International Journal of Scientific Research & Engineering Technology (IJSET)
Vol.13 pp.16-21

Grouping recommender system users in distinct technology diffusion classifications for rating predictions

Tranos Zuva [#1], Keneilwe Zuva [*2]

[#] *Department of ICT*
Vaal University of Technology
Vanderbijlpark, South Africa
[1]tranosz@vut.ac.za

[*]*Department of Computer Science*
University of Botswana
Gaborone, Botswana
[2]zuvak@ub.ac.bw

***Abstract*—Internet has become a core pillar of the modern information society, the number of visitors to the Internet (about 58 percent of the global population) and the rapid growth in the amount of accessible digital information have created a potential threat of information overload. Therefore, timely access to items of interest on the Internet is hindered. Recommender systems are capable of mapping available content/items to user's interest, choices or detect behaviour in order to handle this issue of information overload encountered by Internet users. The system needs to have adequate information about the user to recommend relevant items to the user. This paper endeavours to group users with approximately the same risk levels of being recommended irrelevant content/items. The theory of diffusion of innovation was used in grouping users into categories. The MovieLens dataset was used in the experiment. The risk of calculating the correct rating of a user in each category was obtained then we tested whether the risk values came from different samples using F-test. The results indicated that the users in this dataset may be grouped in only three categories of the Technology Diffusion Model. We concluded that our method of knowing the users of a recommender system will improve the recommendation of relevant content/items to the users. We recommend researchers to then decide the recommender system algorithm/s to use for different categories of users being in rating prediction, ranking and/or recommendation stages.**

***Keywords*—recommender systems, ranking, rating prediction.**

I. INTRODUCTION

The number of visitors to the Internet and the rapid growth in the amount of accessible digital information have created a potential threat of information overload. These interferes with timely access to items of interest on the Internet. Recommender systems were introduced that are capable of mapping available content to user's interest, choices or detected behaviour in order to handle this issue of information overload encountered by Internet users. Based on the user's profile, recommender system has the ability to predict whether a particular user would prefer an item or not. Recommender systems are valuable to both service providers and users(Isinkaye, Folajimi et al. 2015). They lower transaction tariffs for locating and choosing items in an online setting. They have also proven to enhance decision making process and quality (Xu and Chen 2017). In e-commerce environment, recommender systems boost earnings, for the fact that they are efficient technique of selling more products(Schafer, Konstan et al. 2001, Lee and Hosanagar 2014). In scientific libraries, recommender systems assist users by allowing them to manouver beyond catalog searches (Isinkaye, Folajimi et al. 2015). Therefore, the demand to use productive and correct recommendation approaches within a system that will provide suitable and reliable recommendations for users cannot be over-emphasized.

Recommender systems firstly solve the rating prediction problem or matrix completion problem where the system predict the rating value that a user would give for a given item at the same time completing the incomplete rating user-item matrix(Li, Singh et al. 2016, Wang, Guo et al. 2018). This problem has been researched widely and very impressive results have been obtained. Secondly the recommender system should rank the items that need to be recommended to the user (Ricci, Rokach et al. 2015). This ranking recommender system problem has been researched on and the solution tested on users in general. In most cases hybrid algorithms have been opted for in order to solve the problems encountered in using a single algorithm.

In our view different ranking algorithms should be mapped to different group of users. In recommender system we propose that ranking algorithms should take into consideration the characteristics of different groups of users of recommender systems. In this research we endeavour to group the recommender systems users in different groups using the technology diffusion theory and then calculate the risk level of each group.

The arrangement of this paper is as follows: the recommender system prediction algorithms are discussed in section 2, the ranking algorithms in section 3, proposed use of ranking algorithms in a recommender system in section 4, results in section 5 and finally conclusion.

II. BACKGROUND

The purpose of recommender systems is to furnish users with personalized items, which are commonly ranked in a descending order of predicted importance. Three fundamental steps are used by these systems to make recommendations: preferences acquisition (user's input data used to acquire preferences), recommendation computation

ISSN 1737-9296

(proper methods used to compute recommendations) and recommendation presentation (the user presented with the recommendation) (Zuva, Ojo et al. 2012, Jannach and Adomavicius 2016). Existing recommendation systems can be classified into three fundamental categories based on various techniques used in recommendation computation, that is, Collaborative Filtering (CF), Content-Based Filtering (CBF) and Hybrid Filtering (HF)(Zuva, Ojo et al. 2012).

