An Accuracy Improvement of Detection of Profile-Injection Attacks in Recommender Systems using Outlier Analysis

Jiten H. Dhimmar
Research Scholar
Computer Engineering
Parul Institute of Technology
Vadodara, India

Raksha Chauhan
Assistant Professor
Computer Science & Eng.
Parul Institute of Technology
Vadodara, India

ABSTRACT

E-Commerce recommender systems are affected by various kinds of profile-injection attacks where several fake user profiles are entered into the system to influence the recommendations made to the users. We have used Partition around Medoid (PAM) and Enhanced Clustering Large Applications Based on Randomized Search (ECLARANS) clustering algorithms of detecting such attacks by using outlier analysis. In user rating dataset, attack-profiles are considered as outliers in these algorithms. Firstly, we have used PAM and ECLARANS clustering algorithm in detecting the attack-profiles. These both algorithms have been applied for evaluating the performance of the system in identifying the attack profiles when they enter into the system. Experiments show that an accuracy of ECLARANS algorithm for detection of profile-injection attack for E-commerce recommender system is more than PAM clustering algorithm.

Keywords

PAM, ECLARANS, Outlier Analysis, Recommender System, profile-injection attack, Attack-profile.

1. INTRODUCTION

A number of website consists of a recommender system to help users by offering a bunch of items that are going to interest them from a large collection of items. In simple terms, recommender systems make it easier to handle the problem of information overload on the web by creating customized recommendations to the users [1].

Content-based and collaborative filtering are two main techniques initiated in creating recommender systems. In content-based recommender systems, items are recommended depending upon target user's ratings and content of the items. Collaborative Filtering creates recommendations for a certain user by regarding feedback of other users. A particular user is compared to another user in the rating databases to identify his neighbors-users with equal tastes. Collaborative Filtering is used by a number of e-commerce site and in the field of information filtering [1]. Collaborative recommender systems are recognized to be greatly affected by profile injection attacks, attacks that contain the insertion of fake profiles into the ratings databases for changing system's recommendation characteristic [2].

People are often needed to make decisions about items without major information of the variety of choices available. Accordingly, we frequently find recommendations from others relating to which movies to watch, which magazines to read or which vehicle to buy etc. Collaborative recommendation algorithms work in the same manner and can be applied to filter information and recommend personalized content that suit the actual requirements and tastes of distinct users [3]. These algorithms have been effectively applied in numerous online settings and work by collecting preference data and feedbacks from users and by using this information to create recommendations for others people.

While people are relatively proficient in determining the reliability of colleagues and associates and valuing recommendations from such type of sources respectively, it is significantly very hard to make decisions regarding users of online environments given their anonymous or pseudo-anonymous character. Since it is practically not possible to identify prior to motivations and integrity of those who frequently use online systems [3], it does not have any guarantee that the choices expressed for items represent the valid viewpoints of users.

It is possible to establish a number of identities within one particular system therefore the potential for profile injection attacks or shilling attacks to occur exists. These attacks contain the generation of several attack profiles which are generally designed to reflect the valid recommendations of genuine users for specific items, while the target item is given a biased rating with the purpose of promoting or demoting recommendations created for the item. These attacks are known as product push and nuke attacks, respectively [3]. Since it is shown that the existence of even small amount of attack profiles can tremendously bias recommendations, it is crucial that online systems are prevented against these types of attack.

In Random Attack a pre-specified rating is given to the target item and random ratings are given to the filler items whereas in average attacks, rating of every filler item represents the mean rating for obtaining that item. Some other attack types are known as Segment Attack, Bandwagon Attack, Reverse Bandwagon Attack and Love/Hate Attack. The last one is quite a simple attack and requires no system knowledge where the attack profile holds smallest or largest rating value for target items and largest or smallest rating value for filler items for nuke or push attack [4].

In literature, the researchers have proposed several outlier detection methods. They usually are categorized into various groups known as depth based approach, density based approach, distance based approach and clustering based approach. In clustering based approach, the clusters having small number of items are taken into account as the clusters that contain outliers assuming that outliers are a small percentage of the total data. The biggest advantage of this approach over the various other approaches is that the outlier detection is fully unsupervised [4]. Besides that, clustering-based techniques are able to use in an incremental mode [5].

