Models for Recommender Systems in Web Usage Mining Based on User Ratings

Gopinath Ganapathy and K.Arunesh, *Members, IAENG*

Abstract **- In web applications, recommender systems apply statistical and knowledge discovery techniques to predict and make recommendations to the users. Automatic predictions on the interests of the users are made by the collection of ratings and other information from many other users. Collaborative filtering recommender systems make such predictions. Most of the recommendation techniques are based on navigation behaviors and ratings that might be implicit or explicit. Traditional collaborative filtering techniques are quite vulnerable to injection attacks as many provide noisy ratings that can be detrimental to the quality of predictions and also in sensitivity and sparsity problems. To alleviate these issues this paper presents two unique recommendation models namely RANK-RECO and TEST-RECO using ranking and testing measures respectively. These models evolve into algorithms that were experimented and results were provided.**

Index Terms**—Recommender System, Usage Mining, Injection attacks, Auto predictions, usage behavior**

I. Introduction

WEB users are interested in recommendation technology choices to meet a variety of special needs and tastes. User satisfaction and loyalty enhance more retailers in e-Commerce sites such as Amazon, Google, Netflix and last.fm which have embedded Recommendation Systems (RS) into their applications. Web users Recommender Engines and technology are the emerging technology today. However, it is very hard to recommend to the users as their tastes and desires are transient and fractal. Many approaches have been suggested for RS [1] as it assumes research interest. Recommendations are generally content-based or usage based. The Collaborative filtering approach is usage based. Personalization approach supports recommendations. Usually the Recommender Engines such as Pandora, Stands, and Aggregate Knowledge use either deep structured analysis of in item or social behavior analysis or behavioral analysis around the item before offering recommendations. In this paper, after analyzing various approaches of recommender systems and different recommender engines, two new models have been suggested and implemented for RS.

Manuscript received December 31, 2011; revised March 17, 2011.

Gopinath Ganapathy, Head, Department of Computer Science Bharathidasan University, Tiruchirappalli, India. Mobile: 91-9842407008; fax: 0431-2331662; e-mail: gganapathy@gmail.com.

K.Arunesh, Department of Computer Science, Sri.S.R.N.M College, Sattur, Tamilnadu, India. Mobile: 91-9443380679, IAENG member, ID No 108985, e-mail: arunesh_naga@yahoo.com.

II. Recommender Approaches

The approach to the recommender system is based on either individual's past behavior, which is personalized recommendation, or on the past behavior of similar users, which is social recommendation or on the items of interest, which is item recommendation. The combination of the three approaches can also be used for predictions. It is illustrated in Table 1.

Amazon uses all the three approaches that are based on individual behavior, item and the behavior of other users for its recommendations. For the predictions, a few commercial sites focus on some specific methods for RS. For instance Pandora.com deeply analyses the items. It's a site for music and the technology adopted in the system recommends songs to the users based on the structural data. Basically the technology is based on a deep structural analysis of music files. The subtle musical patterns are detected and the groups are formed based on the patterns. So, the recommendations are made based on the structure of songs.

Table 1: Approaches to Recommendation Systems

Usage behavior based Recommendations	
1.	Actual Items, pages (Personalized recommendations)
2.	Related items (item recommendations)
3.	Similar users tastes (social recommendations)

A strand, another recommendation Engine, recommends the products based on social behaviors of the users. The users' online experience is personalized and it generates suggestions based on the social feedbacks of the users. The techniques used in Strands take into consideration the user's tastes and suggest them to get things they want. This social recommender engine is able to provide real-time recommendations of products and services through computers, mobile phones and other inter-connected devices. It deploys its technology through different products, finance, and social media, mostly like music and video, and business that helps to discover things.

ISBN: 978-988-18210-6-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

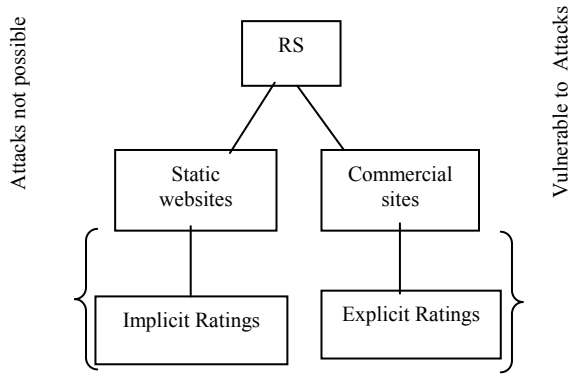

Fig 1. Rating for RS

Aggregate Knowledge is another engine that focuses on analyzing the context of the users visit i.e., the traffic source, semantics, landing page, visitors' demographics etc., and the behavior of the visitors i.e., page views, clicks, time spent on pages etc.,. From this information multiple algorithms based on Behavioral Patterns and Contextual Patterns in the servers extract the best recommendations for the clients' website.

