A Time-Based Online Customer Behaviors Analysis Model

Zeqing Qin¹*, Zesong Wang²

¹ School of Computer Science, Hubei University of Technology, Wuhan, Hubei, 430068, China
² School of Computer Science and Information Engineering, Hubei University, Wuhan, Hubei, 430062, China

*Corresponding author’s e-mail: zeqing@hbut.edu.cn

Abstract. With the advent of the big data era, reviews are common when shopping online. To better understand the connection between reviews and the quality of the real product, we have to use mathematical language to simplify the problem. This relationship can help both the buyer to get what they want and the firms to make better products with attractive features. To help us understand how the indicators like star ratings can reflect the product's popularity, we construct a series of models to illustrate them. We first use the NLP method to process text type data, such as product review information, and convert these text data into sentiment scores, with the help of the k-means++ algorithm, we cluster the data into 3 categories. These categories can help the company to track the performance of their products by simply calculating which category the product belongs to. Next, a Time-based Reputation Judgment Model combined with AHP and Markov Chain Model was established to describe the change of reputation over time. In this way, a company can easily predict the future trend of their merchandise sales.

1. Introduction

With the rapid development of computer technology, online shopping has become the choice of more and more people. Compared to offline shopping, online shopping saves time and effort. Shopping online, however, can also be an unpleasant experience. To circumvent this risk, users often refer to reviews from users who have previously purchased the product. In this paper, we will use the data of these users to find the patterns of user evaluation and help online sales companies develop online sales strategies.

In the online marketplace it created, Amazon provides customers with an opportunity to rate and review purchases. Individual ratings - called “star ratings” – allow purchasers to express their level of satisfaction with a product using a scale of 1 (low rated, low satisfaction) to 5 (highly rated, high satisfaction). Additionally, customers can submit text-based messages – called “reviews” – that express further opinions and information about the product. We selected three products on Amazon, namely pacifiers, hairdryers, and microwave ovens for analysis, and constructed evaluation models to provide online companies with substantial construction suggestions. The notations are listed below in Table 1.
2. Modeling

2.1. Semantic Extraction

Since the reviews provided in the data files are textual, they cannot be quantified directly. Therefore, we need to extract the emotion and attitude of each review separately and quantify them with a ranking mechanism. As the Naive Bayesian classifier is very simple and efficient and highly sensitive to feature selection which fits the requirement of extraction of emotion[1]. Hence, we use this algorithm in supervised learning to classify and quantify customer reviews. Classification can be used to predict labels for unlabeled documents[2]. Using labeled documents, that is, document objects with type as training examples, and statistically predict the labels of new documents based on their similarity to the training examples.

The system first learns the joint probability distribution from input to output by using the given training set, with the assumption that the feature words are independent of each other. Based on the trained model, we can input X and calculated the output Y whose posterior probability is the max.

Specifically speaking, there is a sample data set \( D = d_1, d_2, \ldots, d_n \) consisting of reviews, the feature data set corresponding to the sample data is \( X = x_1, x_2, \ldots, x_n \) and the class variables are \( Y = y_1, y_2, \ldots, y_n \) which means \( D \) can be categorized as \( y_m \). In this paper, \( X \) is the semantic tag data set and the \( Y \) is whether the review can be categorized in one of the emotion tags in the semantic data set. Besides, \( x_1, x_2, \ldots, x_n \) are random and independent with each other. In the algorithm, the prior probability \( P_{\text{prior}} = P(Y) \), posterior probability \( P_{\text{post}} \), can be calculated from the \( P_{\text{prior}} \), proof \( P(X) \) and class conditional probability \( P(X|Y) \) as the following equation:

\[
P(Y|X) = \frac{P(Y)P(X|Y)}{P(X)} \tag{1}
\]

With a giving class \( y \), equation 1 can be rewritten as equation 2.

\[
P(Y|X = y) = \prod_{i=1}^{d} P(x_i|Y = y) \tag{2}
\]

Based on the equation 1 and 2, we can calculate the posterior probability:

\[
P_{\text{post}} = \frac{P(Y)\prod_{i=1}^{d} P(x_i|Y = y)}{P(X)} \tag{3}
\]