2. RELATED WORK

In paper [1], PAM algorithm is use for division of existing rating data into distinct user groups and cluster updating algorithm is use to dynamically upgrade the clusters while new rating data comes into the system. These two approaches are uses to improve the sparsity and scalability of collaborative filtering.

In paper [6], discriminate between genuine and attack profiles and identify profile injection attacks in movie collaborative recommender system. Improving the filler size increases precision and recall and so inserting higher than 250 attack profiles in training sets is unable to increase the precision and recalls extremely.

This paper [7] describes a classification technique to the problem of detecting and dealing with profile injection attacks. In general, several different attributes are determined distinguish properties contained in attack profiles. Three well-known classification algorithms SVM, C4.5 and kNN are normally used in represent the combined advantage of attributes and have an effect on choice of classifier has with regards to increasing the durability of the recommender system.

In this paper [8], they propose a new methodology called Angle-Based Outlier Detection (ABOD) and some variants determining the difference in the angles between the difference vectors of a point to any other points for detecting outliers in a big set of data items. They proposed a novel, parameter-free approach to outlier detection depending upon the difference of angles between pairs of data points. The angles become much greater for points at the border of a cluster. Although angle remains comparatively small by comparison with the angles for real outliers.

Identification of profile-injection attacks on recommender systems are already examined by a number of researchers. Supervised classification techniques have been used in [2] so that you can find out attack profiles from genuine user profiles.

In their paper [9], authors applied hierarchical clustering method in detecting outliers. They have actually compared the overall performance of different hierarchical clustering algorithms using this. The authors of paper [10] have proposed a two-stage method of outlier detection by making use of minimum spanning tree combined with clustering. In paper [11], object of the small clusters obtained by the clustering algorithm are considered as outliers.

In paper [11], authors applied k-means algorithms used for outlier detection. After then in paper [4], Parthasarathi Chakraborty and Sunil Karforma have applied PAM algorithm for the purpose of outlier detection of profile-injection attacks in recommender systems because it is more robust than k-means [4] in presence of outliers. By using PAM algorithm they have enhanced accuracy of profile injection attacks.

3. PROPOSED WORK

Here our proposed work is to improve an accuracy of detection of attack profile in recommender system. For that we are taking ECLARANS algorithm instead of PAM algorithm for the detection of outlier in profile injection attack. We are using ECLARANS algorithm because this algorithm give better accuracy then other clustering based outlier detection algorithm [5] and [12].

3.1 Existing Work

In paper [4] experiment they have used MovieLens dataset (movielens.umn.edu). The data set used included 1,00,000 ratings from 943 users and 1682 movies (items), with every user rated a minimum of 20 items. The item sparsity is easily calculated as 0.9369. The ratings in the MovieLens dataset are integers in the range of 1 to 5.

In push attack highest rating is given to the target items [6] and in nuke attack smallest rating is considered for it. The remaining of the items in attack profiles are randomly selected from the total items in database called filler items [6].

When percentage of filler items is 70%, the performance of PAM algorithm in detecting the attack profiles is 100% i.e. almost all the attack-profiles exist in outlier clusters. When percentage of filler items is 60%, 68% attack-profiles exist in outlier clusters. In case of attack-profiles with 40% percent of filler items, 16% attack-profiles detected correctly.

3.2 Flow of Existing Work

Flow of Existing work for the detection of attack profiles:

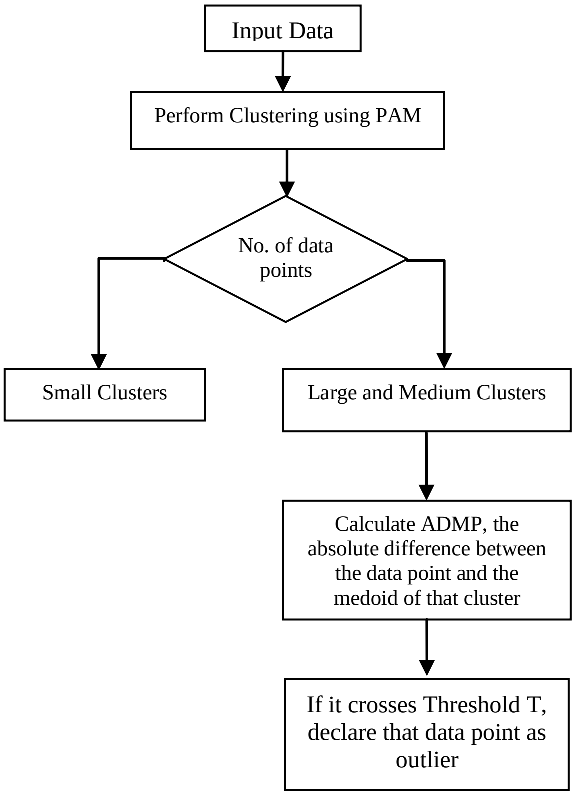
Input Data
Perform Clustering using PAM
No. of data points
Small Clusters
Large and Medium Clusters
Calculate ADMP, the absolute difference between the data point and the medoid of that cluster
If it crosses Threshold T, declare that data point as outlier

Fig 1: System Architecture (Existing Work) [12]

Existing PAM algorithm Steps for outlier detection

(i) Arbitrarily select k objects as medoid points out of n data points (n>k).
(ii) Repeat
(iii) Associate each remaining data object in the given data set to most similar medoid.
(iv) Randomly select a non-medoid object, Orandom.
(v) Compute the total cost S of swapping medoid object Oj with Orandom.
(vi) If S < 0 then swap Oj with Orandom to form the new set of k-medoid objects
(vii) Until no change.

3.3 Flow of Proposed Work

Flow of Proposed work for the detection of attack profiles:

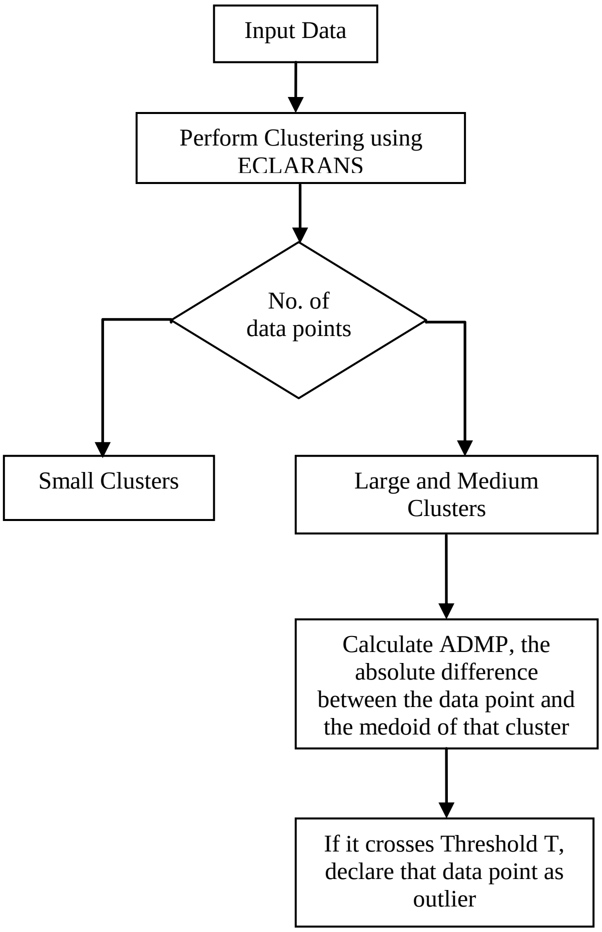
Input Data
Perform Clustering using ECLARANS
No. of data points
Small Clusters
Large and Medium Clusters
Calculate ADMP, the absolute difference between the data point and the medoid of that cluster
If it crosses Threshold T, declare that data point as outlier

Fig 2: System Architecture (Proposed Work)

Now, proposed ECLARANS algorithm Steps for outlier detection:

(i) Input parameters numlocal and maxneighbour. Initialize i to 1 and mincost to a large number.
(ii) Calculating distance between each data points
(iii) Choose n maximum distance data points
(iv) Set current to an arbitrary node in n: k
(v) Set j to 1
(vi) Consider a random neighbor S of current and calculate the cost differential of the two nodes.
(vii) If S has a lower cost, set current to S
(viii) Increment j by 1. If j <= maxneighbour, go to step 6.
(ix) Otherwise, when j > maxneighbour, compare the cost of current with mincost. If the former is less than mincost, set mincost to the cost of current and set best node to current.
(x) Increment i by 1. If i > numlocal, output best node. Otherwise, go to 4.