RS are classified into two categories based on web users navigation behavior [2], [3] and web users Ratings [4], [5], [6]. Navigation behavior is the visiting of pages in a particular session of a user. Web users' ratings are based on the ratings given by the users to the products in e-Commerce sites, Fig. 1. and Fig. 2. show the behavior, ratings and security. Implicit, explicit and both the ratings for web pages, products and users are possible in RS [6]. Generally, RS provide information about the items which are recommended by the system.

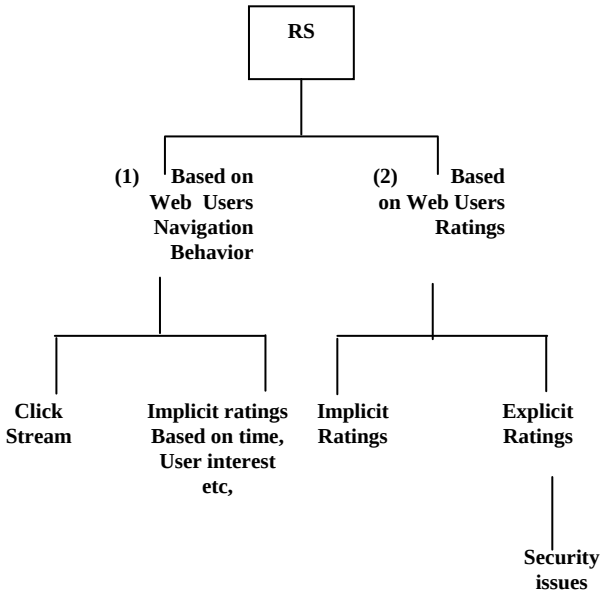

Fig. 2. RS Categories.

The descriptions of the item, other user's reviews, critics, average ratings and predicted personalized ratings for the given user are included when the system recommends. RS provide a way for users to rate the items. Fig. 3. shows the Netflixs' interface and Fig. 4. and Fig. 5. show Amazons' interface. Intelligent and item recommendations need explicit ratings because there are too many product attributes. Attributes such as price, color, style, brand etc., assume different level of importance at different times for the same customer. Users browsing behavior is also changed based on their intention. A classic example is that, a user searches Amazon for new books a day. The same user may search for another product in the following day. If the user explicitly rates the pages and the products, it will support good recommendations.

Fig. 3. Netflixs' suggestions based on users navigation behavior.

Hence, when the system has enough rating data, likes, dislikes, views or valuing, the algorithm would predict the users' taste and would recommends products or services.

III. Related Work

With the significant development in web mining engineering domain, many advanced sophisticated feature techniques, such as Probabilistic Latent Semantic Analysis (PLSA) [7], [8], [9], [10], association rule mining [11], Robust Collaborative Filtering [9], [12], [13], K-Nearest Neighbor algorithms [10], [13], K-Means clustering [10] and matrix factorization [9], [14] are recently utilized to address web usage mining by the researchers. Researcher's argue that, it will improve the quality of web applications, such as web personalization and recommendation systems [1], [15], [16], [17], [18].

ISBN: 978-988-18210-6-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Fig. 4. Amazons' recommendations for books.

