International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 26 Issue 5 – AUGUST 2019.

Characterizing and Predicting Early Reviewers for Effective Product Marketing on E-Commerce Websites

K. PRIYANGA [#1] and R. RAJA YOGESWARI [*2]

[#] *M. Phil. Scholar, Dept. of Computer Science, PRIST University, Thanjavur Campus, India*

[*] *Computer Science & Computer Application, PRIST University, Thanjavur Campus, India*

***Abstract*— Online reviews have become an important source of instruction for users before manufacture an informed procure decision. Early reviews of a product tend to have a high effect on the ensuing product sales. In this paper, we take the initiative to study the behavior characteristics of early reviewers through their posted reviews on two real-world large e-commerce platforms, i.e., Amazon and Yelp. In specific, we divide product lifetime into three uninterrupted phase, namely early, majority and straggler. A user who has posted a review in the early stage is contemplating as an untimely observer. We quantitatively characterize early reviewers based on their rating behaviors, the helpfulness scores received from others and the correlation of their reviews with product popularity. We have found that (1) an early observer tends to assign a higher average rating score; and (2) an early observer tends to post more helpful reviews. Our analysis of product reviews also indicates that early reviewers' ratings and their received helpfulness scores are likely to influence product popularity. By viewing review posting process as a multiplayer competition game, we present a novel margin-based embedding model for early reviewer divination. Extensive experiments on two different e-commerce datasets have shown that our proposed approach outperforms a number of aggressive baselines.**

Index Terms*— *Early reviewer, Early review, Embedding model.

I. INTRODUCTION

The emergence of e-commerce websites has enabled users to publish or share purchase experiences by posting product reviews, which usually contain useful opinions, comments and feedback towards a product. As such, a majority of customers will read online reviews before making an informed purchase decision. It has been reported about 71% of global online shoppers read online reviews before purchasing a product. Product reviews, especially the early reviews (i.e., the reviews posted in the early stage of a product), have a high impact on subsequent product sales we call the users who posted the early reviews early reviewers. Although early reviewers contribute only a small proportion of reviews, their opinions can determine the success or failure of new products and services. It is important for companies to identify early reviewers since their feedbacks can help companies to adjust marketing strategies and improve product designs, which can eventually lead to the success of their new products. For this reason, early reviewers become the importance to monitor and attract at the early stimulation phase of a company.

The pivotal role of early reviews has attracted extensive attention from marketing professional to convince consumer purchase neutral. For example, Amazon, one of the largest e-commerce company in the world, has advocated the Early Reviewer Program, which helps to acquire early reviews on products that have few or no reviews. With this program, Amazon shoppers can learn more about products and make smarter buying decisions. As another related program, Amazon Vine2 invites the most trusted reviewers on Amazon to post opinions about new and prerelease items to help their fellow customers make informed purchase decisions. Based on the above conversation, we can see that early reviewers are especially important for product marketing. Thus, in this paper, we take the originality to study the deportment characteristics of early reviewers through their posted reviews on illustrative e-commerce platforms, e.g., Amazon and Yelp. We aim to conduct effective analysis and make accurate prognostication on early reviewers. This problem is strongly related to the adoption of innovations. In a generalized view, review posting process can be considered as an adoption of innovations, which is a theory that seeks to explain how, why, and at what rate new ideas and technology spread. The analysis and detection of early adopters in the diffusion of innovations have attracted much attention from the research community. Three fundamental elements of a diffusion process have been studied: attributes of an innovation, communication channels, and social network structures. However, most of these studies are theoretical examination at the macro level and there is a lack of quantitative explorations. With the rapid growth of online social platforms and the availability of a high volume of social networking data, studies of the diffusion of innovations have been widely conducted on social networks. However, in many application domains, social networking links or communication channel are unobserved. Hence, existing methods relying on social network structures or communication channels are not suitable in our current problem of predicting early reviewers from online reviews.

To model the deportment of early reviewers, we develop a upstanding way to indicate the assumption process in two real-world large review datasets, i.e., Amazon and Yelp. More specially, given a product, the reviewers are sorted according to their timestamps for publishing their reviews.