4. IMPLEMENTATION AND COMPARATIVE ANALYSIS

Here implementation done in 3 Phase

(1) Percentage of filler items is 40

(2) Percentage of filler items is 60

(3) Percentage of filler items is 70

Where k=3 (Number of clusters)

Input: movielens datasets

Output: cluster

Performance Evaluation Factor: Accuracy.

4.1 Percentage of Filler Items is 40

In experiment we have used MovieLens dataset [13]. The data set used contained 25000 ratings from 547 users and 1407 movies (items) where filler items are 40 percentages. The ratings in the MovieLens dataset are integers ranging from 1 to 5. This is a tab separated list of user id | item id | rating | timestamp. The time stamps are unix seconds since 1/1/1970 Coordinated Universal Time (UTC). In this MovieLens dataset filler Items is 10000 and target items is 15000.

4.1.1 Implementation of PAM algorithm for 40% Filler Items

After applying PAM algorithm in this datasets following size of clusters was generated.

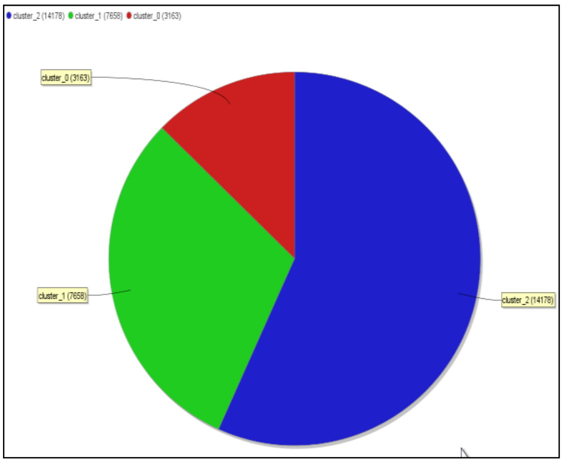

Fig 3: Cluster Size for 40% Filler Items using PAM algorithm

After applying the PAM algorithm on the user rating profiles, we identify those user profiles as attack profiles that belong to the small clusters. Following is the definition of small cluster given in [10] we identify outliers as the data objects that belong to a cluster having size lesser than half the average number of points in the k clusters.

Here cluster 0 and cluster 1 are considered as small clusters because its size is lesser than half the average number of points in the k clusters.

cluster 0 : 3163 items < 8334 items

cluster 1 : 7658 items < 8334 items

cluster 2 : 14178 items > 8334 items
Its means that which user id is included in cluster 0 and 1 are consider as attack profiles. Total attack profiles detected using PAM algorithm for 40% Filler Items is 338 from 547 total users.

4.1.2 Implementation of ECLARANS algorithm for 40% Filler Items

After applying ECLARANS algorithm in this datasets following size of clusters was generated.

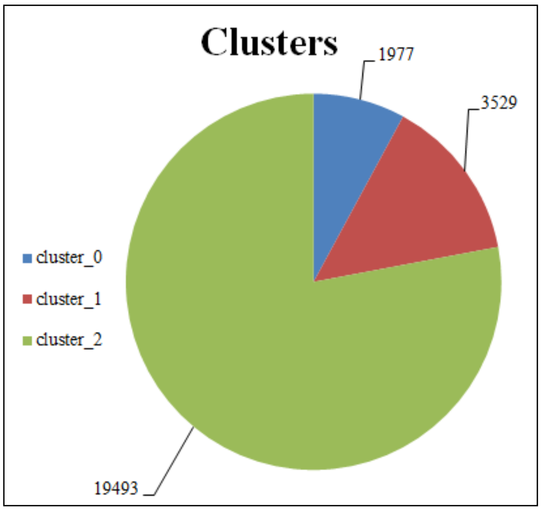

Fig 4: Cluster Size for 40% Filler Items using ECLARANS algorithm

As per the definition of small cluster, here cluster 0 and cluster 1 are considered as small clusters.

cluster 0 : 1977 items < 8334 items
cluster 1 : 3529 items < 8334 items
cluster 2 : 19493 items > 8334 items
Its means that which user id is included in cluster 0 and 1 are consider as attack profiles.

