

MSc in Big Data Technology

Age and Sentiment Based E-Commerce Clothing Recommendation System (Project Proposal)

MSBD5003: Big Data Computing

BANERJEE, Rohini (20543577)

JUANG, Stefan Warren (20537803)

MEHROTRA, Priya (20580317)

PARK, Seoyoung (20537188)

Introduction

E-commerce has built their success story on top of their large user base and its feedbacks. From the point of arrival to the shop site to the moment of product handoff, every minute holds crucial hidden information. A strong user data support base is now the very backbone of every e-commerce business ^[1].

While it is important to collect user information, active interruption to the customers tend to cause more harm than good ^[2]. To gain an edge in the competition, we propose a less intrusive machine learning model. Furthermore, we aim to improve user identification as well as bring a more intimate shopping experience to the users.

Task Description

Via this project, we hope to develop a concise e-commerce recommendation system based on customer maturity and product sentiment. By learning the relationship between past users' comments and the wordings they tend to use, we project our model to determine the maturity of a person. With a delicate balance of past user data and model inference, we will generate pseudo data which can help us implicitly determine similarities between different users.

Based on the maturity group and user sentiment analysis, we aim to provide a novel e-commerce recommendation system.

Dataset Description

For this project, we have chosen a women's clothing e-commerce dataset from Kaggle ^[3]. With 23486 rows and 10 feature variables, the dataset revolves around the reviews written by customers. Each row corresponds to a customer review and includes the following:

1. Clothing ID: Integer categorical variable referring to the specific piece being reviewed.
2. Age: Positive integer variable of the reviewer's age.
3. Title: String variable describing the title of the review.
4. Review Text: String variable containing the body of the review.
5. Rating: Positive ordinal integer variable for the product score granted by the customer (1: Worst and 5:Best).
6. Recommended IND: Binary variable stating where the customer recommends the product (1: Recommends, 0: Does not recommend).
7. Positive Feedback Count: Positive integer denoting the number of other customers who found this review to be correct.
8. Division Name: Categorical name of the product's high level division.
9. Department Name: Categorical name of the product's department name.
10. Class Name: Categorical name of the product's class name.

Technologies Used

In order to build a robust recommendation system, an integration of multiple technologies is necessary. We propose to use the following technologies to complete the project on a wholesome level:

1. Spark SQL and Pyspark Graphframe: For data pre-processing, data cleaning and data integration.
2. Tensorflow/MLlib: Data modeling and data mining ^[4].
3. Tableau and Matplotlib: Pattern spotlight and data visualization.
4. Google Cloud Engine: To provide cloud computation as well as storage.

Conclusion

With the growing amount of user information on the internet and the significant rise in the number of users, it is important for companies to search, map and provide customers with the relevant information according to their preferences and tastes. In e-commerce businesses, this recommendation is often what converts browsers to buyers. Through our project, we hope to build a personalized recommendation system which yields both customer satisfaction and business revenue.

References

- [1] Wei, Kangning, Jinghua Huang, and Shaohong Fu. "A survey of e-commerce recommender systems." Service systems and service management, 2007 international conference on. IEEE, 2007.
- [2] Sarwar, Badrul M., et al. "Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering." Proceedings of the fifth international conference on computer and information technology. Vol. 1. 2002.
- [3] <https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews/home>
- [4] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen, et al., "Mllib: Machine learning in apache spark," Journal of Machine Learning Research, vol. 17, no. 34, pp. 1–7, 2016.