International Journal For Technological Research In Engineering
Volume 3, Issue 12, August-2016 *ISSN (Online): 2347 - 4718*

SURVEY AND ANALYSIS OF DATA MINING AND SEGMENTATION ANALYSIS FOR FREQUENT ITEM SETS

Geeta Devi[1], Nisha Yadav[2]
[1]M.Tech (CSE), [2]Assistant Professor, Department of CSE, MRKIET(Rewari)

Abstract: Data mining which is also known as Knowledge Discovery in the databases (KDD) is an important research area in today's time. It is plays an important role to retrieve valuable, hidden and predictive information from huge data sources. It is a powerful technology with great potential to analyze significant information which can be used for strategic decision making. In this paper we focused and present on the raw data analysis of the itemset number which is usually used by sites like Amazon, Flipkart, Sanpdeal, etc. In this paper we present the study and literature survey of Data mining and present work direct to the analysis of data which contains the item set number sell on the given date. The analysis is over the result the incidence of the item or couple of items frequency for selling and chances of a product sell when a couple of product already sold out. The Proposed GUI is constructed in MATLAB which is comparatively easy tool for designing the tools for this analysis.

Keywords: Opinion mining, Sentiment Analysis, Apriori Algorithm, Min-max Normalization

I. INTRODUCTION

Opinion mining is that the part of study that dissects individual opinions, sentiments, assessments, mentality, and feelings from transcription. It's force in a very range of analysts from distinctive areas of exploration together with IP, data mining, machine learning, phonetics, and even scientific discipline. Sentiment analysis or opinion mining can be defined as sub discipline of NLP (Natural Language Processing), computational linguistics and text mining that focuses on computational study of opinions, sentiments and emotions expressed in text. It covers a wide range of potential applications which includes ecommerce, politics, movies, tourism, and social networking sites etc. Hence research has been taking place in this domain for so many years [21].Data mining includes many rules for extracting data and these include association rule, clustering etc. Data mining has its applications in many fields. Some of them are science, engineering, medical, education, banking telecommunication etc [11][20].

II. LITERATURE SURVEY

In 2012, [10] Mishra et al. has classified different types of opinion mining technique which can be used for remembering classifying establishing opinion orientation of the opinionate text. The main task of opinion mining to conclude negative or positive opinion document and sentences, author observe that in compound sentence it is tedious job for opinion mining. Various task of opinion mining are (1): Task of opinion mining document level. (2): Task of opinion mining sentence level. (3): Task of opinion mining feature level. In 2014, [4] Balaji Jagtap et al proposed automate product feedback assessment system. Data collect from customer is larger in size to concluding result is tricky task. They were evaluate sentiment analysis with help of HMM and SVM base hybrid sentiment classification model with feature extraction method. Analysis helps to provide better understanding of customer opinion about product which also significant to improve the superiority of education. Hybrid approach works well for composite data.

In 2014, [7] Ritu et al. are Implement APRIORI algorithm and FP-Growth algorithm. Main propose of his work is to improve in FP-Growth algorithm and compare with both. The FP-Growth algorithm overcomes the two major disadvantage of apriori that is multiple scan of database and candidates generation. It does its work in two passes. Firstly construct the FP-tree and then generate conditional FP-trees from it. In 2015, [6] C Priyanka et al. have proposed a fuzzy logic based sentiment analysis model for fine grained classification of customer reviews into weak positive, moderate positive, strong positive, weak negative, moderate negative and strong negative classes. The model has been designed in such way that the scores from the popularly known sentiment based lexicon, SentiWordNet has been effectively used insteal of manual scoring which is tedious and laborious task. Experimental results have shown to produce accuracy in the range of 72% to 75% when tested on three datasets containing reviews of electronic gadgets. In 2016,[5] Kesavaram Kumar et al proposed approach is based on three special stages: (i) feature identification, (ii) a new SCA based method for Score calculation computed on various features in the customer opinions (iii) summarize and display the features along with appropriate score values in the order of document based. Here firstly identified the features based on ontology based feature identification method. Secondly calculated the TF-IDF values of each features and evaluate score values for each features. Finally this system ranked the features on the origin of their score rates. The performance estimation result shows that the average accuracy of correctly classified feature was found to be 81.11%. This result indicates that the proposed techniques are effective in performing their tasks. In 2016,[8] Praveen Kumar et al.have expressed importance of sentimental analysis and various approaches to achieve bag-of –words in convenient and easy way using opinions techniques and with briefly represented by a diagram. In this paper we have illustrate important field of sentiment analysis for emotion detection, building resources, transfer learning etc. In paper we also covered different view of sentimental analysis approach, their comparison with showing table and in last we have compared different sentimental analysis approach.