Total attack profiles detected using ECLARANS algorithm for 40% Filler Items is 142 from 547 total users.

4.2 Percentage of Filler Items is 60

In experiment we have used MovieLens dataset. The data set used contained 37500 ratings from 548 users and 1454 movies (items) where filler items are 60 percentages. The ratings in the MovieLens dataset are integers ranging from 1 to 5. In this MovieLens dataset filler Items is 22500 and target items is 15000.

4.2.1 Implementation of PAM algorithm for 60% Filler Items

After applying PAM algorithm in this datasets following size of clusters was generated.

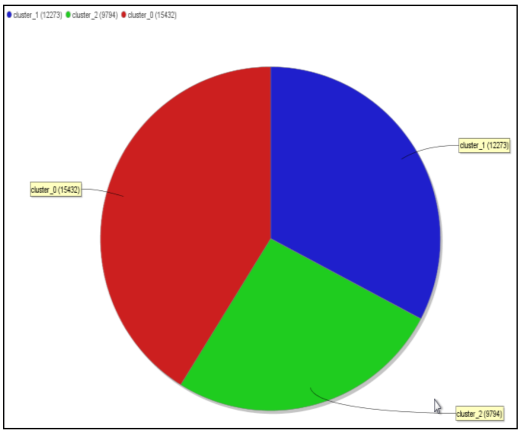

Fig 5: Cluster Size for 60% Filler Items using PAM algorithm

As per the definition of small cluster, here cluster 1 and cluster 2 are considered as small clusters.

cluster 0 : 15432 items > 12500 items
cluster 1 : 12273 items < 12500 items
cluster 2 : 9794 items < 12500 items
Its means that which user id is included in cluster 1 and 2 are consider as attack profiles.

Total attack profiles detected using PAM algorithm for 60% Filler Items is 181 from 548 total users.

4.2.2 Implementation of ECLARANS algorithm for 60% Filler Items

After applying ECLARANS algorithm in this datasets following size of clusters was generated.

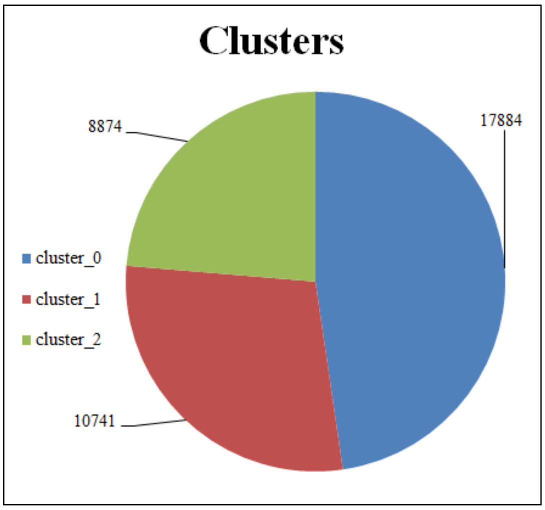

Fig 6: Cluster Size for 60% Filler Items using ECLARANS algorithm

As per the definition of small cluster, here cluster 1 and cluster 2 are considered as small clusters.

cluster 0 : 17884 items > 12500 items
cluster 1 : 10741 items < 12500 items
cluster 2 : 8874 items < 12500 items
Its means that which user id is included in cluster 1 and 2 are consider as attack profiles.

Total attack profiles detected using ECLARANS algorithm for 60% Filler Items is 157 from 548 total users.

4.3 Percentage of Filler Items is 70

In experiment we have used MovieLens dataset. The data set used contained 60000 ratings from 640 users and 1556 movies (items) where filler items are 70 percentages. The ratings in the MovieLens dataset are integers ranging from 1 to 5. In this MovieLens dataset Filler Items is 42000 and target items are 18000.

