Social media-based recommender system.(Clustering/Classification)

Shubham Bhatt
Geodesy and Geomatics
University of New Brunswick
Fredericton, New Brunswick
sbhatt@unb.ca

Akshay Saxena
Geodesy and Geomatics
University of New Brunswick
Fredericton, New Brunswick
akshay.saxena@unb.ca

Shahin Nisha
Geodesy and Geomatics
University of New Brunswick
Fredericton, New Brunswick
shahin.nisha@unb.ca

***Abstract*— Online retail has a vast and growing following and significance, allowing customers to shop at any time, locate the greatest deals, and have a positive shopping experience. As a result, ample technologies have been used to develop Consumer and its Relationship Management solutions in order to better understand customer behavior demands and requirements, engage and influence them to improve their shopping habits, and increase retailer profitability. Current marketing solutions use an overly broad strategy, pushing and suggesting, in most cases, the most popular or most purchased things while ignoring customer centricity and individuality.**

A recommendation system for online retail shops is proposed in this research, which is based on a multi clustering technique of items and customer profiles in both online and physical stores. The proposed approach is based on mining techniques, which allow for the prediction of newly acquired customers' buying behaviors, hence resolving the cold start issues those systems has at the current state of the technologies.

I. INTRODUCTION

Fashion retailers are becoming more attractive and important to allow customers to shop at any time, get the best deals, and have a positive shopping experience. As a result, we use many technologies to develop a customer relationship management solution to better understand, engage, influence, improve the shopping experience and increase retailer profitability. Today's marketing solutions often take an over-the-top approach that suggests the most popular or bestsellers, ignoring the focus and personality of the customer. In this research, we propose a fashion store recommendation system. based on a multi-clustering method of items and customer profiles in both online stores and physical stores. The proposed solution relies on mining techniques that make it possible to predict the buying behavior of newly acquired customers, thereby solving the cold start problems typical of traditional systems.

The particular competitiveness of suppliers strongly will depend on the conquered reputation, brand relevance and on the marketing activities they execute. The particular latter aspect is exploited to raise the sales and so a retailer, through marketing, should be capable of promoting customers to buy more items or even more valuable items. These days, consumers tend to buy more on ecommerce and the COVID-19 situation also stressed this disorder. On-line shopping offers the probability to buy whenever you want; customers buy where they find the best offer, online as well as offline, and perhaps they are also influenced by an increasing amount of information from sites, communities, and sociable networks. To keep a customer is therefore an extremely difficult achievement, and in some measure, it can get easily out of control.

II. RELATED LITERATURE WORK

A recommendation system architecture based [1] is proposed to give semantic recommendations for each user profile. The vastness and heterogeneity of devices and their composition offer innovative services and scenarios that require a new challenging vision in interoperability, security and data management [2]. First, we focus on the system's general functionality and system architecture. We then describe the interface and design of two deployed news agents that are part of the described architecture. While the first agent provides personalized news through a web-based interface, the second system is geared towards wireless information devices such as PDAs (personal digital assistants) and cell phones. Based on implicit and explicit user feedback, our agents use a machine learning algorithm to induce individual user models. Motivated by general shortcomings of other user modeling systems for Information Retrieval applications, as well as the specific requirements of news classification, we propose the induction of hybrid user models that consist of separate models for short-term and long-term interests[3]. In [4], The categorization approach was used to identify target consumers while minimising suggestion errors by identifying individuals to whom the recommendations should be targeted based on which kinds of purchases they had made in a given time period. Cluster Analysis has been used for segmenting a heterogeneous population into a number of subgroups [5]. It is feasible to analyse and organise the data set using Clustering Analysis in order to make recommendations. The solution was created as part of Regione Toscana's Feedback project, and it was tested on real retail company Tessilform. It was validated against real data from December 2019 to December 2020, showing that using the proposed recommendation tool stimulated customers, resulting in a 3.48 percent increase in buyer attention and purchase[1]. The proposed solutions have been shown to work in the presence of a small number of customers and items (as occurs in retail shops with high-value items) and when suggestions are mediated by assistants (as occurs in fashion retail shops). Furthermore, the proposed solution tackles and resolves deficiencies and difficulties[6].

