Priyanka Shenoy, Manoj Jain, Abhishek Shetty, Deepali Vora / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 3, Issue 2, March -April 2013, pp.676-679

Web Usage Mining Using Pearson's Correlation Coefficient.

Priyanka Shenoy1, Manoj Jain1, Abhishek Shetty1, Deepali Vora2
1Final Year Student, Information Technology
2Head of Department, Information Technology
Vidyalankar Institute of Technology, Wadala.

Abstract

Recommender systems apply knowledge discovery techniques to the problem of making personalized recommendations for information, products or services during a live interaction. In this paper we analyze three different recommendation generation algorithms. We look into three different techniques for computing similarities for obtaining recommendations from them. On the basis of various parameters we conclude Collaborative filtering using Pearson's Correlation Coefficient provides better quality than the others.

I. INTRODUCTION

Web Usage Mining is a web mining technique which analyses user's behavior on the web and helps in generation of recommendations. Recommender systems are being widely used in many application settings to suggest products, services, and information items to potential consumers. For example, a wide range of companies such as Amazon.com, Netflix.com, Half.com and Procter & Gamble have successfully deployed commercial recommender systems and reported increased Web and catalog sales and improved customer loyalty.

II. RECOMMENDATION ALGORITMS

The main techniques used to generate recommendations are as follows:-

- Collaborative Filtering.
- Content Based Filtering.
- Hybrid Filtering.

The technique used for in depth analysis is this paper is Collaborative Filtering.

III. COLLABORATIVE FILTERING

The basic idea of collaborative based algorithm is that it provides item recommendation on the opinions of likeminded people. In a typical CF scenario there are a list of 'm' users and list of 'n; items. Each user expresses his/her opinion regarding the item list. This opinion can be given directly or indirectly through transactions.
Collaborative filtering is used to find similarity in two forms-

- Prediction- It is a numerical value of the likeliness of occurrence of item. The predicted value is in the same scale as opinion value (eg. 1-5)
- Recommendation- It is a list of items that the user will like the most. This interface is known as Top- N recommendation.

Collaborative, content-based, demographic, utility-based and knowledge-based techniques. Based on users' ratings a similarity between users can be calculated. Predictions are then made by using these similarities of a user to other users.

Collaborative techniques can further be divided into-

• Memory Based Collaborative Filtering.
• Model Based Collaborative Filtering.
• Item Based Collaborative Filtering.

a. Memory Based Collaborative Filtering
Memory Based Algorithms utilize the entire user-item database to generate prediction. These methods employ statistical methods to obtain set of nearest neighbors. Once the neighborhood is formed systems use different algorithms to combine preferences and produce Top-N search.

b. Model Based Collaborative Filtering
Provide item recommendation by first developing a model of user ratings. Probabilistic approach is used for the same. Building of models is performed by various machine learning algorithms like Bayesian Network, Clustering and Rule Based Approaches.

c. Item Based Collaborative Filtering
Unlike User Based approach item based looks into set of items the user has rated and links it to other items. Once the most similar items are found prediction is then computed by taking weighted average of target users rating.

IV. PEARSON'S CORELLATION COEFFICIENT

One of the most often used similarity

metrics in collaborative-based systems is Pearson's correlation coefficients.
Pearson's Algorithm is a type of memory based collaborative filtering algorithm. Pearson's correlation reflects the degree of linear relationship between two variables, i.e. the extent to which the variables are related, and ranges from +1 to -1.
A correlation of +1 means that there is a perfect positive linear relationship between variables or in other words two users have very similar tastes, whereas a negative correlation indicates that the users have dissimilar tastes.
Pearson's correlation coefficients are used to determine the degree of correlation between an active
users a and another user u.

A commonly used formula for Pearson's correlation coefficients is:

$$w_{a,u}=\frac{\sum_{i=1}^{n}(r_{a,i}-\overline{r}_a)\cdot(r_{u,i}-\overline{r}_u)}{\sqrt{\sum_{i=1}^{n}(r_{a,i}-\overline{r}_a)^2}\cdot\sqrt{\sum_{i=1}^{n}(r_{u,i}-\overline{r}_u)^2}}$$

Where:
a: the active user
u: another of the users of the system
n: the number of items that both the active user and ALL recommender users have rated
ra: is the average ratings of the active user ru is the average of another user u's ratings
w(a,u) the degree of correlation between user a and user u

Then the prediction for user a on item i denoted by p(a,i) is calculated as follows:

$$p_{a,i}=\overline{r}_a+\frac{\sum_{u=1}^{m}(r_{u,i}-\overline{r}_u)\cdot w_{a,u}}{\sum_{u=1}^{m}|w_{a,u}|}$$

Where m is the total

number of users.
Example of Algorithm.
Given the user-item matrix of Table 2, what would be the system's recommendation to User D for Item4

	Item1	*Item2*	*Item3*	*Item4*	*Item5*
User *A*	4	4	1	4	3
User *B*	2	1	4	2	5
User *C*	3	1	3	2	1
User *D*	5	4	2	3	3
User *E*	3	?	5	3	2

Table 1: User-item matrix for exercise 3.

