ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume: 02 Issue: 04, April 2017, pp. 36 – 43

Semantic Ranking Based Service Recommendation System using MapReduce on Big Datasets

Dr. Kogilavani Shanmugavadivel
Assistant Professor (SRG), Department of CSE, Kongu Engineering College, India
Email: kogilavani.sv@gmail.com

Dr. Thangarajan Ramasamy
Professor, Department of CSE, Kongu Engineering College, India
Email: rthangs@kongu.ac.in

Dr. Malliga Subramanian
Professor, Department of CSE, Kongu Engineering College, India
Email: mallisenthil@kongu.ac.in

Dr. Kanimozhiselvi Chenniangirivalasu Sadhasivam
Associate Professor, Department of CSE, Kongu Engineering College, India
Email: kanimozhi@kongu.ac.in

Abstract: *Recommendation system provides most relevant service to the user. Currently, information through online increases a lot which leads to the congestion of data in online and there is a possibility of getting less prediction on service. In previous work, recommendation of services to the user is not based upon the service needed by the user at a time. The proposed system deals with the implementation of personalized rating to the services for hotel reservation system and booking of cars. Candidate service is taken as keywords from the Domain Thesaurus which consists of semantically annotated words. Active user provides their needed service as a preference for each application. Keywords with positive opinion are considered and similar user's opinions are taken from the reviews using keyword extraction method. After keywords extracted, similarity is computed between active user preferences with reviews of the previous user using jaccard and cosine similarity measures. Personalized rating to the most similar keywords is provided to the user as a recommended service using MapReduce framework with single node and multi node environment. The result shows that the execution time of the system is decreased in multi node setup compared to single node environment.*

Keyword: *MapReduce; Preferences; Semantic Ranking; Service Recommendation System;*

1. INTRODUCTION

In Internet, storage of data increases day by day which leads to difficult in analyzing using data mining techniques. The sources of data can be a data warehouse, database, the web, other information repositories or data which are retrieved and maintained in the system dynamically [1]. This leads to inefficient in retrieving large amount of data and scalability issues. When datasets are large in size, a wide distribution of data is needed and complexity arises which leads to the development of parallel and distributed data-intensive mining algorithms [2]. Big Data Analytics used for computing such large dataset in parallel using MapReduce environment [3].

Opinion Mining also referred as sentiment analysis involved in analyzing the text in the document and provides the recommendation 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.

Traditional system provides recommendation for each 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. In hotel reservation system, if one user is concerned about particular service and another user is looking for various service in the same hotel. Then ranking of service provided for the recommendation of both the users will be same. It is not a good recommended system and people will not satisfy with the recom-

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume: 02 Issue: 04, April 2017, pp. 36 – 43

mendation. Moreover, in hotel reservation system the raking of service and service recommendation list to the users are the same and do not consider the user preferences in recommending the service.

Recommendation system can be classified as content based, collaborative based and hybrid recommendation system. Content based recommendation system provides recommendation based on the user preference from the previous user reviews. Collaborative Filtering (CF) techniques recommend service based on the reviews of the previous user, by checking the similarity with the current user. CF is further classified as item-based CF and user-based CF. In item-based CF rating is provided based upon the related item rating by the same user and in user-based CF rating is predicted based upon the same item rating provided by the similar user. Hybrid recommendation system combines recommendation of both content and CF based recommendation.

Cloud computing is an effective platform to facilitate parallel computing in a collaborative way to tackle large-scale data. Big Data Analytics deals with the problems in large dataset. [6]. The main characteristics of Big Data are volume, variety, veracity and velocity. In Big Data, the large dataset are partitioned into small dataset. Each dataset 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. Most widely using Big Data Analytics tools is Hadoop. It is the open source tool for MapReduce framework written in Java, originally developed by Yahoo. Nowadays everything acts as a service, so creating and recommending the service using big data analytics in the social networking will be more efficient and accurate. The File System used for storing large datasets are Hadoop Distributed File System (HDFS). In this by simply adding the servers can be achieved growth in storage capacity and computing power [7].

This paper is organized as follows. Section 2 discusses about related work on recommendation system. Section 3 presents an overview of the proposed recommendation approach. The experimental and evaluation measures are discussed in section 4. Finally section 5 concludes this paper.