4.3.1 Implementation of PAM algorithm for 70% Filler Items

After applying PAM algorithm in this datasets following size of clusters was generated.

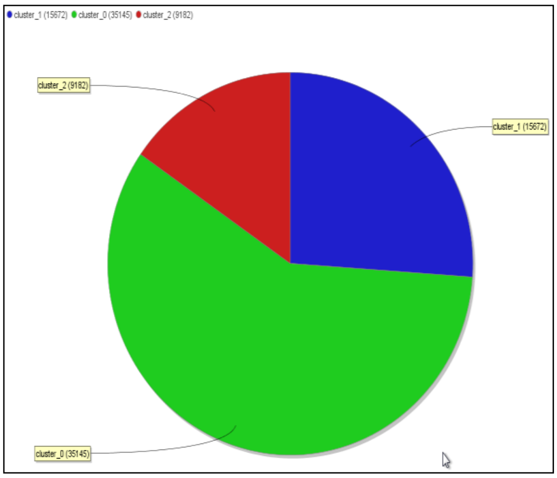

Fig 7: Cluster Size for 70% Filler Items using PAM algorithm

As per the definition of small cluster, here cluster 1 and cluster 2 are considered as small clusters.

cluster 0 : 35145 items > 20000 items

cluster 1 : 15672 items < 20000 items

cluster 2 : 9182 items < 20000 items

Its means that which user id is included in cluster 1 and 2 are consider as attack profiles.

Total attack profiles detected using PAM algorithm for 70% Filler Items is 257 from 640 total users.

4.3.2 Implementation of ECLARANS algorithm for 70% Filler Items

After applying ECLARANS algorithm in this datasets following size clusters was generated.

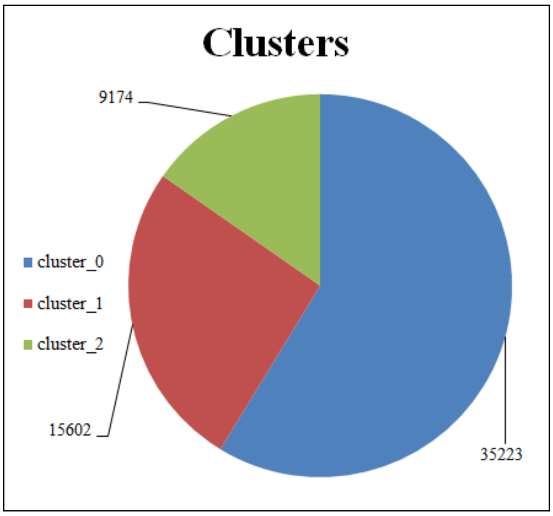

Fig 8: Cluster Size for 70% Filler Items using ECLARANS algorithm

As per the definition of small cluster, here cluster 1 and cluster 2 are considered as small clusters.

cluster 0 : 35223 items > 20000 items

cluster 1 : 15602 items < 20000 items

cluster 2 : 9174 items < 20000 items

Its means that which user id is included in cluster 1 and 2 are consider as attack profiles.

Total attack profiles detected using ECLARANS algorithm for 70% Filler Items is 255 from 640 total users.

Table 1: Comparison of PAM and ECLARANS clustering algorithm for the detection of profile-injection attacks

Percentage of Filler Items	40%	60%	70%
Total Users(Profiles)	547	548	640
Total Attack Profiles	55	123	257
Total Attack Profiles detected using PAM Algorithm	338	181	257
Total Attack Profiles detected using ECLARANS Algorithm	142	157	255
Performance of PAM algorithm to detect attack profiles [4]	16%	68%	100%
Performance of ECLARANS algorithm to detect attack profiles	39%	79%	100%

In existing system PAM clustering algorithm was used for the detection of profile-injection attacks. In above table 1, we can observe that ECLARANS algorithm more correctly detect attack profiles compare with PAM cluster algorithm when size of filler items is less. So that an accuracy of ECLARANS algorithm for detection of profile-injection attacks for recommender system is more than PAM clustering algorithm.

