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.2, Special Issue 5, October 2014

Product Segmentation for Opinion Mining Using Probabilistic Principle Component Analysisin Customer Behaviors

P.Saravanakumar[1], Dr. A. Vijaya[2]

Assistant Professor, Department of Computer Applications, K.S.Rangsamy College of technology,Tiruchengode, Tamil Nadu, India[1]

Assistant Professor, Department of Computer Science, Govt. Arts College, Salem, Tamil Nadu, India[2]

ABSTRACT: Opinion mining plays a significant aspect in data mining to obtain the opinion of the user with regard to a product. Product is reviewed by the users to collect additional information about the product before they purchase that provides a strong decision to the users while purchasing a product. Works conducted on multiple reviewer-level features identified the measures for reviewers with a certain extent to subjectivity. At the same time the method Random Forest predicted the impact of reviews but did not worked with segmentation on the basis of different user opinions. The existing Variable Clustering (VC) algorithm, works on with the market segmentation for retailing based on customers' lifestyle. But VC algorithm provided with the segmentation method did not guide full-proof method for different product decision. To guide different users with variety of products, Opinion Pattern Mining Segmentation (OPMS) based on the Probabilistic Principle Component Analysis (PPCA) report is proposed in this paper. OPMS segments the pattern based on different user opinion (i.e.,) behavior where the opinion is obtained using the result of PPCA report. PPCA report determines the maximum likelihood for the user estimation on the product reviews. PPCA report usage in opinion pattern mining reduces the dimensionality on the segmentation process using the covariance matrix. Efficient segmenting of user profiles obtains the users behavioral patterns (i.e.,) opinion pattern mining with increased threshold rate and decrease the false positive. Threshold and false positive rate are examined through factor analysis in the PPCA report. Probabilistic PCA in proposed work update the product reviews based on the user behavioral reviews. Experimental work uses the OpinRank Review Dataset information for Opinion Pattern Mining and improves the segmentation efficiency up to 8 % when compared with VC algorithm. OPMS is experimented on the factors such as Opinion Decision Threshold, False Positive Rate, Segmentation efficiency and User's Product Trend Ratio Level.

KEYWORDS: Opinion Pattern Mining Segmentation, Probabilistic Principle Component Analysis, Covariance Matrix, False positive, User Behavior Pattern, Product Review

I. INTRODUCTION

Segmentation has turned out to be the primary conceptual model both in marketing theory and in practice. With the increasing use of online reviews, customers post the reviews of the products and dedicated review sites. These reviews provide excellent sources for obtaining the opinions of the valuable consumers about the products, which are very useful to both potential customers and product manufacturers. Techniques are now being developed to exploit these sources to help companies and individuals to gain such information effectively and easily. Taiwan's economy as described in [10] accompanied a model for the country's developing market. Association rule approach and clustering analysis for data mining was carried out to analyze the customers buying product in Taiwan. Knowledge extracted from data-mining results still needed to perform effective segmentation operation for promotion of customer result to the organization.

Discovery of customer relationship between huge databases has been recognized to be useful in discerning marketing, decision analysis, and business management. An important application area of opinion mining relationship is the market basket analysis, which demonstrates the buying behaviors of customers. The buying behavior searches for set of items that are frequently purchased in a given chronological order. In commerce, customer function acts significantly as a trade asset. In order to gain maximum knowledge about the customer, most of the marketing professionals involved in sales are aware of the need for businesses that include obtaining the experiences of the customer's with the help of pattern discovery.

The main idea behind the framework of pattern mining is to apply an efficient segmentation method that distinguishes the customer likeness and unlikeness of the product. By doing so, pattern mining helps to repeatedly determine the relative amount by obtaining the assessment results. However, access to such information is not straightforward since customer knowledge is largely concealed. Though they are available, but are highly un-accessible to obtain the entire volume of data that should be extracted to measure the potentiality. The greatest opportunity to access the knowledge is to use the reduced dimensionality data for building the long-term relationships with customers in a more comprehensive manner.

