

$$w_{pk_i} = TF * IDF \; (4)$$

The weight of APK and PPK is used to calculate the cosine similarity is defined as follows,

$$\text{sim(APK,PPK)} = \cos\left(\vec{W}_{AP}, \vec{W}_{PP}\right)$$

$$= \frac{\vec{W}_{AP} * \vec{W}_{PP}}{\|\vec{W}_{AP}\|_2 * \|\vec{W}_{PP}\|_2} \quad (5)$$

where, $\vec{W}_{AP}$ and $\vec{W}_{PP}$ be the weight of the keyword in the keyword set of the active user and previous user

F. Personalized Rating

Using CF algorithm [5], rating of each service is provided based on the cosine similarity value. The previous keyword set which is most similar to the active keyword set is filtered out from cosine similarity. Rating of each keyword using cosine similarity is calculated and provides the top-k rated service to the active user. The personalized rating for each service of the active user is calculated as follows:

$$pr = \bar{r} + \sum_{PPK_j \in R} sim(APK, PPK_j) * (r_j - \bar{r}) \; (6)$$

$$k = \frac{1}{\sum_{PPK_J \in R} sim(APK, PPK_J)} \quad (7)$$

where, $\bar{r}$ be the average rating of service, r_j be the corresponding rating of the different previous user, sim (APK,PPK_j) be the similarity of APK and PPK of cosine similarity value. K is the normalizing factor and R is used to store the previous user after each filtration.

IV. IMPLEMENTATION ON MAPREDUCE

MapReduce [5], [7], [13] used to execute video data in parallel manner. MapReduce used for implementing keyword and opinion extraction, similarity method, raking of services in parallel. It reduces time in running the algorithm.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

V. EXPERIMENTAL EVALUATION

The video dataset used in the experiment is real video dataset [18] which consist of 400mb of different hotels with overall rating of each hotel. The accuracy is measured by the parameters of precision, recall and F-measure as shown below,

$$Precision = \frac{|ExtractedValues \cap TrueValues|}{|ExtractedValues|} \quad (8)$$

$$Recall = \frac{|ExtractedValues \cap TrueValues|}{|TrueValues|} \quad (9)$$

$$F-measure = \frac{2 * Recall * Precision}{(Recall + Precision)} \quad (10)$$

Keyword extraction by apriori algorithm shows the accuracy in Fig.2,

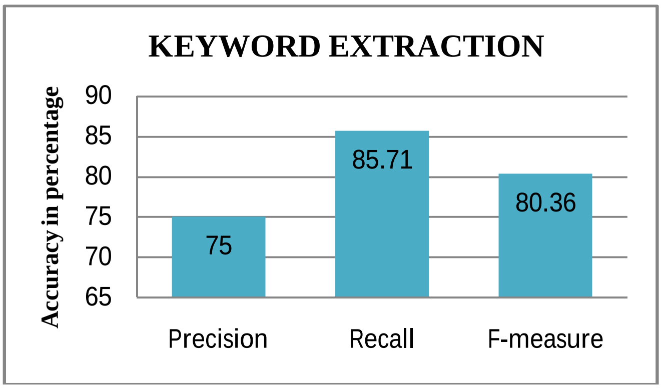

Figure2. Keyword Extraction

By naïve baysen, opinion orientation is analyzed and accuracy is measured for the precision, recall, F-measure is shown in Fig.3,

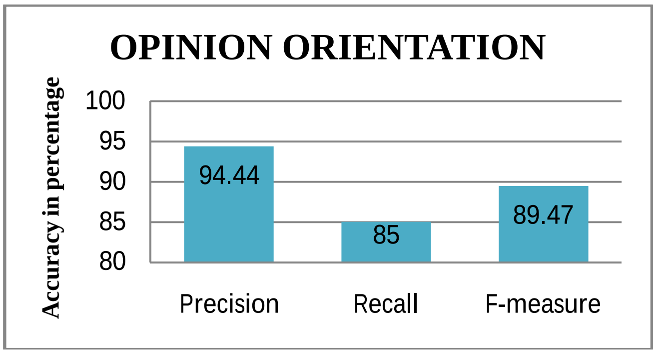

Figure 3. Opinion Orientation

Keyword extraction gives accuracy of 80.36% using frequent itemset mining. Opinion orientation provides 89.47% of accuracy using naïve bayes for hotel video dataset.Fig. 4, shows the outcome of number of keywords based on

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

threshold count in terms of number of keywords threshold. Keywords are extracted for the threshold of 1, 2, 3 and 4, in which some of the keywords are not related to hotel keywords. If the threshold count is greater than 12, there is a chance to ignore some of the keywords. So, the threshold count is set from 5 to 12.