Characteristic information about items (keywords, categorifies, etc) and users (preferences, profiles, etc) are used in content based algorithms. Recommendations are made using a user's item and profile features. The assumption is that if a user was interested in an item in the past, they will once again be interested in a similar item in future. Historical interactions or explicitly asking users about their interests are used to construct user profiles. There are other systems, which utilize user personal and social data in recommending but are not considered purely content-based. Excessive specialization is one issue that has emerged that is making obvious recommendations(Jannach and Adomavicius 2016).

User-item interactions are the basis for collaborative based algorithms. These systems assume that if one item is liked by two users, then if the second user likes a second item, that very same second item could also be an interested to the first user. Hence, the objective is to use historical interactions in order to predict new ones. The similarity of users or items help to predict the next item the user might want, this makes same items to be recommended (Zuva, Ojo et al. 2012).

The main idea in using these algorithms is to predict what the users would want in their lives. The recommender system recommends an item that has a high prediction rating and if the user buys the item it is important that the user then gives feedback in order to evaluate the performance of the system. In experimenting with different algorithms there is need to have a dataset of user-ratings of items of similar nature such movies.

Cross validation model is one way of evaluating recommender systems' performance where a dataset is split into training and test data. There are many evaluation methods that are classified as cross validation models such as holdout method, K-fold cross validation and Leave-one-out validation (Ignatov, Poelmans et al. 2012).

When a recommender system algorithm has been decided, it is experimented with on a defined dataset and then implemented. A hybrid algorithm is used to make obsolete or reduce the magnitude of shortcomings and problems that might exist in other recommendation algorithm assuming that it was used individually to perform some recommendation tasks (Bela Gipp and Hentschel 2009). Users are different, therefore they should be in groups that would allow a specific algorithm to be used. Hybrid algorithms should be in place to deal with differences that exist in recommender system users(Ignatov, Poelmans et al. 2012).

The following section deals with how the recommender systems user may be partitioned into groups using theory of diffusion.

III. Classification of Users

Using theory of diffusion, the users can be put into the following categories;

A. Innovators (2.5%)

The first users to adopt an item or service are the innovators. They are willing to take risks. Their risk tolerance can lead them to adopt items or service which may sooner or later fail(Singh 2013, Aizstrautaa, Gintersa et al. 2015).

B. Early Adopters (13.5%)

The second fastest category of users who adopt an item or service are early adopters. Among the other adopter categories, these users have the highest degree of opinion leadership(Singh 2013, Aizstrautaa, Gintersa et al. 2015).

C. Early Majority (34%)

After a varying degree of time, early majority category users adopt an item or service. The innovators and early adopters time of adoption is significantly shorter than for the early majority. The adoption process of early majority tend to be slower, their social status and contact with early adopters is above average, and rarely hold positions of opinion leadership in a system (Singh 2013, Aizstrautaa, Gintersa et al. 2015).

D. Late Majority (34%)

Individuals in the late majority category will adopt an innovation after the average member of the society. These individuals approach an innovation after the majority of society has adopted it and they have a high degree of skepticism. Late majority are commonly suspicious about an innovation, their social status is below average, their financial lucidity and contact with others in late majority and early majority is very little as well as very little opinion leadership(Singh 2013, Aizstrautaa, Gintersa et al. 2015).

E. Laggards (16%)

The last adopters of an innovation are individuals in this category. They show little to no opinion leadership, this is in contrast to some of the previous categories. These individuals typically have reluctance to change-agents and gravitate towards the advanced in age. Laggards typically tend to be focused on "traditions", their social status and financial fluidity is inclined to be lowest, they turn to be the oldest as compared to all the other adopters and have contact with only family and close friends and they have very little to no opinion leadership as well(Singh 2013, Aizstrautaa, Gintersa et al. 2015).

The following diagram summarizes the categories of recommender systems users.

ISSN 1737-9296

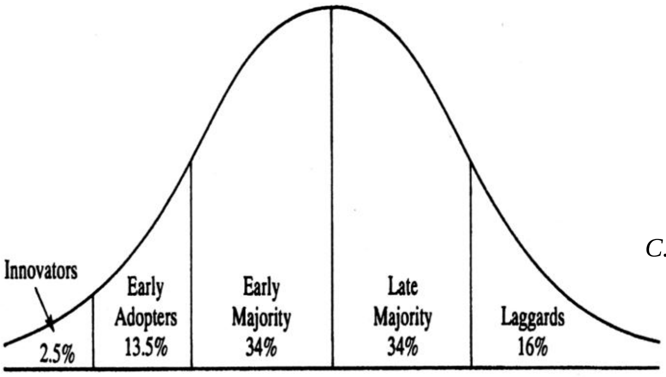

Fig. 1. Categorization of recommender systems users according to diffusion of innovation model.

