Master Thesis Proposal - AGR Dynamics

AGR Dynamics is looking for individual that has the passion and drive to solve a real-world problem that every retailer struggles with.

We strongly recommend that the project is programmed in R and SQL2016, (https://blogs.technet.microsoft.com/dataplatforminsider/2016/03/29/in-database-advanced-analytics-with-r-in-sql-server-2016/).

Problem Description

Merchandisers often lack the tools to plan promotions or new product launch and how it is going to effect the sales of other related items ("Cannibalization" or "Halo effect"). By using machine learning models the Sibyl solution can easily group together the most similar items where the merchandizer can plan positive and negative effects of other similar items to reflect the total sales. In a similar manner when merchandizers want to have an overview of highly correlated items when planning a promotion of an item to have the capability of easily increase the estimated sales for the similar items, which is based on the customer behavior and basket analysis. The model can massively improve promotional performance, ensuring that the right products are in the place at the right time to support promotional activity.

Proposed solution

Related items: The idea is to build a content-based recommender model algorithm, which is a Machine Learning model used to find similar items based on the attributes of the new/related item, such as product name, description, price, colour, brand, vendor, etc. The product profile contains both structured and unstructured data where machine learning model can deal with numeric, categories, and free text data. To extract useful features from text data natural language processing could be used. All this information would be used to train the model. The model creates a weighted vector of item features, where the weights denote the importance of each feature and can be computed from individually content vectors using a variety of techniques. Different feature extraction could be used such as TF-IDF, LDA, Neural network, Bayesian classifier.

Basket analysis: Create model that groups together highly correlated items that help planners to plan promotions and the affect on other highly correlated items.

The focus of the Content-based recommender systems is on the characteristics of the items in order to recommend new/related items with similar properties. As an example, a product profile could include a description, price, colour, brand. Usually item profiles include text data. These text data need to be converted into features. In literature, numerous feature extraction techniques have been proposed, including Bag-of-words TF-IDF, LSA, PLSA, and LDA. As the metadata is known in advance, recommendations are also available for new items where no collaborative data has been collected. A downside of the Content-based models is that they can result in restricted content analysis owing to the "singular product" problem, in which the introduction of a product unrelated to all previous products goes unnoticed by the recommender.

Study plan and Deliverables

1. Read the background papers and get familiar with the dataset and R and SQL2016
2. Build content-based recommender engine with real-world dataset in R and SQL2016
3. Do basket analysis and calculate correlation between items, group highly correlated items together with real-world dataset in R and SQL2016.
4. Evaluation of the model and create framework towards production
5. Test model on second real-world dataset from retailer

The framework will be build efficient, user-friendly, modularity and robustness.

The first practical follow-up of the project will be rewriting part of the system to keep standards of a user-friendly, modularity and robustness of the recommender systems.

We created a framework of next steps towards production, where throughout the thesis a strong focus was on making the code and algorithms efficient and robust.