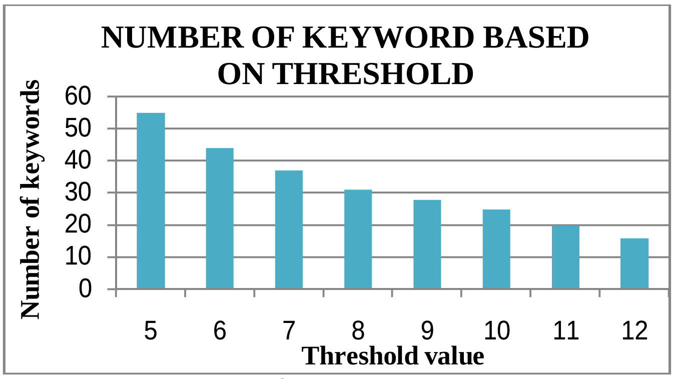

Figure4. Number of keywords based on threshold

In Fig. 5, provides the results of keyword orientation in terms of number of opinions for 5keywordin reviews is shown below,

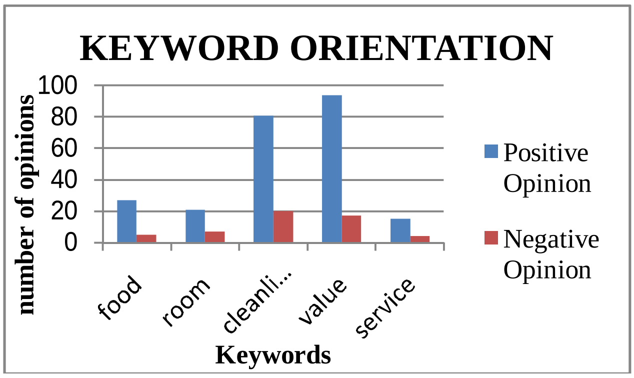

Figure5. Number of Positive and Negative opinions of each keywords

The result is taken for similarity computation of APK with PPK keyword set using jaccord and cosine similarity method. The APK consist of 3 keywords (cleanliness, food, value). From the computation, cosine similarity provides the highest value for keyword between APK and PPK is shown in Fig.6.

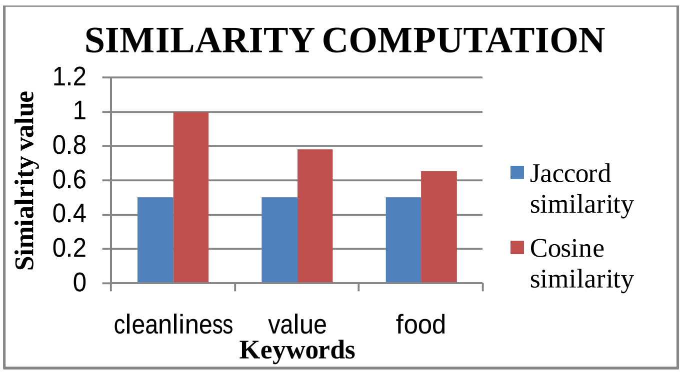

Figure 6. Similarity Computation of jaccord and cosine similarity

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

Rating of keyword for the most similar is rated between (0-5), where the highest value gives most needed keyword to the user. Semantic based service recommendation provide mostaccurate rating than Keyword Aware Service Recommendation (KASR) [13] as shown in Fig. 7,

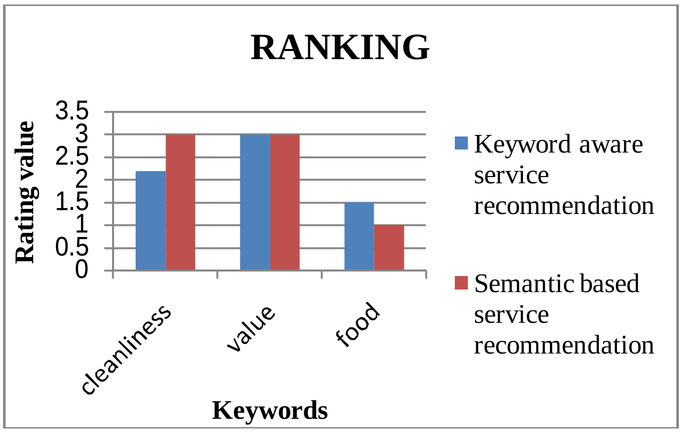

Figure 7. Ranking of keywords

Execution time for a single mapper is higher for both similarity methods. By increasing the number of mapper, execution time is decreased as shown in Fig. 8,