Following, we divide the product lifetime into three consecutive stages, namely early, majority and laggards. A user who has posted a review in the early phase is considered as an early commentator. In our work here, we mainly focus on two tasks, the first task is to analyze the overall characteristics of early reviewers compared with the majority and laggard reviewers. We characterize their rating behaviors and the helpfulness scores received from others and the correlation of their reviews with product popularity. The second task is to learn a forecast model which forecast early reviewers given a product. To analyze the characteristics of early reviewers, we take two important metrics associated with their reviews, i.e., their review ratings and helpfulness scores assigned by others. We have found that (1) an early reviewer tends to assign a higher average rating score to products; and (2) an early reviewer tends to post more helpful reviews. Our above findings can find relevance in the classic principles of personality variables theory from social science, which mainly studies how innovation is spread over time among the participants : (1) earlier adopters have a more favorable attitude toward changes than later adopters; and (2) earlier adopters have a higher degree of opinion leadership than later adopters. We can relate our findings with the personality variables theory as follows: higher average rating scores can be considered as the favorable attitude towards the products, and higher helpfulness votes of early reviews given by others can be viewed as a proxy estimate of the perspective leadership. Our analysis also indicates that early reviewers' ratings and their received helpfulness scores are likely to influence product popularity. We further explain this finding with the herd behavior widely studied in economics and sociology. Herd behavior refers to the fact that individuals are strongly influenced by the decisions of others.

II. LITERATURE REVIEW

Ting Bai, Jian-Yun Nie[1] provided a an early reviewer tends to assign a higher average rating score; and (2) an early reviewer tends to post more helpful reviews. Our analysis of product reviews also indicates that early reviewers' ratings and their received helpfulness scores are likely to influence product popularity. In viewing review posting procedure as a multiplayer competition game, we propose a novel margin based embedding model for early reviewer forecast. Experimenting on two different e-commerce datasets have shown that our proposed system outperforms a number of competitive baselines.

Julian McAuley, Alex Yang[2] Provided a Online audits are regularly our first port of call while considering items and buys on the web. While assessing a potential buy, we may have a particular inquiry as a main priority. To answer such inquiries we should either swim through colossal volumes of buyer audits planning to discover one that is pertinent, or generally suggest our conversation starter straightforwardly to the network by means of a Q/A framework. In this paper we would like to meld these two ideal models: given a huge volume of beforehand addressed questions about items, we trust to consequently realize whether an audit of an item is significant to a given question. We define this as a machine learning issue utilizing a blend of-specialists compose system—here each audit is a 'specialist' that gets the opportunity to vote on the reaction to a specific question; all the while we take in an importance capacity with the end goal that 'applicable' audits are those that vote accurately. At test time this scholarly importance work enables us to surface audits that are important to new questions on request.

Matthew J. Salganik, Peter Sheridan Dodds, Duncan J. Watts [3] provided Collaborative filtering has proven to be valuable for recommending items in many different domains. Here, we explore the use of collaborative filtering to recommend research papers, using the citation web between papers to create the ratings matrix. We tested the ability of collaborative filtering to recommend citations that would be suitable for additional references to target a research paper. We analyzed six methods for selecting citations, evaluating this through offline demonstration against a database of over 186,000 research papers hold in Research Index. We also performed an online demonstrate with over 120 users to measure user opinion of the effectiveness of the algorithms and of the utility of such recommendations for common research tasks. We came across large differences in the accuracy of the algorithms in the offline experiment, especially when balanced for coverage. In the online experiment, users felt they received quality recommendations, and were enthusiastic about the idea of receiving recommendations in this domain.

Julian McAuley, Christopher Targett, Qinfeng ('Javen') Shi, Anton van den Hengel[4] intrigued here in revealing connections between the appearances of sets of items, and especially in displaying the human idea of which objects supplement each other and which may be viewed as satisfactory options. We accordingly try to demonstrate what is an on a very basic level human idea of the visual connection between a couple of articles, as opposed to just displaying the visual similitude between them. There has been some enthusiasm generally in displaying the visual style of spots, and objects. We, interestingly, are not looking to show the individual appearances of objects, yet rather how the presence of one question may impact the attractive visual characteristics of another.