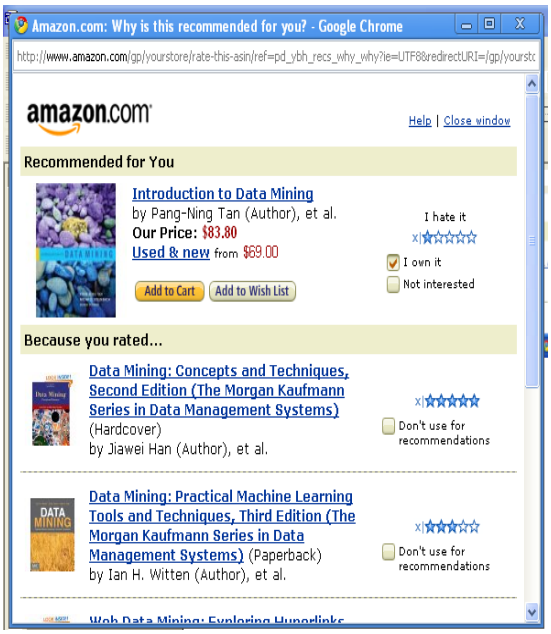

Fig. 5. Amazons' interface for recommendations

In web usage mining, web based prediction system commonly used content based filtering and collaborative filtering system [19], [20], [21], [22]. Content based filtering system generally generates recommendation based on the pre-defined or pre constructed user profiles by comparing the similarity of web content to these profiles. Collaborative filtering system makes recommendations by using the rating of current user for items or products or web pages via referring other user's preference that is closely similar to the current user. As of today, CF system has been widely adopted in web recommendation applications [23], [24], [25] and the researches show that they are user–based CF. If the rating for the products is given by the users explicitly, the CF algorithms are vulnerable to the insertion of biased data [9], [10], [26]. So the researches concentrate on trustworthy and Robust CF Recommender systems [27], [28]. Researchers focus and propose web usage mining as an alternative method for web recommendations because it extracts the knowledge based on the web users behavior that is Navigation behavior and explicit ratings given by the users.

PLSA is an efficient approach to capture the latent or hidden semantic relationships and knowledge among the co-occurrence activities, the system characterize the web user segments and provide dynamic and personalized recommendations [8]. The Markov model and click-stream tree concept which are combined in [17] and a hybrid recommendation model is designed for web users and it recommends web pages. Differentially private RS was proposed in [15] and the model guarantees the privacy of the web users.

A. *Proposed Approach*

In this work, it is proposed to analyze a web recommendations frame work based on rank correlation and pair-t-test. The web user recommendations based on CF is exploited by the attackers and CF is vulnerable to attacks.

IV. Suggested Models

This work introduces two models called Rank measure Recommendations (RANK-RECO) and Testing measure Recommendations (TEST-RECO). In the first method the intelligent system is constructed based on the web users' ratings. The ratings are converted into ranks. The numbers of ratings for the items by the users' are same, the similar tasters are identified in this method. In the second method the sparsity problem is considered, ie., large number of items are presented in e-Commerce applications, but user has select and rated very few items. This leads to sparse entries in the rating vector. In such cases the significant user tastes are predicted based on paired t-test for difference of means.

A. *Recommender Model RANK-RECO (Ranking Measure)*

Let (x_i, y_j); i, j = 1, 2,…, n be the rank of the i th rating in the rating vector given by the users for the products. Ratings given by the users for the products are considered as x and y, takes the values 1, 2… N.

Hence $\bar{X} = \bar{Y} = \frac{1}{n}(1+2+3.....n) = \frac{n+1}{2}$

$$\sigma x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{n^2-1}{12}$$

and $\sigma x^2 = \frac{n^2-1}{12} = \sigma y^2$

In general, $x_i \neq y_i$, Let $d_i = x_i - y_i$,

There fore, $d_i = (x_i - \bar{x}) - (y_i - \bar{y})$

Let us take R as the rank correlation Co-efficient between A and B.

ISBN: 978-988-18210-6-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$$\text{So, } R = 1 - \frac{\sum_{i-1}^{n} d_i^2}{2\pi\sigma x^2} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \quad \text{----- 1}$$

which is the Spearman's formula for the rank correlation coefficient.

Ratings given by the users' are between one and five stars. So the rating vector will contain the numeric values from 1 to 5. Repetition of the same rating is possible for the products. The rating may be tied, in such situation, Let m of the users, say $(K+1)^{th}$, $(K+2)^{th}$,……., $(K+m)^{th}$ are tied, then each of these m users ratings is assigned a common rank, which is the arithmetic mean of the rank K+1, K+2, ……K+m.