II. RELATED WORK

Pattern discovery approach as demonstrated in [8] presented a modern and effectual discovery and evolving technique. The process of updating ambiguous patterns improved the accuracy because the newly discovered patterns were restructured. Semantic Knowledge-Based framework as presented in [11] demonstrated the pattern discovery processing with the help of the real-world information. The framework triggered information among the different status and examined the agents but failed to handle more complex scenarios. Semantic Knowledge-Based framework did not deal with user communication behavior connected to the real world environment.Machine learning based methodology as shown in [17] built an application that was competent of recognizing and broadcasting the semantic relations but additional source of patients information were not integrated. Identifying and classifying medical related information on the web was not effective in providing valuable information to the research community (i.e.,) patients and also to the end user.The long-term relationships with customers (i.e.,) vehicle users in [15] offered a pattern-based classification. Classifying trajectories were not resourceful and effectual for performing the opinion pattern-based classification. Dark Block Extraction (DBE) as demonstrated in [20] robotically estimated the clusters using product review data sets. Dark Block Extraction developed the cluster structure using pair wise variation matrix but c-means clustering with Spatio temporal data were not carried out for the product object data clustering.

III .THE PROPOSED PROBABILISTIC PRINCIPLE COMPONENT ANALYSIS

Opinion pattern mining with PPCA report aims to establish the review of different user behavioral with orientation results and produce the result with lesser false positive and increased threshold rate. The PPCA user opinion identification is significantly used to segment the user behavioral patterns.

A. Probabilistic Principle Component Analysis Report

Based on the observation of the user behavior, PPCA report is obtained that exhibit lesser dimensionality while performing user behavior pattern segmentation. Each user behavior $U1, U2, U3 \dots .Un$ represents different dimensionalities that have to be segmented for obtaining different patterns according to the user behavior. The center of dimensionality reduction in segmentation is formularized as,

$$U_n = \frac{1}{n}\sum_{i=1}^{n} U_i - U_{i+1} \qquad (1)$$

U_n denotes 'n' user behavioral patterns whereas U_i, U_{i+1} are the obtained behavior patterns of each user. Each user behavior is taken into account for providing efficient opinion pattern mining without any dimensionality reduction. Each user's carried out the step in (1) for efficiently segmenting the user behavioral by avoiding dimensionality reduction.

$$U_i' = U_i(CM1) + U_{i+1}(CM2) \dots U_r(CMr) \qquad (2)$$

$U_i$denotes each behavioral pattern with the covariance matrix $'CM1'$ whereas U_r is the behavioral pattern with the covariance matrix 'CMr'. The value of 'CMr' is computed until the opinion of last user is obtained. Here less correlated value denotes the dimensionality reduction while performing opinion pattern mining segmentation.

B. **Maximum Likelihood with Dimensionality Reduction**

Once the behavioral pattern of the user is obtained using PPCA, the maximum likelihood with dimensionality reduction has to be obtained. The report obtained from PPCA determines the maximum likelihood for the user estimation (i.e.,) type of product reviews using the Eigen value vector principles. Let us assume that $\lambda_1, \lambda_2, \ldots\ldots \lambda_n$ denotes the Eigen value for each user behavior likelihood in order to construct the maximum likelihood function. The Eigen value with the maximum likelihood function in OPMS is used for the construction of covariance matrix. The maximum likelihood with the Eigen value function is described as,

$$Maximum\ likelihood\ Function = \frac{1}{n}\sum_{1}^{n+1} \lambda_1, \lambda_2, \ldots\ldots \lambda_n \qquad (3)$$

The maximum likelihood on the opinion pattern mining using the Eigen values $\lambda_1, \lambda_2, \ldots\ldots \lambda_n$ is performed on each user behavioral pattern. Maximum Likelihood mapping take place using the principle subspace of the observed patterns. The PPCA report transforms the pattern into reduced dimensionality pattern while performing the opinion pattern segmentation using the covariance matrix. The covariance matrix is denoted as,

$$\sum_{i,j} = cov\ (x_i\ , x_j) = E[(x_i - \lambda_i)(x_j - \lambda_j) \qquad (4)$$

Where,x_i ,x_j are arbitrary scalar vector points whereas λ_i and λ_j are Eigen values used in covariance matrix for dimensionality reduction. In PPCA report i, and j position is the covariance between i^{th} and j^{th} vector of arbitrary elements. Covariance matrix with different user behavior using probabilistic principle component analysis performs the process of dimensionality reduction.