2. RELATED WORK

Recommendation provided to the people having similar interests and preferences (i.e. stable ones) from previous reviews [5]. It performs similarity computation using k-nearest neighbors. It uses active user history profile as rows, their comments as column and forms a rating matrix. Cosine similarity measure is used to provide the weight of the rank matrix, which is the number of interactions between rows and columns. Finally, calculate the rating for each item from the rank matrix of the neighbor user. The entire process is implemented in MapReduce framework to overcome scalability problem. It takes high computational time when dealing with large amount of input data. So improvement must be done on Hadoop platform to decrease the computation time when dealing with these algorithms.

In item-based recommendation system using CF, rating is predicted based on relevant items rating by the same user [8]. User-item matrix is formed by finding relationship between different items and provides recommendation to the each user. By considering the reviews of relevant item the similarity between item-item is calculated using cosine based similarity, correlation based similarity and adjusted cosine based similarity. Finally, predicted rating for the target user is calculated.

Keyword based service recommendation system [9] which takes the preferences from the previous user keyword set and computes the similarity with the active user keyword set. Using personalized rating for each service is considered and recommends the top rated services. It does not consider the positive and negative preferences. In order to make more accurate the bigrams of words is taken.

3. 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 [9].

First step is to form candidate service list for the application along with domain thesaurus i.e. semantic words [10]. After the collection of reviews, a review sentence is given to data preprocessing. 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 review and classified as n-level orientation scale [11]. Then it used to identify the number of positive and negative opinions of each keyword by using supervised learning algorithm. 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. Finally, personalized rating for each service of the active user is calculated as shown in Figure 1 and recommend top-k rating to the active user [12]. The main steps of semantic based service recommendation system are described as follows:

3.1 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 process. Then each word is tagged using POS tagger.

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume: 02 Issue: 04, April 2017, pp. 36 – 43

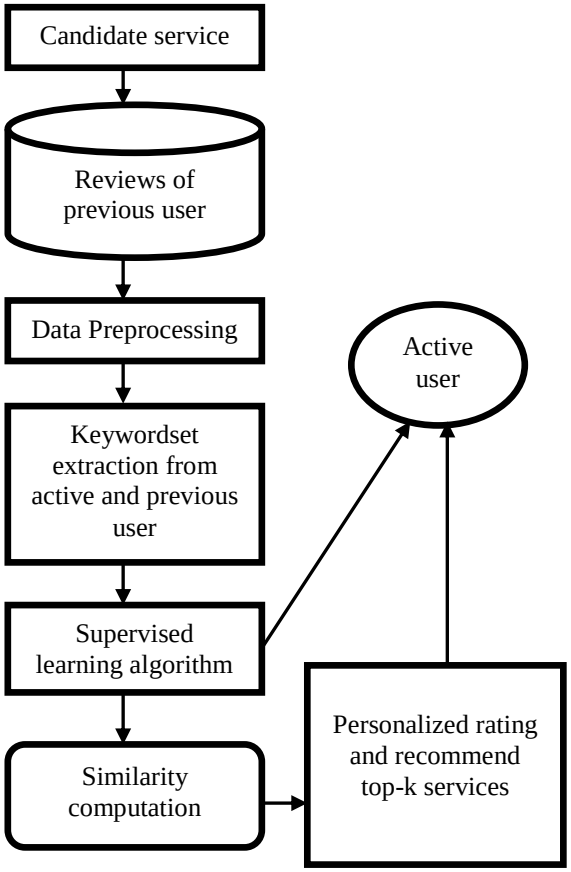

Figure 1 Proposed system design

3.2 Keyword Extraction

Active users give their preferences of service as keywords by selecting from the candidate service [13]. From the active user preference services, keyword set is formed as Active Preference Keyword set (APK). Then correspondingly previous reviews will be transformed as Previous Preference Keyword (PPK) set along with semantic words. Keywords tagged as noun by POS tagger are extracted from the datasets [14] and check for the most frequent keyword using minimum support count. The keyword extraction algorithm is shown as follows:

3.2.1 Keyword Extraction Algorithm

```
keyword extraction (pos tagged input reviews)
      if word is noun then
          extract (word)->e(i)
      endif
count frequency f(i) of each word in e(i)
set a minimum support count c(i)
      if  f(i)>c(i)
          display (word)
      else
          remove (word)
      endif
```

3.3 Keyword Orientation

Bayes theorem calculates probability using supervised term counting based approach on sentence level opinion mining. It is used to identify keyword orientation by determining whether a given review is a positive, negative or neutral using opinion word

The algorithm from [16], gives the probabilities of each label according to the words in reviews. Steps are as follows:

- Create two databases, first with the positive and negative opinion words and next with the review sentences.
- Split the sentence into the combination of words. First the 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 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
```

TABLE I EXAMPLE FOR HOTEL REVIEW

S.No.	Steps	Example		
1	A review text	room is good		
2	Stop Words Removal	room good		
3	POS Tagging	room_NN good_JJ		
4	Nouns, Adjectives, and Adverbs	room, good		
5	Keyword Extraction	room		
6	Opinion Words	Good		
7	Sentence Orientation	Positive		
8	Keyword Orientation	Keyword	Positive	Negative
		room	1	0

Example for labeling the word either as positive or negative is shown in Table I using supervised term counting based approach by naïve Bayes. Table I de-

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume: 02 Issue: 04, April 2017, pp. 36 – 43

scribes the sequence of previous steps carried out in paper for hotel review.

3.4 Jaccard Similarity Measure

Jaccard similarity is an approximation method used for finding similarity between APK and PPK. It 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. To calculate the similarity between APK and PPK, the jaccard similarity measure is given in algorithm is follows:

$$sim\ (APK,PPK) = \frac{|\ APK \bigcap PPK\ |}{|\ APK \bigcup PPK\ |} \quad (1)$$

return sim (APK,PPK)

In above Equation (1), similarity between APK and PPK is given as, number of common keywords in APK and PPK divided by the number of all the keywords in APK and PPK.

3.5 Cosine Similarity Measure

It is an exact similarity method to find the highest similarity between active preference keyword set and previous preference keyword set. The number of occurrence of the particular keyword 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. $w_{ij} = 0$). 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. It can 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], [9].

TF-IDF in which Term Frequency (TF) takes the distinct keywords and number of times the particular keyword appears in the review and in the active keyword set is given by the following Equation (2),

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

where, N_{pki} number of times particular keyword appears in the keyword set, g is the number of keywords in the preference keyword set. 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 the following Equation (3),