The following section deals with the methodology used to partition a given recommender system dataset of user-item matrix into technology diffusion categories

IV. METHODOLOGY

In our endeavour to calculate the risk of calculating the rating prediction of users to an item/service we partitioned the recommender system users into technology diffusion categories. Dataset used is from [MovieLens] (http://movielens.org). It contains 20000263 ratings and 465564 tag applications across 27278 movies. "The data set may be used for any research purposes under the following conditions.

- The user may not state or imply any endorsement from the University of Minnesota or the GroupLens Research Group.
- The user must acknowledge the use of the data set in publications resulting from the use of the data set
- The user may not redistribute the data without separate permission.
- The user may not use this information for any commercial or revenue-bearing purposes without first obtaining permission from a faculty member of the GroupLens Research Project at the University of Minnesota".

The data set has the following heading shown in the table below:

userId	movieId	rating	timestamp

The method used to classify the recommender system users is through the timestamp column and to calculate the risk of calculating the predicated user ratings of an item/service. The following algorithm was used:

A. *Sort the dataset using timestamp*

B. *Calculate the average time that it takes for most of the users to watch a movie.*

$$\bar{X} = \frac{1}{n} * \sum_{i=1}^{n} x_i$$

where:

$\bar{X}$ = average (or arithmetic mean)
n = the number of ratings in each category
x_i = the value of each user rating in the category

C. *Calculate standard deviation it takes for most of the users to watch a movie.*

$$sd = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2}$$

where:

n = the number of ratings in each category

$\bar{x}$ = average (or arithmetic mean)

x_i =the value of each user rating in the category

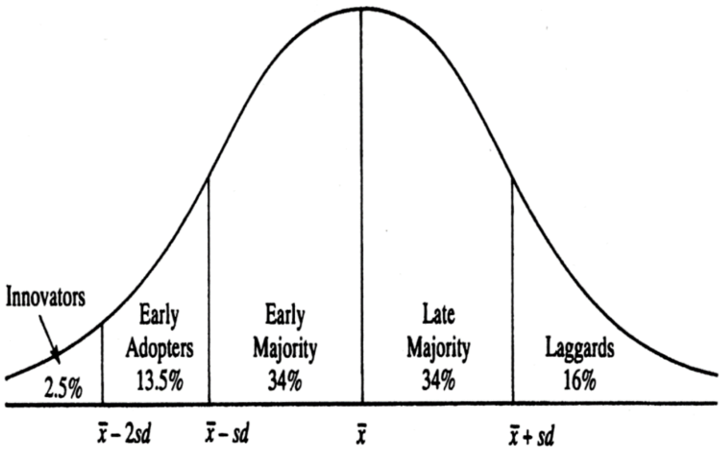

Fig. 2. Technology diffusion categories.

D. *Divide the list of users using the average time calculated in B and standard deviation in C into diffusion categories as in Fig. 2.*

In Fig. 2 the horizontal axis constitutes the average time which is $\bar{x}$ and sd is the standard deviation. A sample of the movies was taken using purposive sampling method. The movies that were watched by almost all the users were used in this experiment

E. *Calculate the variance of their ratings in each category of diffusion*

The following hypothesis were formulated:

ISSN 1737-9296

International Journal of Scientific Research & Engineering Technology (IJSET)
Vol.13 pp.16-21

The null hypothesis (H0) is that there is no difference between the groups and equality between variance.

The alternative hypothesis (H1) is that there is a difference between the variance and groups.

The assumptions of F-test taken into consideration are:

- Normality – That each sample is taken from a normally distributed population
- Sample independence – that each sample has been drawn independently of the other samples

V. RESULT

The risk values for each category of technology diffusion are shown below:

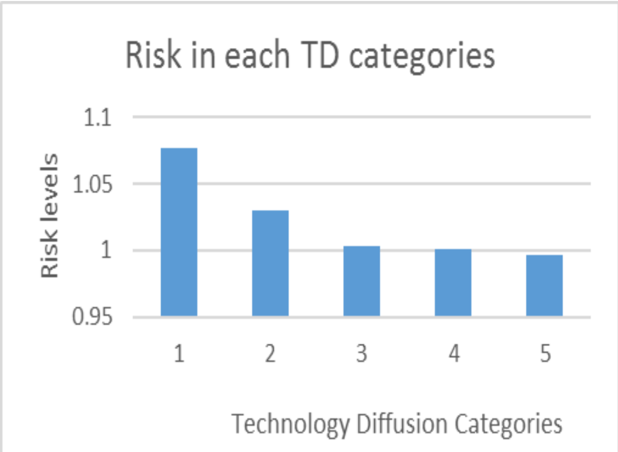

Fig. 3. Results of in each category of technology diffusion categories.