	Item1	*Item2*	*Item3*	*Item4*	*Item5*
User *A*	4	4	1	4	3
User *B*	2	1	4	2	5
User *C*	3	1	3	2	1
User *D*	5	4	2		3

Table 2: Example of user-item matrix.

a: the active user (D)
u: another of the users of the system (A, B, C)
n: the number of items that ALL users have rated (items 1, 2, 3 and 5)
ra: the average of the active user's ratings, in other words, D's ratings, which is
(5+4+2+3)/4 = 3.5
rA: the average of user A's ratings which is
(4+4+1+3)/4 = 3
rB: the average of user B's ratings which is
1+4+5)/4 = 3
: the average of user C's ratings which is 1+3+1)/4 = 2

To obtain the average of each user we take into consideration Item1, Item2, Item3 and Item5 as these are the items that both recommender and active users have rated. Thus, n which indicates the number of items that both active and recommender users have rated, is 4.

ra,i is the rating given by D to item i and ru,i is the rating given by the recommender users to item i.
By simply applying Pearson's correlation coefficient formula given above the degree of Correlation between user D and any of the other users can be calculated.

For instance, the degree of correlation between user A and user D is as follows:

$$w_{A,D}=\frac{[(5-3.5)\cdot(4-3)]+[(4-3.5)\cdot(4-3)]+[(2-3.5)\cdot(1-3)]+[(3-3.5)\cdot(3-3)]}{\sqrt{(5-3.5)^2+(4-3.5)^2+(2-3.5)^2+(3-3.5)^2}\cdot\sqrt{(4-3)^2\cdot(4-3)^2\cdot(1-3)^2\cdot(3-3}}$$

$$w_{A,D}=\frac{(1.5\cdot1)+(0.5\cdot1)+(-1.5\cdot-2)+(-0.5\cdot0)}{\sqrt{(1.5)^2+(0.5)^2+(-1.5)^2+(-0.5)^2}\cdot\sqrt{(1)^2\cdot(1)^2\cdot(-2)^2\cdot(0)^2}}=$$

$$w_{A,D}=\frac{(1.5)+(0.5)+(3)+(0)}{\sqrt{2.25+0.25+2.25+0.25}\cdot\sqrt{1+1+4+0}}=$$

$$w_{A,D}=\frac{5}{\sqrt{5}\cdot\sqrt{6}}=0.9$$

In the same way the correlations between the other users and user D are calculated as Follows:

$$w_{B,D}=\frac{-5}{\sqrt{5}\cdot\sqrt{10}}=-0.7 \text{ and } w_{C,D}=\frac{0}{\sqrt{5}\cdot\sqrt{4}}=0$$

The recommendation for user D for item 4 will then be as follows:

$$p_{A,item4}=3.5+\frac{[(4-3)\cdot0.9]+[(2-3)\cdot(-0.7)]+[(2-2)\cdot0]}{0.9+0.7+0}=$$

$$p_{A,item4}=3.5+\frac{(1\cdot0.9)+[(-1)\cdot(-0.7)]+[0\cdot0]}{0.9+0.7+0}=3.5+\frac{1.6}{1.6}=4.5$$

COMPARISION BASED ON SOME PARAMETERS.

Parameter	Pearson's Coefficient	Bayesian Network	Region KNN
Data size	Large	Less	Very Large
Suitable for type of application	Web Recommendation.	Bio-Informatics.	QOS based recommendation.
Accuracy of results	Accurate	Accurate	Accurate
Quality of recommendation	Good	Good.	Very Good
Ease of implementation	Easy	Difficult	Most difficult
Performance issues	If data is sparse proper recommendation cannot be obtained	One can lose information due to reduction models.	Highly sensitive.
Scalability	Scalable	More Scalable	Scalable
Advantages	Simple and easy to implement.	Scalability, intuitive.	Overcomes scarcity and information loss.
Disadvantages	Cannot perform well if data is sparse. Depends on human ratings.	Expensive.	Very Complicated.

V. CONCLUSION

Recommendation Systems are a powerful tool for extracting additional value for business from its user databases. These Systems help us to find the items the user wants to buy or would like to view. It benefits the users by helping them find the item they like. They are increasingly used as a tool for E- commerce on the web.

Our paper lays an emphasis on Collaborative filtering using Pearson's Coefficient which allows scaling large datasets as well as giving accurate recommendations.

VI. REFERENCES

1. RegionKNN: A Scalable Hybrid Collaborative Filtering Algorithm for Personalized Web Service Recommendation By Xi Chen, Xudong Liu, Zicheng Huang, and Hailong Sun School of Computer Science and Engineering Beihang University

2. An Introduction to Feature Extraction Isabelle Guyon and Andr´e Elisseeff
3. ItemBased Collaborative Filtering Recommendation Algorithms Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl.
4. An Improved Personalized Filtering Recommendation Algorithm Sun Weifeng, Sun Mingyang, Liu Xidong and Li Mingchu
5. Amazon.com Recommendations Greg Linden, Brent Smith, and Jeremy York Amazon.com
6. Evaluation of Collaborative Filtering Algorithms Bakkalaureatsarbeit zur Erlangung des akademischen Grades
7. A Survey of Collaborative Filtering Techniques Xiaoyuan Su and Taghi M. Khoshgoftaar