For this problem, we found two external comment sets according to the type of product as training data, the emotional color of each product review is known, where the value of the positive emotional label is 1 and the value of the negative emotional label is 0. We can use trained models to predict other reviews based on common words in the text. We could represent each review as a vector of adjectives (e.g., good, bad, awesome, awful, ...) since positive reviews (good, awesome) will most likely contain
different adjectives than negative reviews (bad, awful). After semantic extraction, all the valid reviews are labeled with a score between 0 and 1. Table 2 shows some evaluations and their corresponding scores.

| Review                  | Review_id               | Score       |
|-------------------------|-------------------------|-------------|
| Love it.                | R3P1OYS7ZM5Y            | 0.975392407 |
| Great product.          | RVWCIHYH2Z7W3Y          | 0.826613167 |
| Underpowered.           | R3OV8L0FUZD1MY          | 0.50011883  |
| Terrible power!!!       | R3K8WVSNCMWVH3          | 0.179012284 |

2.2. Cluster analysis

In the previous data preparation process, we have used NLP technology to perform semantic extraction and emotional scoring on textual data, that is, the review. In recommendation systems, the k-means algorithm is often used for user clustering and product clustering as collaborative filtering algorithms. For example, *A Survey of Collaborative Filtering-Based Recommender Systems: From Traditional Methods to Hybrid Methods Based on Social Networks* [3], uses k-means to perform cluster analysis on products. The K-means algorithm is widely used in cluster analysis. However, this algorithm requires the initial value of the cluster center, and it is sensitive to outlier data. In this circumstance, the star ratings and comments in this article have many extreme values such as low Star ratings and extreme reviews, and the initial clustering points are also difficult to manually select [4]. Therefore, for this problem, we used the modified K-means++ algorithm instead.

We use the average star rating and average sentence score of the product as two independent variables for cluster analysis. Take the pacifier data as an example, the number of products in cluster 3 is slightly greater than cluster 1 and cluster 2. Besides, the value distribution is relatively uniform in general. In order to verify the credibility of our clustering results, we performed sentence semantic extraction on the top 20 products in each of these three categories and performed word cloud analysis on the most frequent 25 words. The analysis results from left to right are the first category to the third category as figure 1:

![Figure 1. Word cloud of pacifier](image)

From the word cloud, we can see that the negative (derogatory) words in the first product category, that is, the worst product category, account for a relatively large number of words. The words in the second and third categories gradually tend to kind words, this also verifies the reliability of our Cluster Layering Model. Therefore, we can divide the products on the market into 3 categories: A, B, and C, which respectively represent the top, middle, and inferior popularity of products. To make it easier for Company to track which level the company's products belong to after the company's products are online for sales. We also provide a set of product level discrimination formulas based on European distance calculation as equation 4 shows, which is convenient for the Sunshine Company to track their product level.
2.3. Time-based Reputation Judgment Model

We want to figure out how the reputation of a commodity changes regularly over time. To this end, we have established a time-based reputation judgment model. Since there are various types of indicators in the question, such as vine is qualitative data rather than quantitative data, and the dimensions of these indicators are very different, the general evaluation model cannot be applied to this problem accurately. The Analytic Hierarchy Process (AHP) is a multiple criteria decision-making tool that judges and compares the priorities of pairwise comparisons of experts [5], which can help us determine the weighting factor of the indicators in the evaluation system. AHP has been used in almost all applications related to decision-making[6]. For example, in the paper A group AHP-TOPSIS framework for human spaceflight mission planning at NASA[7], the AHP method is used to solve the problem of human spaceflight mission planning. The Markov Model is highly reliable in the problem-based feature description[8]. The essence of a Markov chain is a sequence of random variables $X_1, X_2, X_3, ...$ that satisfy Markov properties, that is, given the current state, the future state and the past state are independent of each other. Therefore, considering the actual situation, we have established a multi-dimensional evaluation system based on the AHP algorithm combined with Markov Chain Model for this problem.