III. ALGORITHM USED IN OPINION MINING

A. Apriori Algorithm

A-priori calculation catches Brobdingnagian info sets amid its starting info passes and uses this outcome because the base for locating different substantial datasets amid succeeding passes. Factor sets having a bolster level over the bottom are known as intensive or visit factor sets and people beneath are known as very little factor sets. The calculation depends on the intensive factor set property that expresses: Any set of a considerable factor set is large and any set of continuous factor set should be visit. The standard calculation for mining all regular factor sets and solid affiliation principles was the AIS calculation. Once a couple of days this calculation was altered and named as apriori. A-priori calculation is, the foremost regulated and important calculation for mining incessant itemsets. A-priori utilizes width 1st inquiry and a Hash tree structure to range somebody factor sets profitably. It creates somebody factor sets of length k from factor sets of length k-1. At that time it prunes the candidates that have an occasional sub style. By dropping conclusion lemma, the contestant set contains all incessant k-length factor sets. After that, it examines the exchange info to make your mind up visit factor sets among the candidates. A-priori is an administered calculation for excavation visit factor sets for Boolean affiliation Algorithms. Subsequent to the Algorithm utilizes earlier info of continuous factor set it's been given the name A-priori. it's an repetitious level savvy ask for Algorithm, wherever k factor sets are utilized to research (k+1)- factor sets. Initially, the arrangement of frequents 1-thing sets is found. This set is indicated by L1. L1 is used to get L2, the arrangement of consecutive 2-itemsets that is used to get L3 et cetera, till not any further continuous kthing sets may be found. The finding of every L_k needs one full sweep of info.

The following are the steps involved in Apriori Algorithm.

Assume for Ck and L_k

Ck denotes candidate itemset of k size

Lk denotes frequent itemset of k size important steps of algorithm are:

- Initially get frequent set L_{k-1}
- Join step: get Ck by doing cartesian product of Lk−1 with itself
- Those itemsets which are of size (k-1) and those are not frequent should not be a subset of a frequent itemset of size k, so those should be removed
- Finally frequent set Lk has been achieved

B. Sentiment Analysis on the frequent words using Senti Word Net

In this step we have a tendency to perform Sentiment Analysis on the frequent words that we have a tendency to go from Apriori algorithmic program by victimization SentiWordNet. It provides a worth for every and each word. Sentiment Analysis deals with the usage of machine-driven techniques for anticipating the introduction of subjective substance on text reviews or comments, with usage in varied fields that has recommendation system and advertising, user intelligence and opinion retrieval. Sentiwordnet is an degree opinion vocabulary and might be thought of as extended from the Wordnet info wherever all term is connected with numerical scores demonstrating positive and negative sentiment information. This examination shows the implications of applying the Sentiwordnet lexical quality to the problem of machine-driven sentiment arrangement of client film reviews or comments.

C. Min-max Normalization

Min-Max normalization is the technique of taking data calculated in its own units and converting it to a value between 0 and 1. We use normalization because star ratings values lies between 1 to 5 and word polarity of SentiWordNet values lies between -1 and +1. Suppose we have some n rows with five variables, A, B, C, D and E, in the data. We use variable B as an example for understanding the normalization concept in the calculations below. All the other variables in the rows are normalized in the similar way. The normalized value of Bi for variable B in the ith row is calculated as: Normalized

(Bi) = Bi – Bmin/Bmax – Bmin.

