IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 10, Issue 1 (Mar. - Apr. 2013), PP 77-82
www.iosrjournals.org

Comparative Analysis of Collaborative Filtering Technique

Urmila Shinde[1], Rajashree Shedge[2]
Computer Department, RAIT, Nerul, Navi Mubai, India

Abstract: *Today it is almost impossible to retrieve information with a keyword search when the information is spread over several pages. The Semantic Web is an extension of the current web in which information is given well-defined meaning. Web personalization is the one application of semantic web usage mining. In this report we have explored comparison of various collaborative filtering techniques. Those techniques are memory based, model based and hybrid collaborative filtering. Our study shows that the performance of hybrid collaborative filtering technique is better than memory based and model based collaborative filtering technique. We have introduced normalization step, which will improve accuracy of traditional collaborative filtering techniques.*
Keywords: *Collaborative Filtering, Hybrid Collaborative Filtering, Semantic web mining, Web mining, Web personalization,*

I. Introduction

The Web has now become the tool for collaborative work and sharing of information throughout the web. Unfortunately the existing web poses a key challenge is that the huge amount of data available is interpretable by humans only, the machine support is limited. It is highly desired that machines should be able to intervene and help while making a search on internet. The major obstacle in achieving this goal has been the fact that as such data available on internet is unstructured, it is disparate and is spread across the web in variant formats. Today it is almost impossible to retrieve information with a keyword search when the information is spread over several pages.

To solve all such cases, semantic web [1] has been proposed. Semantic Web provides a common framework that allows data to be shared and reused across applications, enterprises and community boundaries. Semantic Web Mining can be divided into semantic Web content mining, Semantic Web structure mining and semantic Web usage mining categories. Collaborative filtering algorithm is nothing but an application of semantic web usage mining. A lot of recommendation techniques have been developed recently; Collaborative Filtering (CF) has been known to be the most promising recommendation technique. The main focus of this paper is on web personalization algorithm that is collaborative filtering algorithm. Web site personalization can be defined as the process of customizing the content and structure of a Web site to the specific and individual needs of each user, taking advantage of the user's navigational behaviour. Web personalized retrieval systems are becoming more interesting, especially when not limited to just searching for information but that also are able to recommend the items that would be more appropriate for the user's needs or preferences. This paper presents the detail description of collaborative filtering techniques.

The rest of the paper is organized as follows. Section II presents Literature Survey. Section III deals with Collaborative Filtering techniques along with challenges and characteristics. Section IV gives Overview and Analysis of collaborative filtering techniques. Section V explains the Proposed Approach. Section VI concludes the paper.

II. Literature Survey

2.1 Web Mining

Web mining is the application of data mining techniques to the content, structure, and usage of Web resources. It is thus the nontrivial process of identifying valid, previously unknown, and potentially useful patterns in the huge amount of Web data. Three areas of Web mining are web content mining, web structure mining and web usage mining as shown in Fig. 1.

2.2 Semantic Web Mining

Semantic Web Mining is a series of semantic analysis of information resources and users' question by advanced intelligence theory and technology, through mining its deep semantics, in order to fully and accurately express knowledge resources and user needs.. Parallel to the Web mining, semantic-based Web mining can be divided into Semantic web content mining, Semantic Web structure mining and Semantic web usage mining categories.

2.2.1Semantic Web Content and Structure Mining: In the web mining based semantic network, the differences between content mining and structure mining are almost vanished, so we refer to them collectively as the

semantic Web content and structure mining. Thus, the traditional data mining can easily be transferred to the Semantic Web content and structure mining. This is achieved through:

- Ontology learning
- Mapping and merging ontologies
- Instance learning
- Semantics created by structure

2.2.2 Semantic Web Usage Mining*:* In the Semantic Web environment, we can give clear semantics to user behaviour. On this basis, we can find the users with the same interest, which provides users with ontology-based Web personalized view to improve the Web usage mining results.