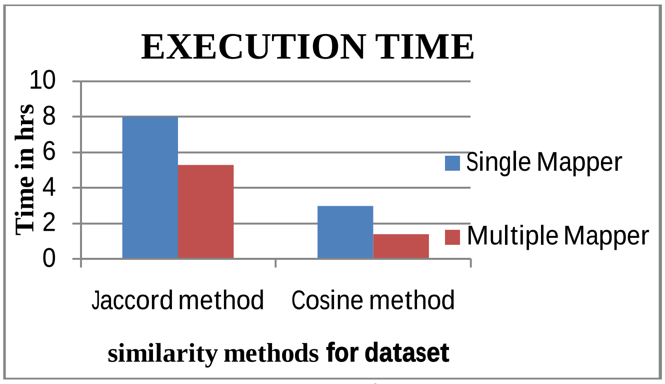

Figure 8. Execution time of similarity methods

VI. CONCLUSION

The proposed system extracts keyword from customer reviews with minimum support threshold. The opinion words are extracted in reviews. Bayes theorem based on probabilities using supervised term counting based approach is used to identify sentence and keyword orientation. The number of positive and negative opinions in review sentences is estimated. And count the number of positive and negative opinion for each keyword in online customer reviews. To validate the performance of the system the rating of each keyword is calculated. The proposed system gives keyword rating but tripadvisor website gives overall rating of the hotel without analyzing the opinions on each keyword. This would make the proposed technique more complete and effective. In future, further implementation is done by increasing the number of node to make the system more efficient and reduces the time in execution.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

REFERENCES

[1] J.Manyika et al. "Big video data: The next frontier for innovation, competition, and productivity,"*McKinsey& Company Publications*, 2011.
[2] C. Lynch, "Big Video data: How Do Your Video data Grow?," *CNI Publication*, vol. 455, no. 7209, pp. 28-29, 2008.
[3] Watkins, Andrew B, "Exploiting immunological metaphors in the development of serial, parallel, and distributed learning algorithms", *Diss. University of Kent at Canterbury*, 2005.
[4] Liu, Bing ,"Opinion mining and sentiment analysis," *Proc. Springer Berlin Heidelberg*, vol.2, pp. 459-526, Jan. 2011.
[5]Zhao, Zhi-Dan, and Ming-Sheng Shang, "User-based collaborative-filtering recommendation algorithms on hadoop," *Proc. IEEE 3rd International Conference onKnowledge Discovery and Video data Mining*, vol. , pp. 478-481,Jan. 2010.
[6]G. Linden, B. Smith, and J. York, "Amazon.com Recommendations: Item-to-Item Collaborative Filtering", *IEEE Trans.Internet Computing*, vol. 7, no. 1, pp. 76-80, Jan. 2003.
[7] M. Bjelica, "Towards TV Recommender System Experiments with User Modeling," *IEEE Trans. Consumer Electronics*, vol. 56, no. 3,pp. 1763-1769, Aug. 2010.
[8] M. Alduan, F. Alvarez, J. Menendez, and O. Baez, "Recommender System for Sport Videos Based on User Audiovisual Consumption," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1546-1557, Dec. 2012.
[9]Sikka R, Dhankhar A, Rana C.,"A survey paper on e-learning recommender system," *International Journal of Computer Applications*, vol. 47,no. 9, pp. 27-30, Jun. 2012 .
[11] Ghemawat, Sanjay, Howard Gobioff, Shun-Tak Leung, "The Google file system," *In ACM SIGOPS Operating Systems Review*, vol. 37, No. 5, pp. 29-43,2003.
[12] Turney and Peter D.,"semantic orientation applied to unsupervised classification of reviews," *In Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 417-424,2002.
[13]Meng, S., Dou, W., Zhang, X., & Chen, J., " KASR: A Keyword-Aware Service Recommendation Method on MapReduce for Big Video data Applications" , *IEEE Trans. Parallel and Distributed Systems, vol.25, no.*12, pp. 3221-3231, Dec. 2014.
[14]Singam, J. Amaithi, and S. Srinivasan, "optimal keyword search for recommender system in big video data application,"*ARPN Journal of Engineering and Applied Sciences*, vol. 10, no. 7, April 2006.
[15] Turney and Peter D , "semantic orientation applied to unsupervised classification of reviews", " *Proc. of the 40th annual meeting on association for computational linguistics*, pp. 417-424, 2002.
[16] Zhang L., Liu B., Lim S. H., & O'Brien-Strain E., " Extracting and ranking product features in opinion documents," *Proc. of the 23rd International Conference on Computational Linguistics: Posters () Association for Computational Linguistics*, pp. 1462-1470, Aug. 2010.
[17] Hu, Minqing, and Bing Liu, "Mining opinion features in customer reviews," *AAAI* ,vol. 4. no. 4, 2004.
[18]http://www.tripadvisor.com
[19]http://www.caranddriver.com
[20]http://www.imdb.com
[21]http://www.yelp.com
[22]http://www.biomedcentral.com
[23]Lam, Chuck, "Hadoop in action,"*Manning Publications Co.*, 2010.