Where,

Bmin = the least value for variable E

Bmax = the highest value for variable E

If Bmax = Bmin then Normalized (Bi) is set to 0.5.

IV. IMPLEMENTATION

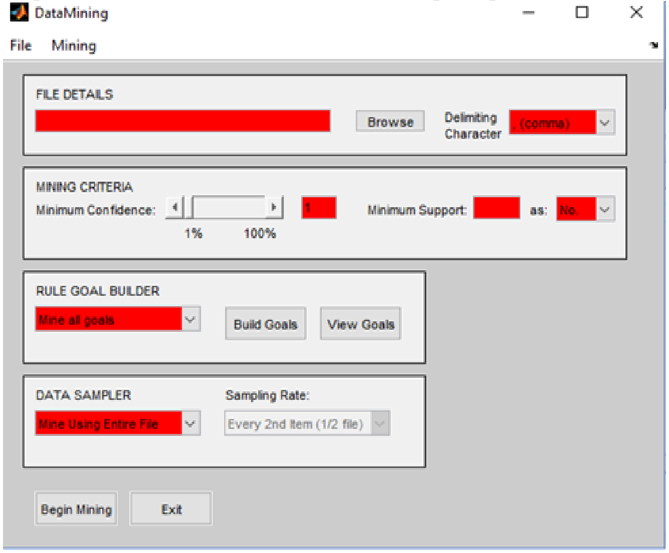
The present work extracts the opinions and reviews of product sell for Amazon data which is free source data collected from Google. It has been using here as a dummy data to analysis of proposed work. Here the technique focusing on different type’s product sells in unit. It has been noted in more than one month for product sell. This will facilitate the Amazon to find more and accurate analysis over their sell so that they can call manufactures to send or enhance the quality of product of the basis of their sell unit. Here the analysis has been collected more than 10 products sell in months. The product categories like as follows: Mobile ,Laptops ,Fabric ,Appliances ,Car & Bike Accessories, Jewellery, Luggage and bags ,Books ,Baby Products ,Software ,Sports and Fitness etc. Here sell of each month as given below as an example. Their variation is quite confusing for manual analysis if we go it 1000 times more data to put. So it is necessary to built up the required mining goal for the specific analysis. Below the list showing the item numbers of above products. The proposed system need to analysis if a consumer bay a laptop and mobile then what is the chance to bay a bag, these types of mining criteria will generally need for mining.

```
35542,65561,12345,0123,0239,09745,09786
 15951,64956,15699,99554,12569,18695
 99876,100002,100004
 23527,100004 ……..
```

Rest of data in Test Data stored in program code folder in .txt format.

A. Procedure for extracting rule and mining the data sets

Step 1: Start MATLAB
Step 2: Fetch the Code folder through current folder option
Step 3: Run the main code on command prompt



FILE DETAILS
Delimiting Character: , (comma)

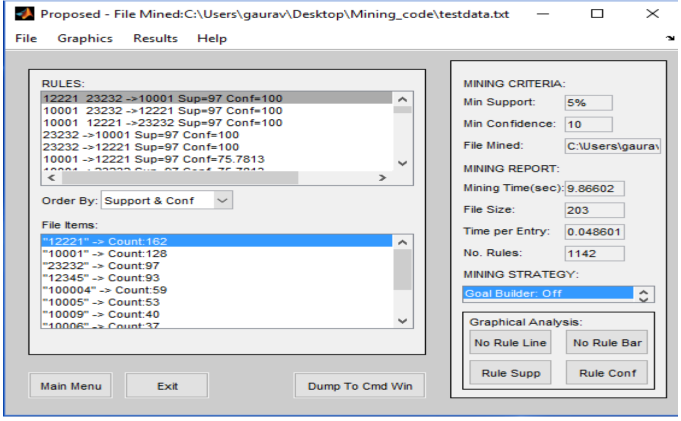
MINING CRITERIA
Minimum Confidence: 1 (1% – 100%)
Minimum Support: as: No.