IV .PSEUDO CODE

```
//Opinion Pattern mining with PPCA report
Begin
Input: User Input pattern 'U1, U2, U3…Un'
Output: Opinion Pattern mining with lesser false positive ratio
For Each User
Step 1: Analyze each user behavior from 'U1, U2, U3…Un'
Step 2: User behavior Segmented into 'S1','S2','S3 …'Sn'
Step 3: Opinion pattern used to attain user product reviews follows
        Step 3.1: Based on the maximum Likelihood $\frac{1}{n}\sum_{1}^{n+1} \lambda_1, \lambda_2, \ldots\ldots \lambda_n$using Eigen Values
        Step 3.2: Reduced the dimensionality using Covariance matrix
        Step 3.3: Covariance matrix $\sum_{i,j} = cov\ (x_i\ , x_j) = E[(x_i - \lambda_i)(x_j - \lambda_j)$ computed
Step 4: Probabilistic update the opinion of different user on different products
End For
Sep 5: Go to step 1
Step 6: Run Principle component Analysis until user 'Un'
End
```

V. SIMULATION RESULTS

Opinion Pattern Mining Segmentation based on the Probabilistic Principle Component Analysis (PPCA)uses JAVA platform with Weka tool for the experimental work. PPCA report uses the OpinRank Review Dataset extracted from the UCI repository for the experimental work. The OpinRank dataset contains user reviews related to car and hotels. The information is collected from the Tripadvisor and Edmunds. The Tripadisor shows the 259000 reviews and Edmunds reviewed the 42,230 reviews.

OpinRank Review Dataset contains the full review of the car model from 2007. The review holds the 140-250 cars for each year. The review data extracted the fields including the dates, author names, favorites and the full textual review. The total review is expected to be 42,230. The review of the hotel for 10 different cities such as Dubai, Beijing, London, New York City, New Delhi, San Francisco, Shanghai, Montreal, Las Vegas, and Chicago are collected. OpinRank Review Dataset has about 80 to 700 hotels in each city. The total number of reviews in the hotel is expected to be 259,000. The experiment is conducted on the factors such as Opinion Decision Threshold, False Positive Rate, segmentation Efficiency, Opinion pattern Mining Time, Dimensionality Reduction Rate and User's Product Trend Ratio Level.

Opinion decision threshold rate describes the acceptance of the customer's product review with the higher threshold rate indicating the higher opinion decision. Opinion decision threshold rate is measured in terms of (threshold %). The false positive rate of the OPMS refers to the expectation of the false positive ratio. False positive ratio is the probability of incorrectly rejecting the null suggestion for particular dataset information. False positive rate of the OPMS is defined as,

$$False\ positive\ Rate\ =\left(\frac{V}{E_0}\right)*100$$

Where V denotes the false positive rating of the product and E_0 denotes the true result obtained from the customers. Segmentation is defined as the important concept in marketing to serve different types of customer. Segmentation efficiency is measured in terms of the success percentage (success %). The opinion pattern mining time using PPCA is defined as the amount of time it takes to construct the pattern mining based on the customer opinion.

$$Pattern\ Mining\ Time\ = P1 - P2$$

Where $P1$ represents the Start time of pattern construction and $P2$ denotes the End Time of Pattern construction. Opinion pattern mining time is measured in terms of seconds (sec). In machine learning approach, dimension reduction in opinion pattern mining segmentation is the process of reducing the number of random variables while reviewing the product by the customers. Dimensionality reduction is measured as,

$$DR\ =\frac{(No.of\ dimensions - Reduced\ Dimensions)}{Processing\ Time}*100$$

User product trend ratio level denotes the amount of accurate product review attained by using the opinion pattern mining segmentation with the probability principle component analysis report.

VI.RESULT AND DISCUSSION

Opinion Pattern Mining Segmentation based on the Probabilistic Principle Component Analysis (PPCA) is compared against the Random Forest based classifier (RF) method and Variable Clustering (VC) algorithm. OPMS is evaluated using the OpinRank Review Dataset from the UCI repository.