III. DATA COLLECTION

Twitter has a huge collection of data as its tweets are completely public and pullable. This makes it possible to perform analytic queries over it. For tweets extraction, Snscrape, a Python library, is a scraper for social networking

services (SNS). It can scrape data like user profiles, hashtags, or searches and returns the relevant posts. From Facebook, it can scrape user profiles, groups, and communities visitor posts, from Instagram data like user profiles, hashtags, and locations, and Twitter-like users, user profiles, hashtags, searches, tweets (single or surrounding thread), list posts, and trends, VKontakte: user profiles, Weibo (Sina Weibo): user profiles

A. ACCESSING TWITTER DATA

Twitter offers two different families of APIs to allow the public to access their data, the Streaming API and the REST API (Representational State Transfer). They differ regarding the functionality they offer and the ways they constrain a user. To access any one of them, certain prerequisites have to be fulfilled. To gain access to Twitter data, it is necessary to register a developer account. Any Twitter user can apply for a developer account. 3 Application requires a verified phone number and email address, as well as a detailed description of how the API will be used. Use of the Twitter API requires agreeing to Twitter's Developer Agreement and Policy, as well as their related policies, including the Display Requirements and Automation Rules. These agreements are in place to ensure a reasonable usage of the Twitter API and the data shared. Once a developer account has been established, a developer's Twitter application needs to be registered by providing a name, description and domain. It authenticates the end-user for the application of Twitter API and gives the end-user the access key and access token via the dashboard for app management.

B. SCRAPING TWITTER DATA

Three packages are required for data scraping: pandas, scrape, and itertools. We selected a specific location and set a radius for the tweets extraction. When filtering by location, you have two choices: utilise the near city tag in conjunction with within radius, or geocode the latitude, longitude, and radius. When scraping tweets by location, however, it's best to include the geocode tag containing longitude and latitude data, as well as a radius to limit the search area. Given the data and characteristics available, this will produce the most accurate results possible.

For the classification of tweets, a set of words is created that can be confidently classified as belonging to a particular category for each of the selected classes(i.e., black Friday sale, Cyber Monday sale, Costco, Walmart and BestBuy). The clusters chosen are shoes, camera, laptop, kitchen, and watches. All the popular models from each cluster word are taken on creating this. This set of words can be used to form a cluster of people who have tweeted about it.

IV. DATA ACCESS , EXPLORATION AND PREPROCESSING

The stored CSV data is loaded in the panda's data frames and all the frames are combined into one panda's data frame. Also, the number of columns is matched and other data pre-processing operations such as dropping the unnecessary columns, and changing the data types are performed on the data frame. The flexibility provided by the panda's data frame allows the manipulation of the data in an easy and effective manner.

The data fetched using the snscrape API contains HTML links that are not required for the analysis. The data also contains many pictures, videos, hash characters, and retweeted data which indicates redundant data and requires to be fixed as it can create inconsistencies in the analysis. The python regex library is used to remove the characters from the textual data. The columns containing only the HTML links data are dropped from the data frame as well as other columns such as the number of likes that are non-essential for the analysis. The tweets data also contains numbers that are to be removed and passed into the regex expression in the exclusion parameter along with other characters.

V. DATA ANALYSIS AND PREDICTION

The Jaccard similarity measures the similarity between two sets of information to determine which members are shared and distinct. The Jaccard similarity is computed by dividing the number of observations by the number of observations in either set. In other words, the Jaccard similarity will be computed because the size of the intersection is divided by the scale of the union of two sets. this could be written in set notation using intersection ($A \cap B$) and unions ($A \cup B$) of two sets:

$$J(A,B)=(|A \cap B|)/(|A \cup B|)$$

where $|A \cap B|$ is the number of members shared between both given sets and $|A \cup B|$ gives the total number of members in both sets (shared and un-shared). Also, the Jaccard Similarity is 0 if the 2 sets don't share any values and 1 if the 2 sets are identical and the set may contain either numerical values or strings. Also, this function can be used to find the dissimilarity between two sets by calculating $d(A,B)=1–J(A,B)$.