RULE GOAL BUILDER
Mine all goals

DATA SAMPLER
Mine Using Entire File
Sampling Rate: Every 2nd Item (1/2 file)

Fig 1: Run proposed Data mining is showing over command prompt.

Step 4: Fetch the Amazon sell data set through browse.
Put minimum confidence level is 10 and minimum support is 5.Build all rule mining rule then next GUI come out as below



RULES:

```
12221  23232 ->10001 Sup=97 Conf=100
10001  23232 ->12221 Sup=97 Conf=100
10001  12221 ->23232 Sup=97 Conf=100
23232 ->10001 Sup=97 Conf=100
23232 ->12221 Sup=97 Conf=100
10001 ->12221 Sup=97 Conf=75.7813
```

Order By: Support & Conf

File Items:

```
"12221" -> Count:162
"10001" -> Count:128
"23232" -> Count:97
"12345" -> Count:93
"100004" -> Count:59
"10005" -> Count:53
"10009" -> Count:40
"10006" -> Count:37
```

MINING CRITERIA:
Min Support: 5%
Min Confidence: 10
File Mined: C:\Users\gaurav

MINING REPORT:
Mining Time(sec): 9.86602
File Size: 203
Time per Entry: 0.048601
No. Rules: 1142

MINING STRATEGY:
Goal Builder: Off

Fig 2: Showing the second GUI and time taken for mining will come out in command prompt.

Step5. Now it is having file and rule item, file item shows the same occurrence of items and their count which shows the maximum occurrence item in given data sets. Here support understands as how the data is frequently appears in datasets. Confidence is the condition that the how many time the rule is true. Here in Rule side bar it is showing number 23527, 99876 will provide 23232 with support of 92 percent and confidence level of 100 percent.

Step 6: For Graphical analysis of data Press all button on right side of second GUI and result of data mining will come out.

Fig.3 show the plot a graph of the scale of the LHS of the foundations against the no of Algorithms extracted. This helps to point out the proportions of Algorithm numbers for variable Algorithm sizes.

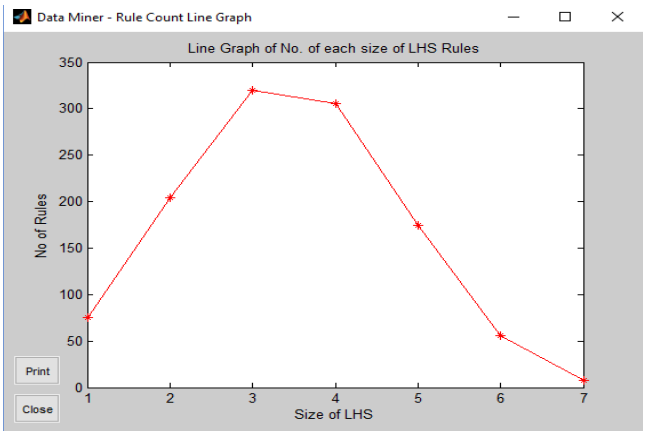


Fig 3: Showing line graph of number of each of LHS rule.

Support is the number of times the items in a rule appear together in a single entry within the entire set. Confidence is the number of times that the LHS of a rule leading to the RHS is true within the data set.