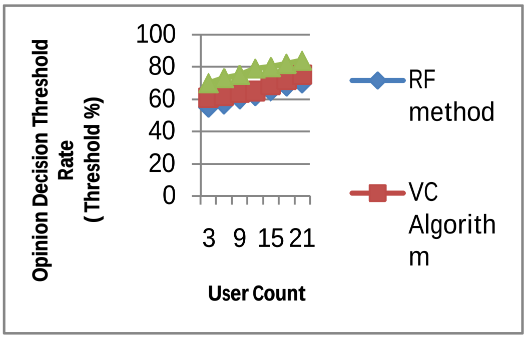

Fig 1 Performance of Opinion Decision Threshold Rate

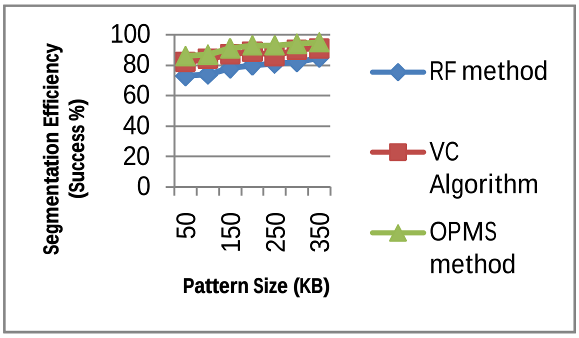

Fig 3 Segmentation Efficiency Measure

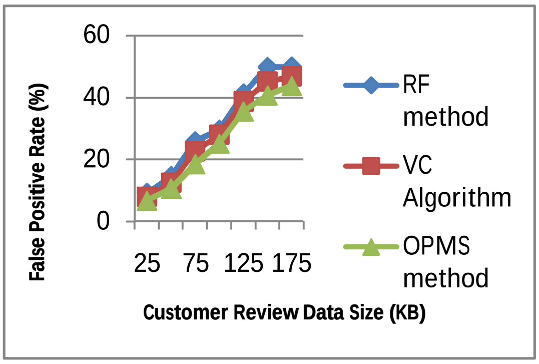

Fig 2 Measure of False Positive Rate

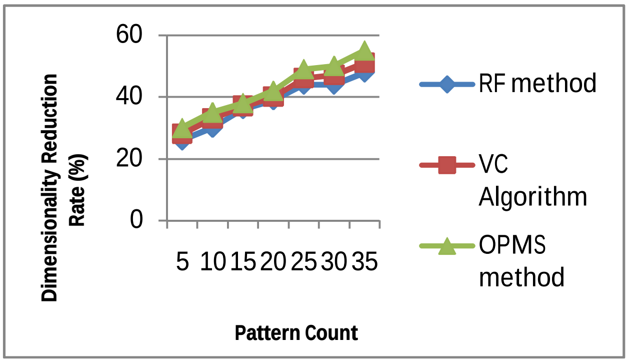

Fig 5 Dimensionality Reduction Rate Measure

VII. CONCLUSION

Opinion Pattern Mining Segmentation (OPMS) based on the Probabilistic Principle Component Analysis (PPCA) report is a valuable method which segments the useful information from large amount of data. The product review result provides clear information about the competitors to the organization using the PPCA report. Opinion pattern mining based segmentation shows significance data mining technologies to reduce the false positive rate because the data operated performed the maximum likelihood mapping of the user behavior. Experimental result indicates that the OPMS outperforms all the existing segmentation work with 16.15 % improved decision threshold rate and system efficiency rate. Probability PCA update the product reviews ratio level by 7.11 % based on the user behavioral reviews. OpinRank Review Dataset from the UCI repository is used to clearly obtain the experimental result of OPMS with existing system on the parametric factors such as opinion pattern mining time, false positive rate, and dimensionality reduction rate.