Daichi Imamori , Keishi Tajima [5] provided approach for concept Due to the dynamicity, new well known records consistently show up and vanish in miniaturized scale blogging administrations. Early identification of new records that will wind up mainstream in future is an essential issue that has a few applications, for example, slant location, viral showcasing, and client suggestion. Estimation of prominence of a record is additionally valuable for approximating the nature of data it posts. Estimation of the nature of data is vital in numerous applications, yet it is for the most part hard to gauge it without human mediation. Comparative thought has additionally been effectively connected to small scale web journals with connecting capacities. These certainties demonstrated that there is high relationship between the notoriety and the nature of data. In this manner, the estimation of forthcoming notoriety of new records, which have not yet settled the prevalence they merit, is additionally

helpful for estimation of the quality.

III. PROPOSED WORK

A. *Frequency based Itemset Mining*

Regular itemset mining is a conventional and significant problem in data mining. An itemset is repeated if its support is not less than a brink stated by users. Conventional regular itemset mining approaches have chiefly regarded as the crisis of mining static operation databases. In the operation data set regular itemsets are the itemsets that happen often. To recognize all the regular itemsets in a operation dataset is the objective of Frequent Itemset Mining. Within the finding of relationship rules it created as a phase, but has been simplified autonomous of these to several other samples. It is confronting to enlarge scalable methods for mining regular itemsets in a huge operation database as there are frequently a great number of diverse single items in a distinctive transaction database, and their groupings may form a very vast number of itemsets.

B. *Utility based Itemset Mining*

By seeing the circumstance of usage as précised by the user a high utility itemset is the one with utility value larger than the minimum brink utility. A wide topic that wraps all features of economic utility in data mining is known to be utility-based data mining. It includes the work in cost-sensitive education and dynamic learning as well as work on the recognition of uncommon events of high effectiveness value by itself. By maintaining this in mind, we at this point offer a set of algorithms for mining all sorts of utility and frequency based itemsets from a trade business deal database which would considerably aid in inventory control and sales promotion. Consideration of a utility based mining approach was motivated by researchers due to the limitations of frequent or rare itemset mining, which permits a user to suitably communicate his or her views regarding the usefulness of itemsets as utility values and then find itemsets with high utility values higher than a threshold. Identifying the lively customers of each such type of itemset mined and rank them based on their total business value can be done by these set of algorithms. This would be enormously supportive in developing Customer Relationship Management (CRM) processes like campaign management and customer segmentation. In all types of utility factors like profit, significance, subjective interestingness, aesthetic value etc the utility based data mining is a newly absorbed research area. This can add economic and business utility to existing data mining processes and techniques. A research area inside utility based data mining identified as high utility itemset mining is intended to discover itemsets that introduce high utility.

C. *Correlation feature selection*

Feature selection is a preprocessing step to machine learning which is constructive in diminish dimensionality, detach immaterial data, increasing learning perfection, and improving result comprehensibility.

1) *Steps of feature selection*

A feature of a subset is good if it is highly correlated with the class but not much correlated with other features of the class.

Steps: a. Subset generation: We have used four classifiers to rank all the characteristics of the data set. Then we have used top 3, 4, and 5 characteristics for classification.

b. Subset evaluation: Each classifier is applied to generated subset.

c. Stopping criterion: Testing process continues until 5 characteristics of the subset are selected.

d. Result validation: We have used 10-fold cross acceptance method for testing each classifier's accuracy.

D. *Classification techniques*

1) *NBTree*

NBTree is a simple hybrid algorithm with Decision Tree and Naïve-Bayes. In this algorithm the smple concept of recursive partitioning of the schemes remains the same but here the difference is that the leaf nodes are naïve Bayes categorizers and will not have nodes predicting a single class.

2) *Naïve Bayes*

The Naïve Bayes classifier technique is used when dimensionality of the inputs is high. This is a easy algorithm but gives good output than others. We are using this to find the dropout of students by calculating the probability of each input for a predictable state. It trains the weighted training data and also helps prevent over fitting.

