**Status Report**

***Term Project: Predict the relevance of search results on homedepot.com***

*by Weiyi Tuo, Kaiyang Lv, Yichao Chen, Jingyi Lin, Ying Ma*

*March 14, 2016*

1. **Introduction of the Kaggle competition**

Kaggle is a data modeling and data analysis competition platform. Companies and researchers are able to publish their data on the website for gathering solutions on their problems. Statisticians and data mining experts from around the world compete with each other to produce the best model. This crowd sourcing model relies on the fact that there are many strategies that can be used to solve almost any predictive modeling problem, and researchers are incapable of judging what methods are most effective for particular problems at the beginning.

In this project, which is to predict the relevance of search results on HomeDepot.com, is focusing on big data analysis and modeling. What the task mainly about is to analysis the datasets Home Depot provided, in order to help find what extent a search result matches the paired search query. Currently, Home Depot utilizes an implicit measurement to gauge how quickly the customers can search out the products they want. By this manually calculation, the process becomes rather slow and quite subjective. In this project, a relevance evaluation model is expected to take place the human inputs, to greatly improve the efficiency of iterations can perform on the current search algorithms.

1. **Team plan**

We already knew the customers’ search term, and some attributes of each product (title/brand/functionality). What we have to do is calculate (through specific similarity calculating model) a relevance score for each product and search term pair. The given score of real relevance can be used to check the accuracy of the model. Our ultimate goal of this project is to continuously enhance the model to improve the accuracy rate, and finally a relatively precise model will be generated.

According to the tips post on courseweb, our team will solve this problem from the following two aspects:

1. To find the correlation between two items, we could use of an item-to-item collaborative filtering approach. This essentially means that for each item X, we build a neighborhood of related items S(X); whenever customers search certain item in the system, then we could find similar items from that their neighborhood.

The way to find similar items is: since the item-to-item approach makes crucial use of similar items, here is a high-level view of how to do it. First, associate each item with the set of customers who have bought/looked at it. The similarity between any two items could then be a normalized measure of the number of users they have in common (i.e., the Jaccard index) or the cosine distance between the two items (imagine each item as a vector, with a 1 in the *i*th element if user *i* has bought the item and 0 otherwise).

A shortage of the item-to-item approach is that you do not get very much diversity or surprise in item-to-item recommendations, so recommendations tend to be kind of "obvious" and boring.

1. Another way is a factorization approach. Rather than looking at individual items in isolation (in the item-to-item approach, if customer A and B both buy a book X, we may make essentially the same recommendations based on X, regardless of what they bought in the past), factorization approach would look at all the items which customers have bought, and try to detect properties that characterize what they like. For example as a customer has bought a lot of science fiction books and a lot of romance books. In our evaluation system, it regards a book which contains both science fiction and romance elements at the same time are the most relevant products.

Some useful aspects of the factorization approach include that this approach is simpler to implement. Another good factor is the factorization approach can update recommendations faster. As soon as customer searches any new item in the system, Home Depot can make a new recommendation in case the factorization has been recomputed.

1. **Team progress**

We have held a group meeting to discuss about the project objectives. In the meeting, we downloaded the datasets provided by Home Depot, and carefully read the data introductions on the website. After scanning all the datasets, we decided to use R in this project. We imported the csv files into R, and roughly checked the datasets. We found the data are in huge size, which contains plenty of information for us to further analysis. Additionally, there are some missing values in the dataset, which means that before the further analysis, we have to firstly clean the data to make them reliable for all the situations.

In order to make further works more smoothly, we generated a schedule plan for the project. Firstly, we have to do deeper data exploration to find some patterns or overlaps between the provided datasets. After that, we can use the ggplot package used in former assignment, to generate some graphs to show the data frequency or distribution. The data variations are valuable aspects to observe the datasets. Later then, we can try to build a data model based on the current data. The most important part of the project is the verification and testing. In the later phase of the project, we need to do many cross-validation works to ensure the accuracy of the model. It is reasonable that we will need to modify our model many times to produce the most optimal one for Home Depot to use as an evaluation tool.