The effect of tying m individuals, suppose that there are S such sets of ranks to be tied in the x-series, the total sum of squares is

$$\frac{1}{12}\sum_{i=1}^{n}(m_i^3 - m_i) = \text{Tx}$$

And for y-series of data

$$\text{Ty} = \frac{1}{12}\sum_{i=1}^{n}(m_j^3 - m_j)$$

For this case of ties the Rank Correlations is given by

$$R = 1 - \frac{6(\sum d^2 + Tx + Ty)}{n(n^2-1)}$$ and the limits for rank correlation coefficient are given by $-1 \le R \le 1$.

Algorithm for Ranking Measure

Input : An active users and New user's rating scale for the different products.
Output : Recommendation based on rank Correlation

Step 1 : The active ratings and the pattern are to be treated 2 dimensioned vectors.
Step 2 : Rank the ratings series X and Y. If the same rank is repeated, ie., tied rank, then calculate the average common rank.
Step 3 : Compute Σd and Σd2, d = x – y. Count The number of times the common rank repeated.
Step 4 : For each common rank repeated for x series and y series, Compute TX and TY. Measure the similarities between the active ratings
TX = TY = $m(m^2-1) / 12$.
Step 5 : Calculate the rank correlation coefficient
$Rs(P_i) = 1 - 6 (\Sigma d^2 + TX + TY) / n(n^2-1)/12$.
Step 6 : Repeat the above steps for all users.

Step 7 : Arrange the calculated recommendation scores based on step 4 in a descending order.
ie., RS = $(r_1, r_2, \ldots, r_n)$ and select the highest N recommendation scores to predict the top–N recommendation sets.
REC (S) = { P_j (RS (P_i) > Rs (P_{i+1})}

B. Recommender Model TEST-RECO (Testing measure)

Let x_i be the rating vector given by the existing user and y_j be the rating given by the new user. (i = 1, 2, …, n).

Consider, when the users rating scale sizes are equal, ie., $n_1 = n_2 = n$ and the two ratings scales are not independent but the ratings are paired together, ie., the pair of observations (x_i, y_i), corresponds to the same i^{th}, rating unit. The problem is to test if the rating scale means differ significantly or not.

To test the significance of similar tastes for the users based on rating scale under null hypothesis, t is computed and t is an unbiased estimate of the common population variance which follows students' t-distribution with degree of freedom (n1+ n2-2).

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

where $\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i \quad \bar{y} = \frac{1}{n_2}\sum_{i=1}^{n_{21}} y_i$

$$s^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum (x - \bar{x})^2 + \sum (y - \bar{y})^2\right]$$

Compare the calculated value of t with the tabulated value at certain level of significance. If calculated |t| > tabulated t, null hypothesis is rejected and if calculated |t| < tabulated t, null hypothesis is accepted at the level of significance adopted.

Algorithm for Testing Measure

Input : Rating scale for different products ie., $n_1 \neq n_2$ in X and Y set.
Output : Recommendations based on pair-t-test

Step 1: The product rating scale is treated as 2 dimensioned vectors. $X = \{r_1, r_2.r_{3,,}..,n_1\}$ and $Y = (r_1, r_2, r_3, \ldots, n_2)$.

Step 2: Compute $\bar{x}$, $\bar{y}$ and s^2

Where $\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i$, $\bar{y} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i$

and $S^2 = \frac{1}{n_1 + n_2 - 2}[\Sigma(x - \bar{y})^2 + \Sigma(y - \bar{y})^2]$

Step 3 : Under the null hypothesis, t is calculated for the ratings, X and Y.

ISBN: 978-988-18210-6-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$$\text{Where } t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Step 4: Compare the calculated value t, with tabulated values
If |t| > tabulated t, then reject s←0.

Step 5: If |t| < tabulated t, then s←1, x and y are in significance.

Step 6: Arrange the calculated recommendation scores in step 2-5 in descending order and select the top-N recommendations.

In the tabulated t values, it may be accepted at 1%, 2%, 5%, 10% and 50% level of significance. Based on significant data, it has been concluded that two ratings differ or do not differ significantly.

V. Experimental Evaluation

In order to evaluate the effectiveness of the proposed methods RANK-RECO and TEST-RECO, preliminary experiment is conducted on real word data set based on rank correlation and paired t-test for difference of means.

Data Set

The data set used is downloaded from Grouplens website, which offers more than one lakh ratings on 1682 movies by 945 users' ratings which are available publically in the data set.