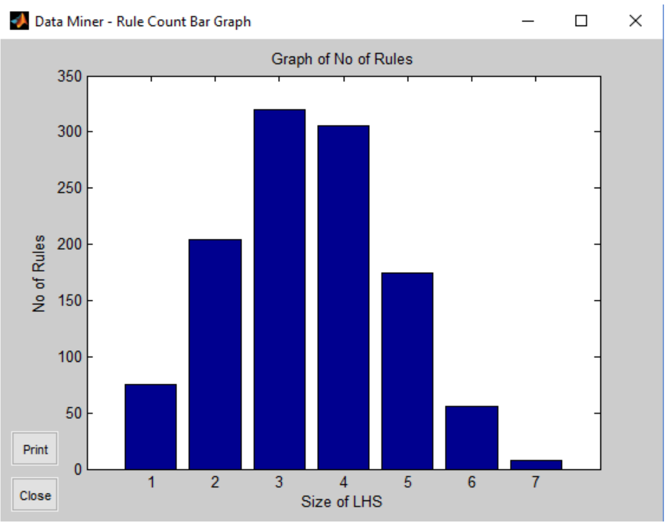


Fig 4: Showing bar graph of number of each of LHS rule

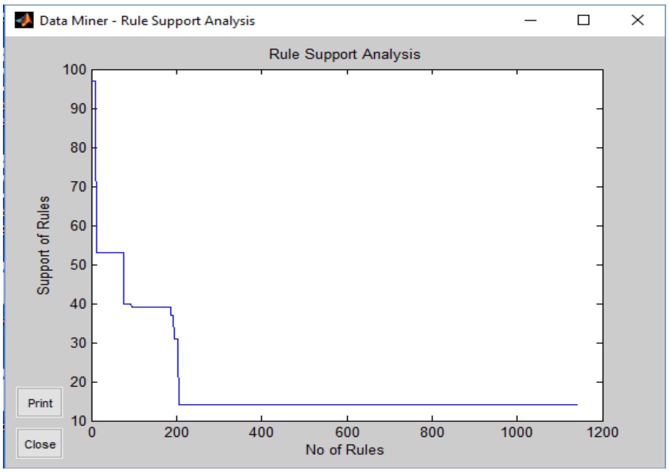


Fig 5: Showing Support of rule Vs Number of rules

Here in the graph it is clear that on increasing the number of rule the support function decreases. Means the possibilities of occurrence of item number together and the it dependent confidence will also decreases as go to higher combination.

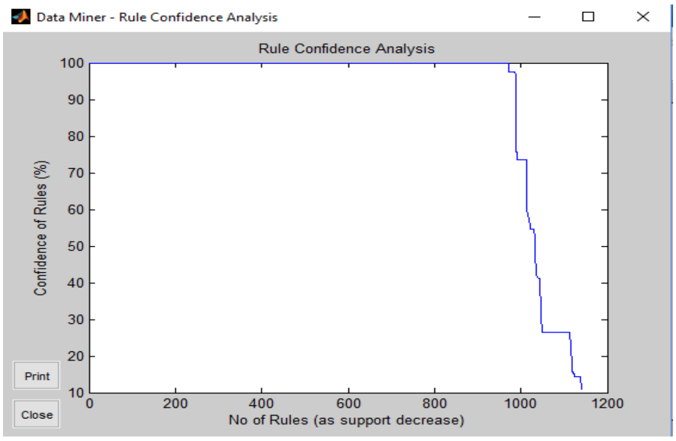


Fig 6: Rule Confidence Analysis

Confidence analysis is the occurrence of some number that is the possibilities of occurrence in group and that will be guessed that it will may occurs or not. So here seventy numbers is find that is showing the same situation .Further confidence will decrease and come to our assigned value.

V. CONCLUSION

With the advancement in web and internet technology, large amounts of data is produced every day, which made it difficult for the customers, manufacturers or even for social network users to get accurate and correct information. Which led introduction of aspect based opinion mining. In this paper, we significantly evaluated the latest models for feature based opinion mining. We highlighted the factors which are generally important for an effective and intelligent aspect mining system. In addition to this, we proposed conceptual model for efficient opinion mining system. Our present model has the capability to cover majority of the factors which determine the effectiveness of aspect mining system.

REFERENCES

[1] Asghar, M.Z., et al., “A Review of Feature Extraction in Sentiment Analysis”, Journal of Basic and Applied Scientific Research, vol. 4(3): pp. 181-186, 2014.

[2] Mukherjee, A. and B. Liu. “Aspect extraction through semi-supervised modeling”, proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. vol. I, pp. 339-348, 2012.

[3] Oeng, L., Y. Choi, and J. Wiebe. 'BenefactivelMalefactive Event and Writer Attitude

[4] Balaji Jagtap, Virendrakumar Dhotre , “ SVM and HMM Based Hybrid Approach of Sentiment Analysis for Product Feedback Assessment”,International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 3, Issue 3, 2278-6856.,2014.