5. CONCLUSION AND FUTURE WORK

In existing system, experiments results show that PAM algorithm only gives accurate results of detecting attack profiles when there is large number of filler items. If we reduce the number of filler items then accuracy will be decrease. We have used Enhanced Clustering Large Applications Based on Randomized Search (ECLARANS) clustering algorithm instead of Partition around Medoid (PAM) clustering algorithm for detecting attack-profiles. Both algorithm have been applied and then examined for comparing the accuracy of the system for distinguishing the attack profiles when they enter into the system. Experiment results of proposed method show that ECLARANS algorithm improves the accuracy of detection of profile-injection attack compare to PAM clustering algorithm for E-commerce recommender system.

We identify that “user profile” as “attack profiles” that belong to the small size of clusters and all the user profiles which included into large size of cluster consider as genuine profile but still there is a presence of attack profiles in large size of clusters. PAM and ECLARANS algorithm can find attack profiles from small number of filler items, but not all attack profiles. In future it’s may be possible to detect all attack profile for small number of filler items as well as large number of filler items by considering detail description of user and movies in movielens datasets.

6. ACKNOWLEDGMENTS

I thanks to Assistant Professor Raksha Chauhan who is from Computer Science and Engineering Department of Parul Institute of Technology, Vadodara. She is also my guide who inspire me for doing such work and due to her support here I am able to write this research paper and she also provide me good knowledge about outlier detection techniques and guide me for doing this work.

7. REFERENCES

[1] Parthasarathi Chakraborty, "A Scalable Collaborative Filtering Based Recommender System using Incremental Clustering", IEEE International Advance Computing Conference, Patiala, India, 6-7 March 2009.

[2] Robin Burke, Bamshad Mobasher, Chad Williams and Runa Bhaumik, "Detecting Profile Injection Attacks in Collaborative Recommender Systems". In Proceedings of the 8th IEEE International Conference on E-Commerce Technology and the 3rd IEEE International Conference on Enterprise Computing, E-Commerce and E-Services, 2006.

[3] Kenneth Bryan, Michael O’Mahony and Padraig Cunningham, “Unsupervised Retrieval of Attack Profiles in Collaborative Recommender Systems”, Technical Report UCD-CSI-2008-03, University College Dublin, April 2008.

[4] Parthasarathi Chakraborty and Sunil Karforma, “Detection of Profile-injection attacks in Recommender Systems using Outlier Analysis”. In International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA), Procedia Technology, Volume 10, 2013.

[5] Deepak Sinwar and Dr. Sudesh Kumar, “Study of Different Clustering Approaches for Outlier Detection”, IJCSC, Volume 4, Number 2 September 2013.

[6] Ghazaleh Aghili, Mehdi Shajari, Shahram Khadivi and Mohammad Amin Morid, “Using Genre Interest of Users to Detect Profile Injection Attacks in Movie Recommender Systems”. In Proceeding of IEEE International Conference on Machine Learning and Applications, 2011.

[7] Chad A. Williams, Bamshad Mobasher and Robin Burke, “Defending recommender systems: detection of profile injection attacks”, SOCA, 2007.

[8] Hans-Peter Kriegel, Matthias Schubert and Arthur Zimek, “Angle-Based Outlier Detection in high-dimensional Data”. In Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 444–452, Las Vegas, NY, USA, 2008.

[9] Antonio Loureiro, Luis Torgo and Carlos Soares, "Outlier Detection Using Clustering Methods: a data cleaning application". In Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector. Bonn, Germany, 2004.

[10] John Peter.S., "An Efficient Algorithm for Local Outlier Detection Using Minimum Spanning Tree", International Journal of Research and Reviews in Computer Science, March 2011.

[11] Al-Zoubi and Moh'd Belal, “An Effective Clustering-Based Approach for Outlier Detection”, European Journal of Scientific Research, Vol. 28, Issue 2, March 2009, pp. 310-316.

[12] S.Vijayarani and S.Nithya, “An Efficient Clustering Algorithm for Outlier Detection”, International Journal of Computer Applications (0975 – 8887), Volume 32– No.7, October 2011.

[13] “MovieLens dataset” available at: https://movielens.org/