Experiment analysis

One to five is the rating integer values, where '1' is considered as lowest and '5' is the highest value. For Amazon recommendations five stars are considered as 'I love it', one is considered as 'I hate it' and two, three, four as 'I like it', 'it's ok' and 'I don't like it'. Similarly Netflix and other domains have their own rating scales.

Table 2 User based CF approach

	M1	M2	M3	M4	M5	CF
U1	5	3	4	2	2	0.7334
U2	4	2	5	3	3	0.6289
U3	2	2	3	4	4	-0.5977
U4	5	2	3	3	3	0.0546
U5	3	2	5	3	3	0.6001
N U	4	4	5	3	3	

For the evaluation of the method, Table 2 is constructed using a simplified user based CF approach. Table 3 and Table 4 are prepared based on RANK-RECO and TEST-RECO models. Users U1..U5, the products M1..M5 movies and the new user's (NU) ratings are given in the tables. For rank correlation method, it is considered that the rating given by the users are as ranks. Rating scales five as rank one and one as the last rank, ie., five. Similarly the other ranks are considered. It is a different approach for recommender systems.

Table 3 User based RANK-RECO approach

	M1	M2	M3	M4	M5	RANK-ECO
U1	5	3	4	2	2	0.75
U2	4	2	5	3	3	0.467
U3	2	2	3	4	4	-0.6
U4	5	2	3	3	3	0
U5	3	2	5	3	3	0.3
N U	4	4	5	3	3	

CF is purely based on the correlation between the user's ratings. ie., the predictions for the new user is provided based on the similar liking users. In e-Commerce applications the user's choices may leads to sparsity. So TEST-RECO model offered better recommendations.

The main advantage of the second method is that, if the number of new user's rating and number of ratings for the items by the existing user's are need not be equal, the significance of tastes between the new user and the existing users can be found. n_1 is considered as new users' rating and n_2 is existing users' ratings for different items. So in such cases the similar tastes between the users' are predicted with the help of pair-t-test. The predictions for the new user are provided based on the similar liking users. In case of sparse ratings the existing CF method is very difficult to find the correlation between users. To overcome this problem the difference of means are calculated. The significantly rated users are identified and the model offered high quality recommendations.

Table 4 User based TEST-RECO approach

	M1	M2	M3	M4	M5	TEST-RECO
U1	5	3	4	2	2	-0.8661
U2	4	2	5	3	3	-0.6325
U3	2	2	3	4	4	-1.3719
U4	5	2	3	3	3	-0.9734
U5	3	2	5	3	3	-0.9734
N U	4	4	5	3	3	

ISBN: 978-988-18210-6-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

VI. Conclusion

This work proposes Rank and Testing measure methods to generate recommendations for web users. Items and page ratings, reviews of customers for the products and items are the standard practices for web based commerce applications. These practices support buying decisions of web users and consumers through valuable independent information. So, the proposed models support taking consumers or web users' decisions into consideration to select their products or web pages.

The experiments show that these RANK-RECO and TEST-RECO models achieve better prediction accuracy. Also it is found that the rank and test measure feature weights are not highly sensitive to slight changes. This paper presents the preliminary results of a work in progress. The promising current results promote further implementations. The proposed recommender models can be extended to domains like health care, world tourism, education system and social web sites for improving the prediction accuracy and sensitivity.

References

[1] Guy Shani and A.Gunawardana "Evaluating Recommender Systems", Microsoft research, MSR-TR-2009-159 Nov – 2009.

[2] D Sule Gunduz, M.Tamer Ozsu "A web page prediction model on click-stream tree representation of user behavior", 2003 ACM 1-58113-737.

[3] Amit Bose, Kalyan Beemanapalli, Jaideep Srivastava, S.Sahar "Incorporating Concept Hierarchies into usage mining based Recommendations", WEBKDD'06 ACM.

[4] N.Ampazis "Collaborative Filtering via Concept Decomposition on the Netflix Dataset", ECAI 2008, pp. 26-30, 2008.

[5] V.Krishnan, P.Narayanashetty, M.Nathan,R.Davies and J.Konstan "Who Predicts Better?-Results from an Online Study Comparing Humans and an Online recommender System", ACM, RecSys'08, 2008, pp 211-218.

[6] F von Reischach, F.Michahelles and A.Schmidt "The Design of Ubiquitous Recommender Systems", MUM '09, ACM 978-1-60558-846-9 09/11 – 2009.