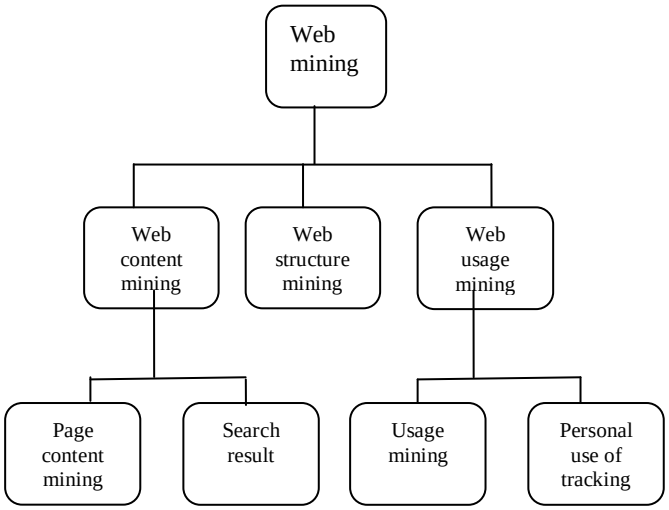

Fig. 1 Web mining techniques

.

2.3 Web Personalization

Web site personalization**[**2][4] is the process of customizing the content and structure of a web site to the specific needs of each user taking advantage of user's navigational behavior. The steps of a Web personalization process include (a) the collection of Web data, (b) the modeling and categorization of these data (pre-processing phase), (c) the analysis of the collected data, and (d) the determination of the actions that should be performed. Fig. 2 represents all these steps.

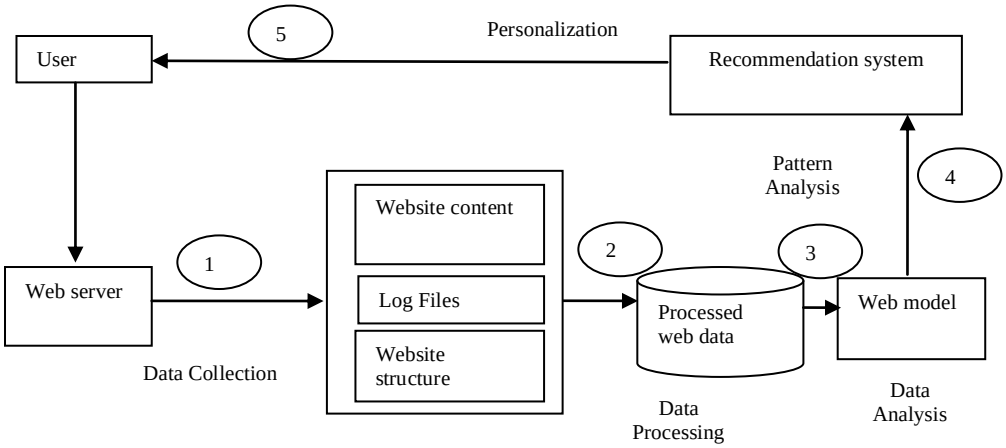

Fig. 2 Web personalization system architecture

The basic personalization techniques are rule Based, Simple Filtering, Content- Based Filtering, Collaborative Filtering.

2.4 Related Work

Here, we present some of the studies for collaborative filtering algorithms. A number of collaborative filtering techniques have been proposed, of which the most popular ones are those based on the correlation criteria and matrix factorization.

Shaina Saini and Latha Banda [3] deals with improving scalability in Collaborative filtering through Genre Interestingness measure approach using Tagging. Due to the explosive growth of data and information on web, there is an urgent need for powerful Web Recommender system (RS). RS employ Collaborative filtering that was initially proposed as a framework for filtering information based on the preferences of users. But CF fails seriously to scale up its computation with the growth of both the number of users and items in the database. Apart from that CF encounters two serious limitations with quality evaluation: the sparsity problem and the cold start problem due to the insufficiency of information about the user. To solve these limitations, they combine many information sources as a set of hybrid sources. These hybrid features are utilized as the basis for formulating a Genre Interestingness Measure (GIM), they propose a unique approach to provide an enhanced recommendation quality from user created tags. This paper is based on the hybrid approach of collaborative filtering, tagging and GIM approach.