The Jaccard Similarity is frequently used in data science. few examples like the text mining method consist of finding similarities between two text documents based on the number of terms used in both. Similarly, Movie recommendation algorithms use the Jaccard coefficient to find similar customers if they rented or rated highly many of the same movies. Also, In e-commerce by using a database of thousands of customers and millions of products, identify similar customers.

The calculation of the Jaccard similarity score is computationally expensive. The first run on 62000 tweets took 14 minutes to run the function that calculated the Jaccard score. The expense of computer resources and time spend to calculate the Jaccard score increased with an increment in the number of tweets passed into the function. To resolve such an issue, Spark is essentially useful. In this approach, the Optimus-spark library is used to fulfill the requirement. The Optimus-spark library syntax is such that a person without a programmer background can also comprehend the usage of the library and functions associated with ease.

The registration of the Jaccard similarity function is important to be utilized with the Optimus-spark library, the udf (user-defined function) function of pyspark library is used for this purpose. After the use of Optimus-spark, the calculation was performed in 25 seconds on 142000 tweets. The optimization of computational expense performed by the Optimus-spark library works well even on scaling the data

i.e., increasing the number of tweets passed into the function. Apart from providing the solution to the high computational requirement, the Optimus-spark library provides several other functionalities such as performing data cleaning i.e., removing unnecessary characters, keywords, and hashtags, the implementation of machine learning algorithms can also be achieved with the library.

The K-Means algorithm is an unsupervised machine learning algorithm that assembles data into k clusters. A user-defined number of clusters can be chosen and the algorithm will then group the data even if this number is not optimal for the specific case. The common method used to evaluate the appropriate number of clusters is the Elbow method. This method runs k-means clustering for a range of clusters k and each value is the calculation of the sum of squared distances from each point to its assigned center(distortions). When the distortions are plotted and appear to be an arm, the "elbow" (the point of inflection) is the most accurate value for k.

VI. CONCLUSION

The number of clusters produced using k-means clustering analysis is 2, based on the selected 5 keywords or product categories i.e., shoes, camera, laptop, kitchen, and watches, as the data obtained is not evenly distributed throughout all the categories. The clusters are determined by the defined items inside the categories list, therefore the machine learning method may be improved by picking and adding a larger number of products to be analysed in the Twitter data.

VII. ACKNOWLEDGMENT

We would like to express our gratitude to Nasrin Eshraghi Ivari, our supervisor, who proposed the topic and provided valuable lessons and resources. We were able to complete all stages of my project because of her direction and suggestions. We'd also want to express our gratitude to Daniel Khadivi, who assisted us throughout the project. Finally, we'd want to express our gratitude to the University of New Brunswick for giving us this opportunity.

VIII. REFERENCES

[1] A. R, . M. Gawich and M. Fernández-Veiga, "Personalized Recommendation for Online Retail Applications Based on Ontology Evolution.," In Proceedings of the 2020 6th International Conference on Computer and Technology Applications (ICCTA '20)., pp. 12-16, 2020.

[2] C. Badii, P. Bellini, A. Difino and P. Nesi, ""Smart City IoT Platform Respecting GDPR Privacy and Security Aspects"," IEEE Access, 2020.

[3] D. Billsus and M. J. Pazzani , "User Modeling for Adaptive News Access.," User Modeling and User-Adapted Interaction, vol. 10, pp. 147-180, 2000.

[4] Y. Cho , J. Kim and S. Kim , "A personalized recommender system based on web usage mining and decision tree induction Expert Syst. Appl., 23," International Conference on Innovative Internet Community Systems, p. pp. 329–342.

[5] P. Bellini , I. Bruno , D. Cenni , A. Fuzier and P. Nesi , ""Mobile Medicine: Semantic Computing Management for Health Care Applications on Desktop and Mobile Devices", on Multimedia Tools and Applications .," Springer, vol. 58, no. 1, pp. 41-79, 2012.

[6] L. A. I. P. P. N. &. G. P. Pierfrancesco Bellini, "Multi Clustering Recommendation System for Fashion Retail," Springer, 2022.

[7] K. Arvai, "K-Means Clustering in Python: A Practical Guide," Real Python, [Online]. Available: https://realpython.com/k-means-clustering-python/.