[7] X.Jin, Y.Zhou and B.Mobasher "Web Usage Mining Based on Probabilistic latent Semantic Analysis', KDD'04, ACM 1-58113 2004.

[8] G.Xu, Y.Zhang and X,Zhou "A Web recommendation Technique Based on Probabilistic Latent Semantic Analysis", WISE2005, LNCS 3806, pp 15-28, - Springer 2005.

[9] B.Mehta, T.Hofmann and W.Nejdl "Robust collaborative Filtering" RecSys'07, ACM 978-1-59593-730-8/07/0010, pp 49-56, 2007.

[10] JJ.Sandvig, B.Mobasher and R.Burke "A Survey of Collaborative Recommendation and the Robustness of Model-Based Algorithms" IEEE CCTC on DE - 2008.

[11] JJ.Sandvig, B.Mobasher and R.Burke "Robustness of Collaborative Recommendation Based On Association Rule Mining" ResSys'07, pp 105-112, ACM - 2007.

[12] B.Mehta, T.Hofmann and P.Fankhauser " Lies and Propaganda: Detecting Spam Users in Collaborative Filtering" , IUI'07, pp 14-21, ACM -2007.

[13] B Van Roy and X.Yan "Manipulation Robustness of Collaborative Filtering" April 1, 2010.

[14] Y.Koren, R. Bell and C.Volinsky "Matrix Factorization techniques for Recommender Systems" IEEE computer society, pp 42-49, Aug-2009.

[15] F.McSherry and L.Mironov "Differtially Private Recommender Systems: Building Privacy into the Netflix Prize Contenders", KDD'09, ACM 978-1-60558-846-9 9/9/06 – 2009.

[16] T.Zhu,R.Greiner,G.Haubl "An Effective complete - web Recommender system", WWW'2003 ISBN 963-311-355-5.

[17] M.Goksedef and S.Gunduz " A Consensus Recommender for Web Users" Springer, ADMA 2007, LNAI 4632,pp 287-299, 2007.

[18] X.Jin, Y.Zhou, B.Mobasher "A Unified Approach to Personalization Based on Probabilistic Latent Semantic Models of Web Usage and Content" AAAI -2004.

[19] Dariusz Krol, M.Szymanski, B.Trawinski "The Recommendation mechanism in an internet information system with time impact coefficient", International journal of C.Sc and Application, 2006 vol.3 issue 3.pp 65-80.

[20] Choochart Haruechaiyasak, M.Shyu,S.Chen " Data mining for building A web- page recommender system", 2005.

[21] Ranieri Baraglia C.Lucchese S.Orlando,M.Serrano, F.Silvestri " A privacy preserving Recommender system ", SAC'06 ACM 2006.

[22] Drachsler, H., Hummel, H. G. K., & Koper, R. "Recommendations for learners are different: applying memory-based recommender system techniques to lifelong learning", in proceedings of SIRTEL workshop on EC-Tel, pp: 17-20, September 2007.

[23] John O'Donovan, Barry Smyth "Trust in Recommender Systems", IUI'05, ACM 2005 1-58113-894, 2005.

[24] Paolo Buono, Maria Francesca Costabile, Stefano Guida, Antonio Piccinno, Giuseppe Tesoro, "Integrating User Data and Collaborative Filtering in a Web Recommendation System" in proceedings of 18th International conference on user modeling, vol.2266, pp: 315-321, 2002.

[25] A.Kumar and P.Thambidurai "Collaborative Web Recommendation Systems – A Survey Approach", GJCST, vol 9, issue 5, pp.30-35, 2010.

[26] M. O'Mahony, N.J.Hurley and G.Silvestre "Recommender Systems: Attack Type and Strategies", AAAI-05, 2005, PP 334-339, 2005.

[27] B.Mehta and T.Hofmann "A Survey of Attack-Resistant Collaborative Filtering Algorithms", IEEE CSTC on Data Engineering - 2008.

[28] B.Mobasher, R.Bruke,R.Bhaumik and C.Williams "Toward Trustworthy Recommender Systems: An Analysis of Attack Models and Algorithm Robustness ", ACM vol 7, No.4, pp 23:1-23:38, 2007.

[29] D.Fleder and K.Hosanagar "Recommender Systems and their Impact on sales Diversity", EC'07, ACM – 2007.

ISBN: 978-988-18210-6-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)