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

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 keyword is calculated by the Equation (4) as follows,

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

The weight of APK and PPK (w_{pk_i}) is used to calculate the cosine similarity in the Equation (5) defined as follows,

$$\text{sim(APK,PPK)} = \cos\left(\vec{W}_{AP}, \vec{W}_{PP}\right) \quad (5)$$

where, $\vec{W}_{AP}$ and $\vec{W}_{PP}$ be the weight of the keyword in the keyword set of the active preference and previous preference. The above Equation (5) can also be written as in Equation (6).

$$\text{sim(APK,PPK)} = \frac{\vec{W}_{AP} * \vec{W}_{PP}}{\| \vec{W}_{AP} \|_2 * \| \vec{W}_{PP} \|_2} \quad (6)$$

In cosine similarity method, similarity between *APK* and *PPK* is given as multiplication of weight vector of active preference with weight vector of previous preference divided by the square root of weight vector of active preference with the weight vector of previous preference.

3.6 Personalized Rating

Using CF algorithm [9], rating of each service is provided based on the cosine similarity value. The previous keyword set which is exact similar to the active keyword set is filtered out from cosine similarity. Rating of each keyword using cosine similarity is calculated and it is used to provide the top-k rated service to the active user. The personalized rating for each service of the active user is calculated as in Equation (7)

$$pr = \bar{r} + k \sum_{PPK_j \in R} sim(APK, PPK_j) * (r_j - \bar{r}) \quad (7)$$

where, $\bar{r}$ be the average rating of service, r_j be the corresponding rating of the different previous user,

sim(APK,PPK$_j$) is the similarity between APK and PPK using cosine similarity measure. *k* is the normalizing factor and *R* is used to store the previous user

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume: 02 Issue: 04, April 2017, pp. 36 – 43

after each filtration and it is calculated using the Equation (8) as follows:

$$k = \frac{1}{\sum_{PPK_J \in R} sim(APK, PPK_J)} \quad (8)$$

4. EXPERIMENTAL EVALUATION

The dataset used in the experiment is real dataset taken from the UCI repository. Table II represents the description about dataset used in this work.

TABLE II DATASET DESCRIPTION

Dataset	Total reviews	Number of users	Category
Hotel dataset	4,35,666	4,4676	688 hotels
Car dataset	2,87,330	863	601 cars

The total input is split into 80% as training data with 20% as test data. The accuracy for keyword extraction is measured by parameters precision, recall and F-measures [16] as shown below,

$$Precision = \frac{|RelevantValues \cap RetrievedValues|}{|RetrievedValues|} \quad (9)$$

$$Recall = \frac{|RelevantValues \cap RetrievedValues|}{|RelevantValues|} \quad (10)$$

$$F-measure = \frac{2 * Recall * Precision}{(Recall + Precision)} \quad (11)$$

