****

# MASENO UNIVERSITY

# SCHOOL OF COMPUTING AND INFORMATICS

# DEPARTMENT OF COMPUTER SCIENCE

CCS 308: **RESEARCH METHODS AND TECHNICAL WRITING**

Research report: **Analysis of Recommendation Algorithms for Ecommerce**

By

Name: **DENNIS MAYEKU**

Reg: **CI/03002/014**

Date: **29 May 2017**

Sign: **………….**

Lecturer: **Dr. Sylvester McOyowo**

**Analysis of Recommendation Algorithms for Ecommerce (Machine Learning and Data Analytics)**

1. **Abstract**

The aim of this research was to investigate several techniques for analyzing large-scale purchase and preference data for the purpose of producing useful recommendations to customers. The research in particular applies a series of knowledge discovery algorithms such as nearest neighbor collaborative filtering, dimensionality reduction and classical data mining on datasets from two online businesses.

The results indicate that collaborative filtering algorithms do better than the classical data mining techniques and for certain density levels of the neighborhood the difference between the two algorithms is very spectacular. Also, it was observed that when the density of the neighborhood decreases, the quality of recommendations reduce as well.

The research concludes that dimensionality reduction techniques hold the promise of allowing collaborative filtering algorithms to scale to large datasets and at the same time produce high quality recommendations.

It is recommended that in the future works, more research should be carried out to understand why low dimensional representation works well for some recommender applications, and comparatively worse for others.

1. **Introduction**

The largest ecommerce sites offer millions of products for sale. Choosing among so many options is challenging for consumers. Recommender systems have emerged in response to this problem. A recommender system for an ecommerce site receives information from a customer about which products she is interested in (e.g. how they rate and search products), and recommends products that are likely to fit her needs.

One of the earliest and most successful recommender technologies is collaborative filtering. It works by building a database of preferences for products by consumers. A new customer, John, is matched against the database to discover neighbors, which are other consumers who have historically had similar taste to John. Products that the neighbors like are then recommended to John, as he will probably also like them. Therefore, collaborative filtering is based on the assumption that people who agreed in the past will agree in the future, and that they will like similar kinds of items as they liked in the past.

Although very popular, this technique, however, faces two main setbacks:

* **Scalability**

Collaborative filtering algorithms are able to search tens of thousands of potential neighbors in real-time, but the demands of modern ecommerce systems are to search tens of millions of potential neighbors. Also, if a site is using browsing patterns as indications of product preference, it may have thousands of data points for its most valuable customers. This much volume of variable data per customer is likely to slow down the number of neighbors that can be searched per second, further reducing scalability.

* **Quality of recommendations**

A balance needs to be developed between scalability and quality of recommendations. Generally, the less time the algorithm spends searching for neighbors, the more scalable it will be. On the contrary, quality of recommendations decreases with a low density neighborhood.

1. **Problem statement**

The research sought to study new and existing algorithms that have the potential to improve both scalability and quality of recommender systems for ecommerce.

The research presents new algorithms that are particularly suited for sparse datasets, such as those that are common in ecommerce applications. The new algorithms have characteristics that are likely to make them perform better in ecommerce applications than many classical algorithms.

1. **Methods**

In performing the experimental validations, the researchers use two datasets. First dataset is from a large ecommerce site, Fingerhut Inc, which sells a wide variety of heterogeneous products, ranging in price from tens to hundreds of dollars. The second dataset is derived from MovieLens which is a content site with tens of thousands of users who come in to rate and review movies. New visitors on the site get recommendations of movies they are likely to like based on reviews from the other members.

For MovieLens, the researchers randomly selected enough users to obtain 100,000 ratings from the database (only users who had rated 20 or more movies were considered).

The dataset was converted into a binary user-movie matrix, R that had 943 rows (i.e, 943 users) and 1682 columns (i.e 1682 movies that were rated by at least one of the users).

The sparsity level, ML, was also computed as 1 – 100,000 / (943 x 1682) = 0.9369

The train set and test set ratio was set at 80% and 20% respectively.

For FingerHut ecommerce store, the researchers retrieved from the database purchase information of 6,502 customers on 23, 554 catalog products with a total of 97,045 purchase records.

The train set and test set ratio was set at 80% and 20% respectively.

The sparsity level, EC, was also computed to be 0.9994

1. **Results**

* **Neighborhood size**

For both datasets, the size of the neighborhood has a significant impact on the quality of the recommendations.

It was discovered that the quality of the top-N recommendations improves with increase in neighborhood size. However, after a certain point, the improvement gains diminish and the quality becomes worse.

The two datasets have different optimal number of neighborhood count. The MovieLens dataset has a peak neighborhood size of 80 – 120 whereas the FingerHut dataset was discovered to range from 170 to 220.

* **Number of dimensions**

In the case of MovieLens dataset, the recommender quality initially improves as the researchers increase the rank of the lower dimensional space, but it quickly reaches its maximum performance and any further increases in the rank of the space leads to worse recommendation results.

For FingerHut, the recommendation quality continues to improve all the way up to 800 dimensions. This is assumed to be as a result of the much larger dimension i.e, 6502 x 23554 compared to 943 x 1682 for MovieLens. Also at 0.9994 compared to 0.9369 for MovieLens, the FingerHut dataset is significantly sparser.

1. **Interpretation of results/Conclusion**

It was discovered by the researchers that dimensionality reduction techniques hold the promise of allowing collaborative filtering algorithms to scale to large datasets and at the same time produce high quality recommendations. However, future work needs to be carried out to unravel why some datasets in contrast perform worse with low dimensional representations.

The cold start problem is common with collaborative filtering algorithms. The algorithms often need a large amount of existing data on a user in order to make accurate recommendations. This was however not a problem with the two datasets since a ton of necessary information was available beforehand.

The sparsity problem is also another hindrance that was resolved during the experiment. The number of items sold on ecommerce sites is extremely large. The most active users will only have rated a small subset of the overall database. Thus, even the most popular items have very few ratings.

Overall, recommender systems can help reduce cognitive load on users, introduce quality control and help provide solutions to large amounts of quality data.

1. **Weaknesses of the Research**

The researchers only mentioned the likelihood of false positives and false negatives but failed to include their occurrence while conducting the recommendation generation process. A false positive is when the algorithm recommends a product the user would not like whereas a false negative is when the system fails to recommend a product that the user would otherwise like.

Detecting error rate for false positives and negatives can be calculated by masking one of the original ratings then computing it systematically to see how the output predicted by the algorithm compares to the original rating. This is achieved by using the RMSE (Root Mean Square Error) method which is the range between actual prediction and observed prediction, therefore helping the researchers determine occurrence of false positives or false negatives.

1. **Recommendations**

Both large scale and small scale businesses can benefit from increased sales with application of smart algorithms to recommend products to their customers. Recommender systems have been responsible for a significant percentage of revenues in music, advertising, movies, news, books, research articles, search engines, expert systems, financial services, insurance and online dating businesses. Recommender systems are a powerful new technology for extracting additional value for a business from its customer databases. These systems help customers find products they really want to buy from a business. Recommender systems benefit customers by enabling them find products they like. Similarly, they help businesses by generating more sales.

1. **References**

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (n.d.). Analysis of recommendation algorithms for e-commerce. *Analysis of recommendation algorithms for e-commerce*. Retrieved April 6, 2017.