Problem Chosen	2020 MCM/ICM Summary Sheet	Team Control Number
C		2020230

A Wealth of Data

Summary

To assist Sunshine Company in developing a scientific and reasonable online sales strategy, our team shall establish a SO-PMI based Review Quantization Model (RQ) and the Comprehensive Rating Model of Product Reputation (CRPR). Based on the data provided and the two models, we analyzed the influence of star rating, reviews and helpfulness rating on the sales of goods and the relationship among them.

After data processing (including index deletion, meaningless data elimination, etc.), we introduce an RQ model to quantify the text review. First, after text segmentation and elimination of stop words, we build a dictionary of positive and negative words as seed words. Then, we use the revised formula to calculate the SO- PMI value of each word in a review and obtain the quantified review.

To study the popularity of products, we introduce a CRPR model. First, we define review values with the combination of quantified review and helpfulness ratings and define identity coefficient using vine and verified_purchase. Then, use the entropy weight method to determine the weight of the index, and obtain the reputation value after linear weighting. Taking the pacifier data as an example, and the weight are 0.49474 and 0.50526, respectively. Finally, the sensitivity analysis of the identity coefficient using hair_dryer data shows that both vine and verified_purchase have almost the same impact on reputation. Moreover, our correlation analysis indicates that reviews are more important than star ratings.

To investigate the change of reputation with respect to time, we use the least square method and obtain the functional relationship of hair_dryer, microwave and pacifier. RMSE are 0.14, 0.16, and 0.09 for hair dryer, microwave, and pacifier, respectively.By analyzing the fitting curve, we conclude that the reputation of hair_dryer and microwave is stable in the later period, while pacifier fluctuates greatly.

To find potential successful products based on the functional relationship between product reputation and time, we establish an LDA(Latent Dirichlet Allocation) based product prediction model. First, we determine the reputation inflection points of 3 data set according to the fitting curves. Take hair_dryer as example, we obtain the corresponding time period as: (2012.12-2013.1, 2015.3-4). Then, we use the LDA to extract keywords from the comment sets in these time periods and employ the PMI to calculate the semantic similarity between words, thereby predicting products success or failure by combining the semantic similarity and the sales volume. Taking hair_dryer set as an example, we find product_id =b003v264ww is a potential successful product.

Also, we conduct cross correlation analysis of star ratings and review values, which indicate that the rating will trigger more reviews. Based on the theme extraction and word frequency of each star levels review by LDA, we observe that some specific words in the comments will affect the star rating.

Keywords: SO-PMI; LDA; Least Squares; Correlation analysis; Sales strategy