Equation (9), *Precision* is given by the intersection of number of values extracted with the number of relevant values obtained divided by the number of extracted elements. Equation (10), *Recall* is given by the intersection of number of extracted values with the number of true values obtained is divided by the number of true values. Figure 2, represents accuracy by precision, recall, F-measures values, which can be calculated for 20 sample reviews. Out of 20 review sentences, 10 are relevant keyword, 12 are retrieved keyword and 9 are intersection of relevant and retrieved keyword for hotel dataset. Similarly for car dataset, 7 are relevant keyword, 8 are retrieved keyword and 9 are intersection of relevant and retrieved keyword.

Keyword extraction for hotel dataset gives higher accuracy of 81%. For car dataset, accuracy measures about 80.36%. From the analysis, it shows that extraction of keyword for hotel dataset is more relevant to the user needs.

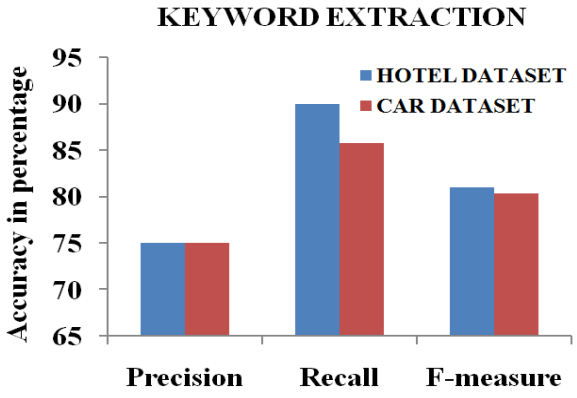

Figure 2 Keyword extraction

Figure 3, shows the outcome of number of keywords based on minimum support count. Keywords are extracted for the support count of 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 and 16, in which some of the keywords are not related to dataset keyword. If the support count is greater than 14, there is a chance to ignore some of the keywords. So, the support count is set from 9 to 14 for both dataset.

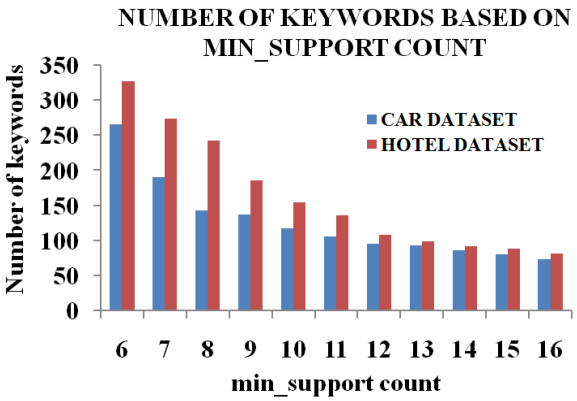

Figure 3 Number of keywords based on min_support count

Keywords with the min_support count of 10 are used for further analyses in the paper. For the support count of 10, hotel dataset contains 154 keywords and car dataset consist of 117 keywords. For each keyword, number of positive and negative label is identified using naïve Bayes. Figure 4, provides the results of keyword orientation in terms of number of positive label for three keyword (*large*, *spacious*, *cheaper*) in hotel reviews. The result is shown for the three keywords with highest positive label from the other keywords. Keywords of *large*, *spacious* and *cheaper* are the most needed keyword by the active user.

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume: 02 Issue: 04, April 2017, pp. 36 – 43

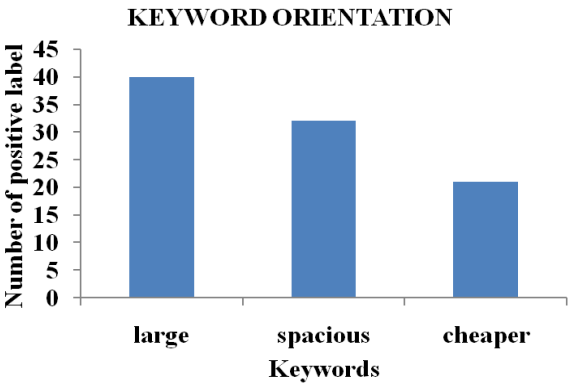

Figure 4 Number of positive opinions for hotel dataset

Similarly, the results of keyword orientation for 4 keywords (*engine, cost, brakes, wheel*) in car reviews is shown in Figure 5. The active user gives preferences as *cost, engine, brakes, motor*, and *wheel* for the car dataset. The result is taken for similarity computation of APK with three different PPK keyword sets using jaccard and cosine similarity measures.

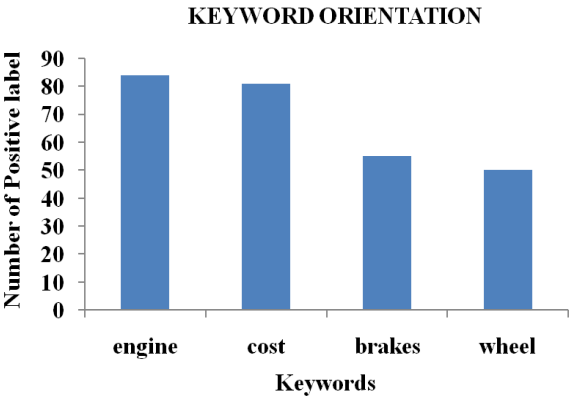

Figure 5 Number of positive opinion for car dataset

Figure 6, specifies the similarity value calculated for jaccard similarity measures and cosine similarity measures for 3 keywords (*large, spacious, cheaper*). The result shows that cosine similarity measures provide the highest value for the keywords than jaccard similarity measures in hotel dataset.