Fig. 3 has the vertical axis indicating the risk values calculated using the standard deviation and the horizontal axis indicating the technology diffusion categories. Bar 1 indicates the risk in calculating rating predication of innovators that is those users that are venturesome, educated and/or that use multiple information sources. Bar 2 indicates risk for the early adopters who are the social leaders, popular and/or educated users. Bar 3 indicates early majority who are thoughtful and/or have many informal social contacts. Bar 4 the late majority those that are sceptical, traditional and/or of lower socio-economic status while Bar 5 indicates the laggards whose neighbours and friends are main information sources and/or at the same time be afraid of debt.

In testing whether the variances of the samples of the users' ratings have a significant difference we used the F-Test. The hypothesis is that the variances are equal. In the event that if F > F Critical one-tail, we reject the null hypothesis.

TABLE I. F-TEST TWO-SAMPLE FOR VARIANCES

	Bar 1	*Bar 2*
Mean	3.699164345	3.592132505
Variance	1.076842875	1.029829319
Observations	359	1932
df	358	1931
F	1.045651795	
P(F<=f) one-tail	0.284392571	
F Critical one-tail	1.139313305*	

This is the case, F < F critical (1.045651795 < 1.139313305*). Therefore, we accept the null hypothesis. The variances of the two populations are equal.

TABLE II. F-TEST TWO-SAMPLE FOR VARIANCES

	Bar 3	*Bar 4*
Mean	*3.521783806*	*3.415330867*
Variance	*1.002711391*	*1.0013219246*
Observations	*4866*	*4866*
df	*4865*	*4865*
F	*0.885636874*	
P(F<=f) one-tail	*1.14804E-05*	
F Critical one-tail	*0.953926004**	

This is the case, F < F critical (0.885636874 < 0.953926004*). Therefore, we accept the null hypothesis. The variances of the two populations are equal.

TABLE III. F-TEST TWO-SAMPLE FOR VARIANCES

	Bar 2	*Bar 3*
Mean	3.521784	3.592132505
Variance	1.002711	1.029829319
Observations	4866	1932
df	4865	1931
F	0.973668	
P(F<=f) one-tail	0.239374	
F Critical one-tail	0.9398*	

This is the case, F > F critical (0.973668> 0.9398*. Therefore, we reject the null hypothesis. The variances of the two populations are unequal.

ISSN 1737-9296

TABLE IV. F-TEST TWO-SAMPLE FOR VARIANCES

	Bar 4	*Bar 5*
Mean	3.415331	3.531209
Variance	1.00132192	0.996515
Observations	4866	2291
df	4865	2290
F	1.136152	
P(F<=f) one-tail	0.000209	
F Critical one-tail	1.061118*	

This is the case, F > F critical (1.136152 > 1.061118*). Therefore, we reject the null hypothesis. The variances of the two populations are unequal.

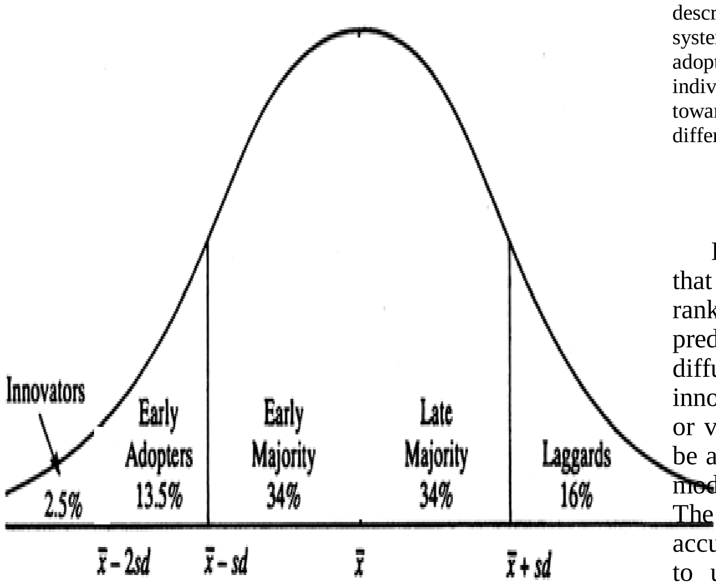

Fig. 4. According to the results recommender systems users may be categorized into these three categories