REFERENCES

[1] AnindyaGhose., Panagiotis G. Ipeirotis., “Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics,” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2010

[2] V.L. Migueis., A.S. Camanho., Joao Falcao e Cunha., “Customer data mining for lifestyle segmentation,” Expert Systems with Applications., Elsevier Journal., 2012

[3] ArchanaTomar., VineetRichhariya., Mahendra Ku. Mishra., “ A Improved Privacy Preserving Algorithm using the Association rule mining in centralized database.,” International Journal of Advanced Technology & Engineering Research (IJATER) ISSN NO: 2250-3536 VOLUME 2, ISSUE 2, 2012

[4] Marco Muselli., and Enrico Ferrari.,“Coupling Logical Analysis of Data and Shadow Clustering for Partially Defined Positive Boolean Function Reconstruction,” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 1, JANUARY 2011

[5] K.VenkateswaraRao., A.Govardhan., and K.V.ChalapatiRao.,“ Spatiotemporal Data Mining:issues, tasks and Applications.,” International Journal of Computer Science & Engineering Survey (IJCSES) Vol.3, No.1, 2012

[6] SwarupaPanmetsa., L.V.S.S., Ch. Raja Ramesh, "Anonymization of the Sequential Patterns in Location Based Service Environments," International Journal of Computer Technology & Research, IJCTR, ISSN 2319-8184,Vol 1, Issue 1, October 2012

[7] NingZhong., Yuefeng Li., Sheng-Tang Wu., "Effective Pattern Discovery for Text Mining," IEEE Transactions on Knowledge and Data Engineering, Volume:24, Issue: 1, Jan 2012

[8] Eric Hsueh-Chan Lu, Vincent S. Tseng., and Philip S. Yu., "Mining Cluster-Based Temporal Mobile Sequential Patterns in Location-Based Service Environments,"IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 6, JUNE 2011

[9] Chih-HaoWena.,, Shu-Hsien Liao., Wei-Ling Chang., Ping-Yu Hsu., "Mining shopping behavior in the Taiwan luxury products market," Expert Systems with Applications., Elsevier Journal., 2012

[10] Emilio Miguelanez., Pedro Patron., Keith E. Brown., Yvan R. Petillot., and David M. Lane., "Semantic Knowledge-Based Framework to Improve the Situation Awareness of Autonomous Underwater Vehicles," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 5, MAY 2011

[11] Panagiotis Papadimitriou., Panayiotis Tsaparas, ArielFuxman., and LiseGetoor., "TACI: Taxonomy-Aware Catalog Integration," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 7, JULY 2013

[12] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo.,"Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 4, APRIL 2013

[13] Jae-Gil Lee., Jiawei Han., Xiaolei Li, and Hong Cheng., "Mining Discriminative Patterns for Classifying Trajectories on Road Networks," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 5, MAY 2011

[14] Wu-Jun Li., and Dit-Yan Yeung., "MILD: Multiple-Instance Learning via Disambiguation," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 1, JANUARY 2010

[15] OanaFrunza., Diana Inkpen., and Thomas Tran., "A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 6, JUNE 2011

[16] Wenjing Zhang., and XinFeng., "Event Characterization and Prediction Based on Temporal Patterns in Dynamic Data System," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING., 2013

[17] Jung-Yi Jiang.,Ren-JiaLiou., and Shie-Jue Lee., "A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 3, MARCH 2011

BIOGRAPHY

P.SARAVANA KUMAR, is currently working as an Assistant Professor in Master of Computer Applications in K.S.Rangasamy College of Technology,Tiruchengode,Namakkal. He is handling PG classes. He is also pursuing Ph.D., in Computer Science under Periyar University, Salem-11. He presented a paper in International conferences and also he published in International Journal. His area of interest is DataMining.

Dr.A.VIJAYA KATHIRAVAN is working as an Assistant Professor in Computer Applications in PG and Research Department of Computer Science, Govt. Arts College (Autonomous), Salem-07, TamilNadu, INDIA. She received her M.Phil. in Computer Science from Bharathiar University, Coimbatore and she awarded her doctoral degree in Computer Applications from University of Madras, Chennai. She has published 6 Books, 3 papers in National Journal, 30 papers in International Journal, 35 Papers in National Conference Proceedings, 38 Papers in International Conference Proceedings and a total of 112 publications. Her research interests include data structures and algorithms, data/text/web mining, search engines, web communities, social network mining, machine learning, Natural Language Processing, Organizational leadership and human resource management.