Abstract

Compotronics - Clustering and Learning based Analytical Suggester for Electronic Goods

Data scraping & Tabulation	1 Week
Clustering & Categorization	1.5 Weeks
Dashboard and GUI creation	2 Weeks
Backend - Learning & Optimization	2 Weeks
Testing	1 Weeks

Proposed Approach

This involves scraping of data from existing e-commerce websites and using it in a local machine. The data will be in structured form. This data needs to be clustered based on retailer and customer needs. Clustering data will not only make it easier to predicts the products needed by the customer but it will also allow organization of data in the database and give the administrators a clarity on sales and division of products. **Decisive hierarchical clustering** algorithms will be used for classification of products.

The dashboard will allow a user to select the electronic product he/she wishes to purchase. This may include cellular phones, laptops, cameras and televisions. Based on the product chosen, the customer will be asked to rate his/her specifications according to the importance level it holds for them. The app then suggests the best choice for the particular product among all available ones according to the weight given to the specification along with the customers' importance level chosen.

The specification given by the customer is used to find out the suggested products using **Neighbourhood models**. Neighborhood models are heuristics based models which uses similarity metrics, for eg : cosine similarity, for finding similar users and items. It is based on, very reasonable, heuristic that a person will like the items that are similar to previously liked items. Rating prediction in item based neighborhood models is given by weighted average of ratings on similar items as shown below:

$$\hat{r}_{u,i} = b_{u,i} + \frac{\sum_{j \in N(i,k,u)} s_{i,j}(r_{u,j} - b_{u,j})}{\sum_{j \in N(i,k,u)} s_{i,j}}$$

where, N(i, k, u) is a set of k items that are similar to i and rated by the user u; $s_{i,j}$ is a similarity function.

The retailer dashboard and admin dashboard will contain the results and graphs of sales and a prediction of future sales. This prediction is done via an Auto Regression (AR) Model. Since the prediction for the daily sale for each store from past data seems to be a time series problem, we manage to solve this problem by building a time series analysis model. The basic method we are using is auto-regression (AR) model. A package called "Forecast" in R, which does autoregression is being used for this purpose.

Alternative Approaches

The application processes data based on fixed parameters. It suggests to a user, similar products on the basis of other customers' ratings and views. Thus by tracking user's shopping patterns queues are created to prioritize products with respect to each other and finally suggestion is given to the user based on this queue.

Project Scope

The project takes into account the sentimental analysis of a customer according to his/ her requirements. It then generates the desired product output by taking care of other users' views and ratings along with the user's basic criterion for the selection of a particular product. This not only help the customers choose a better product based on what they require but also help retailers order and sell the products in a much more efficient way based on analysis of the past data and prediction of sales in the future. This administrator dashboard also allows the programmers to get an insight on sale patterns and trends based on live sales analysis allowing them to select and eliminate retailers and customers based on their application needs.

Project out of Scope

Using web scraping to extract data from an e-commerce website. Selenium and python scripts will be used for the data scraping. Python scripts will be used to read this scraped text data and create SQL queries out of this data thus allowing proper storage of data in the database.