In Figure 7, similarity computation for car datasets is shown, it specifies the similarity value calculated using jaccard similarity and cosine similarity measures for keywords of 5 is (*cost, engine, brakes, motor, wheel*). The result shows that cosine similarity provides the highest similarity value for the keywords than jaccard similarity measures in car dataset.

Rating of keyword for the most similar keywords is rated using personalized rating, where the highest rating gives the most needed keyword to the active user. Semantic based service recommendation provides the most accurate rating as shown in Figure 8.

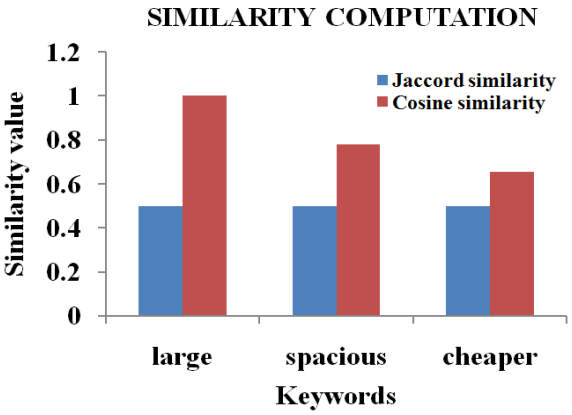

Figure 6 Similarity computation of jaccord and cosine similarity for hotel dataset

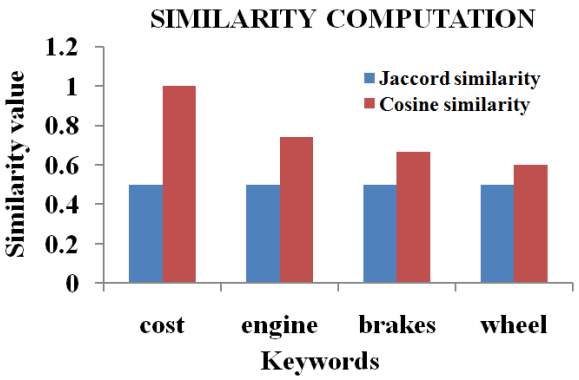

Figure 7 Similarity computation of jaccord and cosine similarity for car dataset

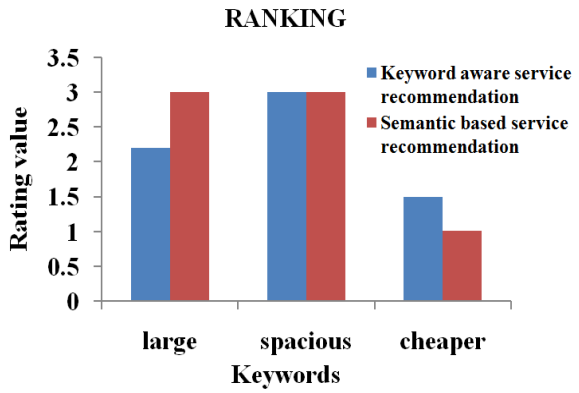

Figure 8 Ranking of keywords for hotel dataset

Figure 8, shows ranking for the keywords *large and spacious* which is higher than *cheaper* as needed by the active user preference. Semantic based service recommendation provides higher and better rating to the active user than Keyword-Aware service recom-

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume: 02 Issue: 04, April 2017, pp. 36 – 43

mendation. Similarly, for car dataset cost, *engine*, *brakes*, *wheel* provides higher rating in semantic based service recommendation than keyword-aware service recommendation as shown in Figure 9.

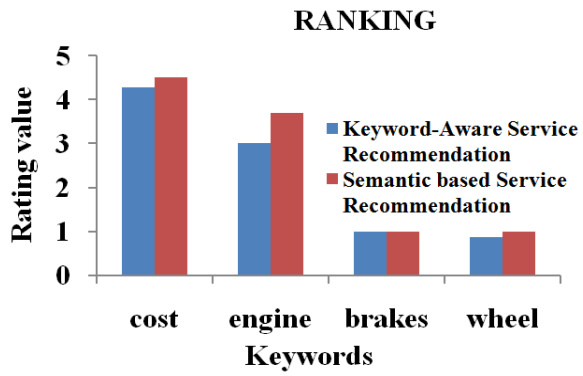

Figure 9 Ranking of keywords for car dataset

Figure 9, shows *cost* and *engine* is the highest rated keyword as needed by the active user. Figure 10, shows the execution time of single and multiple node for both hotel and car dataset. By increasing the number of nodes, execution time is decreased in multiple node clusters.

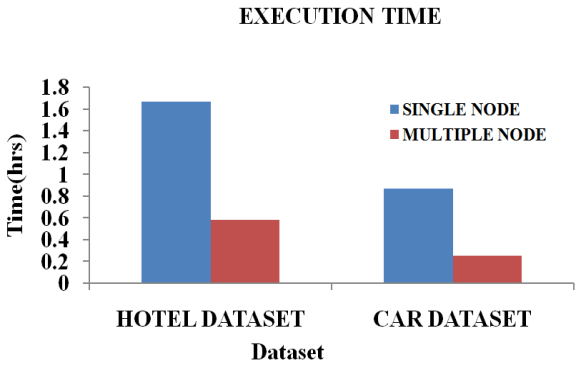

Figure 10 Execution time for dataset

5. 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 website gives overall rating of the product without analyzing the opinions on each keyword. This would make the proposed technique more complete and effective.

In future, further implementation is carried out by increasing the number of nodes to make the system more efficient and reduces the time in execution.

REFERENCES

[1] J.Manyika. (2011), "Big data: The next frontier for innovation, competition, and productivity", *McKinsey& Company Publications*.

[2] C. Lynch. (2008), "Big Data: How Do Your Data Grow?", *CNI Publication*, vol. 455, no. 7209, pp. 28-29.

[3] Watkins, Andrew B. (2005), "Exploiting immunological metaphors in the development of serial, parallel, and distributed learning algorithms", *Diss. University of Kent at Canterbury*.

[4] Liu, Bing. (2011), "Opinion mining and sentiment analysis", *Proc. Springer Berlin Heidelberg*, vol.2, pp. 459-526.

[5] Zhao, Zhi-Dan, and Ming-Sheng Shang. (2010), "User-based collaborative-filtering recommendation algorithms on hadoop", *Proc. IEEE 3rd International Conference on Knowledge Discovery and Data Mining*, vol. , pp. 478-481.

[6] Lam, Chuck. (2010), "Hadoop in action", *Manning Publications Co*.

[7] Ghemawat, Sanjay, Howard Gobioff, Shun-Tak Leung. (2003), "The Google file system", *In ACM SIGOPS Operating Systems Review*, vol. 37, No. 5, pp. 29-43.