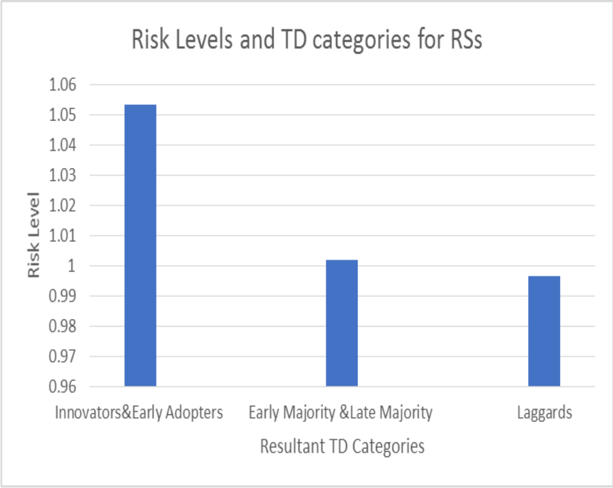

Fig. 5. Risk level in calculating prediction user ratings in technology diffusion categories for recommender systems.

The results from the experiments indicate that the Innovators and Early Adopters can be grouped together. This group of users have the highest degree of opinion leadership.and is willing to take risks. Their risk tolerance can lead them to adopt items or service which may sooner or later fail Early majority and late majority can also be grouped in one category. The adoption process of early majority and late majority can now be described as users who rarely hold positions of opinion leadership in a system and may have a high degree of skepticism. Finally the laggards, last adopters of an innovation. They show little to no opinion leadership.These individuals typically have reluctance to change-agents and gravitate towards the advanced in age. The recommender system should then use different algorithms for different categories of users.

VI. CONCLUSION

In order to personalize the recommender system list so that high accuracy can be achieved on every user, different ranking algorithm and/or different recommender system prediction algorithms should be used in different technology diffusion categories. The results of the study indicate that innovators and early adopters their ratings are so wide apart or vary so much that there is need for suitable algorithm to be able to predict with higher degree of accuracy. There is a moderate variation of the ratings in early and late majority. The laggards can be predicted with some higher degree of accuracy. We recommend that before recommending items to users there is need to classify them then select an appropriate algorithm for each category of users.

REFERENCES

[1] Aizstrautaa, D., et al. (2015). "Applying Theory of Diffusion of Innovations to Evaluate Technology Acceptance and Sustainability." *Procedia Computer Science* 43(2015): 69 – 77.

[2] Bela Gipp, J. B. and C. Hentschel (2009). Scienstein: A Research Paper Recommender System. *Proceedings of the International Conference on Emerging Trends in Computing* (ICETiC'09), Virudhunagar , India.

[3] Ignatov, D. I., et al. (2012). A New Cross-Validation Technique to Evaluate Quality of Recommender Systems. *Perception and Machine Intelligence- First Indo-Japan Conference*, Kolkata, India.

[4] Isinkaye, F. O., et al. (2015). "Recommendation systems: Principles, methods and evaluation." *Egyptian Informatics Journal* 16(3): 261-273.

[5] Jannach, D. and G. Adomavicius (2016). Recommendations with a Purpose. *Proceedings of the 10th ACM Conference on Recommender Systems*, Boston, Massachusetts, USA, ACM.

[6] Lee, D. and K. Hosanagar (2014). Impact of Recommender Systems on Sales Volume and Diversity. *Thirty Fifth International Conference on Information Systems*, Auckland.

[7] Li, B., et al. (2016). Data Poisoning Attacks on Factorization-Based Collaborative Filtering. *29th Conference on Neural Information Processing Systems* (NIPS 2016), Barcelona, Spain.

[8] Ricci, F., et al. (2015). Recommender Systems Handbook. New York, Springer Science+Business Media.

[9] Schafer, J. B., et al. (2001). "E-Commerce Recommendation Applications." *Data Mining and Knowledge Discovery* 5(1): 115-153.

[10] Singh, V. (2013). Innovation diffusion categories and innovation-related needs. *Proceedings of the International Conference on Engineering Design.*

[11] Wang, Z., et al. (2018). Matrix Completionwith Preference Rankingfor Top-N Recommendation. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*(IJCAI-18), Stockholm, Sweden.

[12] Xu, B. and J. Chen (2017). "Consumer Purchase Decision- Making Process Based on the Traditional Clothing Shopping Form." *J Fashion Technol Textile Eng* 5(3): 1-12.

[13] Zuva, T., et al. (2012). "A Survey of Recommender Systems Techniques, Challenges and Evaluation Metrics." *International Journal of Emerging Technology and Advanced Engineering* 2(11): 382-386.

ISSN 1737-9296