[5] Kesavaram Kumar, A. Periya Nayaki, M. Indra Devi, “ Customer Feedback Evaluation System Using Feature Based Opinion Mining”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 3,pp. pp. 238-244, March 2016

[6] C Priyanka et al. “Fine Grained Sentiment Classification of Customer Reviews Using Computational Intelligent Technique”, International Journal of Engineering and Technology (IJET), Vol 7 No 4 ,pp.1453-1468, Aug-Sep 2015.

[7] Ritu1, Jitender Arora, “Intensification of Execution of Frequent Item-set algorithms”, International Journal of Recent Development in Engineering and Technology, Volume 2, Issue 6, pp. 42-45,June 2014.

[8] Praveen Kumar,Umesh Chandra Jaiswal, “A Comparative Study on Sentiment Analysis and Opinion Mining”, International Journal of Engineering and Technology (IJET), Vol 8 No 2 ,pp.938-943,Apr-May 2016.

[9] Annotation',/nProceedings of the 51st Annual Meeting of the Association for Computational Linguistics. vol. 2, pp. 120- 125, 2013.

[10] N. Mishra and C.K.Jha, “Classification of Opinion Mining Techniques,” Int. J. Comput. Appl., vol. 56, no. 13, pp. 1–6, 2012.

[11] Jagmeet Kaur1, Neena Madan “Review of Apriori Algorithm and its Recent Improvements”, International Journal of Emerging Technologies in Computational and Applied Sciences, 12(2), , pp. 150-152, March-May 2015.

[12] M. Eirinaki, S. Pisal and J. Singh, “Feature-based opinion mining and ranking”, Journal of Computer and System Sciences, vol. 78, no. 4, pp. 1 175- 1 184, 20 12.

[13] Moghaddam, S. and M. Ester. 'AQA: aspect-based opinion question answering”, Proceedings of the IEEE 11th international Conference on Data Mining Workshops (ICDMW'15), pp. 89- 96, 2011.

[14] Jingbo Zhu, Huizhen Wang, Muhua Zhu, B. Tsou and M. Ma, “Aspect-Based Opinion Polling from Customer Reviews”, IEEE Transaction on Affective Computing, vol. 2, no. I, pp. 37-49, 2011.

[15] E. Marrese-Taylor, J. Velasquez and F. BravoMarquez, “A novel deterministic approach for aspectbased opinion mining in tourism products reviews”, Expert Systems with Applications, vol. 4 1, no. 17, pp. 7764-7775, 2014.

[16] W. Zhang, H. Xu and W. Wan, “Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis”, Expert Systems with Applications, vol. 39, no. 1 1, pp. 10283- 10291, 2012.

[17] Bagheri, M. Saraee and F. de Jong, “Care more about customers: Unsupervised domainindependent aspect detection for sentiment analysis of customer reviews”, Data-Based Systems, vol. 52, pp. 20 1- 2 13, 2013.

[18] F. Xianghua, L. Guo, G. Yanyan and W. Zhiqiang, “Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon”, Data-Based Systems, vol. 37, pp.

186- 195.2013.

[19] Pefialver-Martinez, F. Garcia-Sanchez, R. ValenciaGarcia, M. Rodriguez-Garcia, V. Moreno, A. Fraga and J. Sanchez-Cervantes, “Feature-based opinion mining through ontologies”, Expert Systems with Applications, vol. 4 1, no. 13,pp. 5995-6008,2014.

[20] Ms. Rina Raval, Prof. Indr Jeet Rajput , Prof. Vinitkumar Gupta, “Survey on several improved Apriori algorithms”, IOSR Journal of Computer Engineering (IOSR-JCE), Volume 9, Issue 4 (Mar. -Apr. 2013), PP 57-61

[21] Roney Feldman, “Techniques and Applications for Sentiment Analysis”, Communications of the ACM, Vol. 56 No. 4, Pages 82- 89,2013.

[22] H. Xu, F. Zhang and W. Wang, “Implicit feature identification in Chinese reviews using explicit topic mining model”, Data-Based Systems, vol. 76, pp. 166- 175, 2015.