[8] G. Linden, B. Smith, and J. York. (2003), " Amazon.com Recommendations: Item-to-Item Collaborative Filtering", *IEEE Trans.Internet Computing*, vol. 7, no. 1, pp. 76-80.

[9] Meng, S., Dou, W., Zhang, X., & Chen, J.(2014), " KASR: A Keyword-Aware Service Recommendation Method on MapReduce for Big Data Applications" , *IEEE Trans. Parallel and Distributed Systems*, vol.25, no.12, pp. 3221-3231.

[10] Turney and Peter D. (2002), "semantic orientation applied to unsupervised classification of reviews", *In Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 417-424.

[11] Hu, Minqing, and Bing Liu. (2004), "Mining opinion features in customer reviews", *AAAI* ,vol. 4. no. 4.

[12] Zhang L., Liu B., Lim S. H., & O'Brien-Strain E. (2010), "Extracting and ranking product features in opinion documents", *In Proc. of the 23rd International Conference on Computational Linguistics:* pp. 1462-1470.

[13] https://archive.ics.uci.edu/ml/datasets/OpinRank

[14] Singam, J. Amaithi, and S. Srinivasan. (2006), "optimal keyword search for recommender system in big data application", *ARPN Journal of Engineering and Applied Sciences*, vol. 10, no. 7.

[15] Khushboo, Trivedi N., Swati K. Vekariya, and Shailendra Mishra.(2012), "Mining of Sentence Level Opinion Using Supervised Term Weighted Approach of Naïve Bayesian Algorithm.", *International Journal of Computer Technology and Applications*.

[16] Naveena V., Kogilavani S.V. (2016), "Service Recommendation Based on Ranking Using Keywords in Hadoop", *Journal of Scientific Research and Reports*, vol. 10, no. 4, pp.1-9.

Authors Biography

Dr. S. V. Kogilavani, is an Assistant Professor (SRG), Department of CSE in Kongu Engineering College. She completed her BE in CSE at Mahendra Engineering College. She completed her M.E in CSE at Kongu Engineering College. She completed her Ph.D in CSE at Anna University, Chennai.

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume: 02 Issue: 04, April 2017, pp. 36 – 43

Her research interests are information retrieval, natural language processing, text mining and big data analytics.

Dr. R. Thangarajan, is an Professor, Department of CSE in Kongu Engineering College. He completed his BE in ECE at Kongu Engineering College. He completed his M.E in CSE at Government College of Technology, Coimbatore. He completed his Ph.D in CSE at Anna University, Chennai. His research interests are natural language processing, Theoretical computer science and information retrieval.

Dr. S. Malliga, Professor, Department of CSE in Kongu Engineering College. She completed her M.E. in CSE department at Kongu Enigneering College. She completed her Ph.D. in Anna University. Her research interests are Network Security, Big Data Analytics and Internet of Things.

Dr. C. S. Kanimozhiselvi, Professor, of Department of CSE in Kongu Engineering College. She completed her M.E. in CSE department at Kongu Enigneering College. She completed her Ph.D. in Anna University. Her research interests are Data Mining, Big Data Analytics, Soft Computing.