3) *Instance-based-k-nearest neighbor*

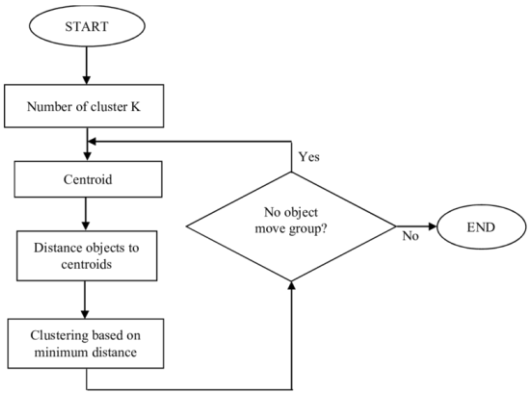
In this technique a new item is divided by comparing the memorized data items using a distance measure. For this we require storing of a dataset. Matching of items is happened by putting them close to original item. Nearest neighbors can be happened by using cross-validation either automatically or manually.

IV. ALGORITHM

A. *K-MEANS ALGORITHM*

When the data space X is RD and we're using Euclidean distance, we can represent each cluster by the point in data space that is the average of the data assigned to it. Since each cluster is represented by an average, this approach is called K-Means. The K-Means procedure is among the most popular machine learning algorithms, due to its simplicity and interpretability. Pseudocode for K-Means is shown in Algorithm 1. K-means is an algorithm that loops until it converges to a (locally optimal) solution.

Within each loop, it creates two kinds of updates: it loops over the responsibility vectors rn and modify them to point to the closest cluster, and it loops over the mean vectors μk and modify them to be the mean of the data that currently belong to it. There are K of these mean vectors (hence the name of the algorithm) and you can think of them as "prototypes" that describe each of the clusters. The basic idea is to find a prototype that describes a group in the data and to use the rn to assign the data to the best one. In the compression view of K-Means, you can think of replacing your actual datum xn with its prototype and then trying to find a situation in which that doesn't seem so bad, i.e., that compression will not lose too much information if the prototype accurately reflects the group.



Flow Chart K-Means Algorithm

1) Methods for k-means clustering

Let $X = \{x_1, x_2, x_3, \ldots\ldots\ldots, x_n\}$ be the set of data values and $V = \{v_1, v_2, \ldots\ldots\ldots, v_c\}$ be the set of place.

1) To select *'c'* cluster place.

2) Adjust the distance between each information mark and cluster place.

3) Attach the data point to the cluster place whose pass from the cluster center is minimum of all the cluster place.

4) Recollect the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i \quad (1)$$

where, *'c_i'* represents the number of data mark in i^{th} cluster.

5) Recollect the distance between each data mark and access new cluster place.

6) If no data mark was changed then stop, otherwise repeat from step 3).

2) DIS ADVANTAGES

1) The learning algorithm requires apriori condition of the number of cluster place.

2) The use of limited position - If there are two greatly extending data then k-means will not be able to intention that there are two clusters.

3) The research algorithm is not unvaried to non-aligned transformations i.e. with different presentation of data we get various conclusion (data represented in form of Cartesian co-ordinates and polar co-ordinates will give other results).

4) Euclidean length part can unevenly weight underlying factors.

5) The learning algorithm maintains the local optima of conform error function.

6) Randomly deciding of the cluster center cannot lead us to the fruitful result. Pl. refer Fig.

7) Suitable only when mean is defined i.e. fails for absolute data.

8) Not able to handle data and exception.

9) Algorithm fails for non-aligned data set.

B. Naïve Bayes

Naive Bayes is a most classification algorithm for binary (two-class) and multi-class classification problems. The technique is simplest to understand when described using binary or categorical input values.

It is known naive Bayes or idiot Bayes because the calculation of the probabilities for each hypothesis are simplified to make their calculation tractable. Rather than attempting to calculate the values of each attribute value P(d1, d2, d3|h), they are assumed to be conditionally independent given the target value and calculated as P(d1|h) * P(d2|H) and so on.

This is a strong assumption that is most unlikely in real data, i.e. that the attributes do not interact. Although the approach performs surprisingly well on data where this assumption does not hold.

Representation Used By Naive Bayes Models

The representation for naive Bayes is probabilities.

A list of probabilities are stored to file for a learned naive Bayes model. This includes:

Class Probabilities: The probabilities of each class in the training dataset.

Conditional Probabilities: The conditional probabilities of each input value given each class value.

Learn a Naive Bayes Model From Data.

Learning a naive Bayes model from your training data is fast.