The four variables: star rating, vine, verified purchase, and review are the independent variables in the time-based reputation evaluation model and the dependent variable $Y$ is reputation, the relationship is shown in equation 5. The final ranking and the weight of each variable can be calculated by the multi-dimensional evaluation model we established.

$$Y = f(SR, VT, VP, RE)$$

We set the four indicators: star rating, vine, verified purchase, review as the result layer of AHP to calculate the weight ratio of these indicators as shown in figure 2.
In the end, we calculated the four impact factors of the star rating, vine, verified purchase, and review according to the 1:1 ratio of sales to CTR. They are: 0.3940, 0.2938, 0.1060, 0.2063. After obtaining the weights of these four variables, we can use the Markov Process to identify the pattern of reputation changing over time. Given the characteristics that the convergence of the transfer matrix in the Markov chain model and the different relationships between the commodities, we establish a Markov Chain Model with a one year to one-quarter cycle according to their difference to analyze the sales status of various typical commodities throughout the year. Additionally, there might be seasonal changes in the market, so we think it is necessary to analyze the market [9].

Since the data of the transfer matrix within a year is relatively sparse, we decided to divide it into four classes $p_1, p_2, p_3$ and $p_4$; the transfer matrix with a period of quarter requires more classification to get better fitting results, the five categories are $q_1, q_2, q_3, q_4, q_5$.

![Figure 2. Overview of Multi-dimensional Evaluation System](image)

![Figure 3. The quarterly standardized scores of products](image)
Take one-step transition probability as an example, each row of the state transition matrix represents a corresponding state. $t_{11}$ means the probability of the first row and first column in the 1-step transition probability matrix. Since $t_{11} = q_{1(1)}/s_1$, $q_{1(1)}$ represents the times $q_1$ changes to $q_1$ in one step and $s_1$ represents the total number of times that the state $q_1$ appears in the sample. Through this method, we can obtain the state transition matrix of the pacifier within the quarter.

$$
t_{11} = \begin{bmatrix}
\frac{3}{8} & \frac{1}{4} & \frac{3}{8} & 0 & 0 \\
\frac{1}{4} & \frac{1}{2} & \frac{1}{8} & 0 & \frac{1}{8} \\
\frac{2}{15} & \frac{2}{15} & \frac{3}{5} & \frac{1}{15} & \frac{1}{15} \\
0 & 0 & \frac{2}{5} & \frac{3}{5} & 0 \\
0 & \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{2}
\end{bmatrix}
$$

3. Conclusion

According to our analysis, finding the regularity and correlation among star ratings, reviews, and helpful votes is the key to the problem. When using the AHP method for weighting, in addition to considering the impact of sales on the population, we also combined an important indicator "click rate" in the recommendation system as the factor layer of the AHP weighting model. This allowed our model to be more comprehensive.

In each model, we have provided a complete set of calculation methods for sunshine company. And we have also highlighted the situations in which sunshine company only needs to change the parameters when it is used, which is more user-friendly.

References

[1] J. Chen, H. Huang, S. Tian, and Y. Qu, “Feature selection for text classification with naïve bayes,” Expert Systems with Applications, vol. 36, no. 3, pp. 5432–5435, 2009.
[2] C. R. Center, “Document for pattern package.” https://www.clips.uantwerpen.be/pages/pattern-vector#classification.
[3] R. Chen, Q. Hua, Y.-S. Chang, B. Wang, L. Zhang, and X. Kong, “A survey of collaborative filtering-based recommender systems: From traditional methods to hybrid methods based on social networks,” IEEE Access, vol. 6, pp. 64301–64320, 2018.
[4] S. Na, L. Xumin, and G. Yong, “Research on k-means clustering algorithm: An improved k-means clustering algorithm,” in 2010 Third International Symposium on intelligent information technology and security informatics, pp. 63–67, IEEE, 2010.
[5] T. L. Saaty, “How to make a decision: the analytic hierarchy process,” European journal of operational research, vol. 48, no. 1, pp. 9–26, 1990.
[6] O. S. Vaidya and S. Kumar, “Analytic hierarchy process: An overview of applications,” European Journal of operational research, vol. 169, no. 1, pp. 1–29, 2006.
[7] M. Tavana and A. Hatami-Marbini, “A group ahp-topsis framework for human spaceflight mission planning at nasa,” Expert Systems with Applications, vol. 38, no. 11, pp. 13588–13603, 2011.
[8] C. Yadav and R. Singh, “Reliability of object oriented systems with markov transfer of control,” in 2011 International Conference on Computer and Management (CAMAN), pp. 1–3, IEEE, 2011.
[9] M. S. Rozeff and W. R. Kinney Jr, “Capital market seasonality: The case of stock returns,” Journal of financial economics, vol. 3, no. 4, pp. 379–402, 1976.