Training is quick because only the probability of each class and the probability of each class given different input (x) values need to be calculated. No coefficients need to be fitted by optimization procedures.

Calculating Class Probabilities

The class probabilities are easy the frequency of instances that belong to each class divided by the total number of instances.

In a most binary classification the probability of an instance belonging to class 1 would be calculated as:

P(class=1) = count(class=1) / (count(class=0) + count(class=1))

In the easiest case each class would have the probability of 0.5 or 50% for a binary classification problem with the same number of instances in each class.

1) ADVANTAGES

1. 1. They are very simple for implementing.

2. For approximate the parameters they only needs a very few amount of training data.

3. In many cases, the results are nice.

2) DISADVANTAGES

1. Modify of loss of accuracy.

2. Naive Bayes classifier cannot change dependencies because dependencies exist between variables.

V. CONCLUSION

We have studied the novel task of early reviewer characterization and prediction on two real-world online review datasets. Our actual analysis strengthens a series of theoretical conclusions from sociology and economics. We found that an early reviewer tends to assign a higher average rating score; and an early reviewer tends to post more helpful reviews. Our experiments also indicate that early reviewers' ratings and their received helpfulness scores are likely to influence product popularity at a later stage. We have adopted a competition-based viewpoint to model the review posting process, and developed a margin based embedding

ranking model (MERM) for predicting early reviewers in a cold-start setting.

REFERENCES

[1] N. Aaraj, S. Ravi, S. Raghunathan, and N. K. Jha, "Architectures for efficient face authentication in embedded systems," inProc. Design, Autom. Test Eur., Mar. 2006, vol. 2, pp. 1–6.
[2] M. D. Marsico, M. Nappi, and D. Riccio, "FARO: Face recognition against occlusions and expression variations," IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 40, no. 1, pp. 121–132, Jan. 2010.
[3] A. F. Abate, M. Nappi, D. Riccio, and G. Tortora, "RBS: A robust bimodal system for face recognition,"Int. J. Softw. Eng. Knowl. Eng., vol. 17, no. 4, pp. 497–514, 2007.
[4] N. J. Belkin, P. B. Kantor, E. A. Fox, and J. A. Shaw, "Combining evidence of multiple query representation for information retrieval," Inf. Process. Manag., vol. 3, no. 31, pp. 431–448, 1995.
[5] R. M. Bolle, J. H. Connell, S. Pananti, N. K. Ratha, and A. W. Senior, "The relation between the ROC curve and the CMC," inProc. 4th IEEE Work. Automat. Identification Adv. Technol., 2005, pp. 15–20.
[6] D. Delgado-Gomez, F. Sukno, D. Aguado, C. Santacruz, and A. ArtesRodriguez, "Individual identification using personality traits,"J. Netw. Comput. Appl., vol. 33, no. 3, pp. 293–299, May 2010.
[7] M. D. Marsico, M. Nappi, and D. Riccio, "HERO: Human ear recognition against occlusions," inProc. IEEE Comput. Soc. Workshop Biometrics—In Assoc. IEEE Conf. Comput. Vis. Pattern Recognit.—CVPR, San Francisco, CA, 18 Jun. 2010, pp. 320–325.
[8] R. Distasi, M. Nappi, and D. Riccio, "A range/domain approximation error based approach for fractal image compression," IEEE Trans. Image Process., vol. 15, no. 1, pp. 89–97, Jan. 2006.
[9] K. Sarkar and H. Sundaram, "How do we find early adopters who will guide a resource constrained network towards a desired distribution of behaviors?" in CoRR, 2013, p. 1303.
[10] D. Imamori and K. Tajima, "Predicting popularity of twitter accounts through the discovery of link-propagating early adopters," in CoRR, 2015, p. 1512.
[11] X. Rong and Q. Mei, "Diffusion of innovations revisited: from social network to innovation network," in CIKM, 2013, pp. 499– 508.
[12] I. Mele, F. Bonchi, and A. Gionis, "The early-adopter graph and its application to web-page recommendation," in CIKM, 2012, pp. 1682–1686.
[13] Y.-F. Chen, "Herd behavior in purchasing books online," Computers in Human Behavior, vol. 24(5), pp. 1977–1992, 2008.