Zeeshan Khawar Malik, Colin Fyfe [4] presents Web personalization research has a combination of many other areas that are linked with it and includes AI, Machine Learning, Data Mining and natural language processing. This work describes the whole era of web personalization with a description of all the processes that have made this technique more popular and widespread. This paper has also thrown light on the importance of this strategy and also the benefits and limitations of the methods that are introduced in this strategy. This paper also discusses how this approach has made the internet world more facilitating and easy-to-use for the user.

Boddu Raja Sarath Kumar [5] presents some standard computational techniques are applied within the framework of Content-boosted collaborative filtering with imputational rating data to evaluate and produce CF predictions. The Content-boosted collaborative filtering algorithm uses either naive Bayes or means imputation, depending on the sparsity the original CF rating dataset. Results are presented and shown that this approach performs better than a traditional content-based predictor and collaborative filters.

Here we studied brief about web mining and relation between semantic web mining and web personalization, along with works related to various authors. Next chapter describes the web personalization technique i.e. collaborative filtering.

III. Collaborative Filtering

In everyday life, people rely on recommendations from other people by spoken words, reference letters, and news reports from news media, general surveys, travel guides, and so forth. Recommender systems assist and augment this natural social process to help people sift through available books, articles, web pages , movies, music, restaurants, jokes, grocery products, and so forth to find the most interesting and valuable information for them. The developers of one of the first recommender systems, Tapestry coined the phrase "Collaborative Filtering (CF),"

CF techniques use a database of preferences for items by users to predict additional topics or products a new user might like. In a typical CF scenario, there is a list of m users $\{u_1, u_2. . . u_m\}$ and a list of n items $\{i_1, i_2, . . . , i_n\}$, and each user, u_i, has a list of items, Iu_i, which the user has rated, or about which their preferences have been inferred through their behaviours. The ratings can either be explicit indications, and so forth, on a 1–5 scale, or implicit indications, such as purchases or click-through. For example, following is the list of people and the movies they like or dislike as shown in Table I, in which Tony is the active user that we want to make recommendations for.

TABLE I
EXAMPLE OF USER-ITEM MATRIX

Users	**Item**			
	Shaktiman	**Aryaman**	**Spider-man**	**Super-man**
Alice	Like	Like		Dislike
Bob		Like	Dislike	Like
Chris		Dislike	Like	
Tony	Like		Dislike	?

There are missing values in the matrix where users did not give their preferences for certain items. There are many challenges for collaborative filtering tasks. CF algorithms are required to have the ability to deal with highly sparse data, to scale with the increasing numbers of users and items, to make satisfactory recommendations in a short time period, and to deal with other problems like synonymy (the tendency of the same or similar items to have different names), shilling attacks, data noise, and privacy protection problems. Early generation collaborative filtering systems, such as GroupLen [6], use the user rating data to calculate the similarity or weight between users or items and make predictions or recommendations according to those calculated similarity values..Well-known model-based CF techniques include Bayesian belief nets (BNs) CF models, clustering CF models [8], and latent semantic CF models [7]. An MDP (Markov decision process)-based CF system [9] produces a much higher profit than a system that has not deployed the recommender.

3.1 Characteristics and Challenges of Collaborative Filtering

E-commerce recommendation algorithms often operate in a challenging environment, especially for large online shopping companies like eBay and Amazon. For CF systems, producing high quality predictions or recommendations depends on how well they address the challenges, which are characteristics of CF tasks as well.

1) Data Sparsity: In practice, many commercial recommender systems are used to evaluate very large product sets. The user-item matrix used for collaborative filtering will thus be extremely sparse and the performances of the predictions or recommendations of the CF systems are challenged. The data sparsity challenge appears in several situations, specifically, the cold start problem occurs when a new user or item has just entered the system; it is difficult to find similar ones because there is not enough information.
2) Scalability: When numbers of existing users and items grow tremendously, traditional CF algorithms will suffer serious scalability problems, with computational resources going beyond practical or acceptable levels. For example, with tens of millions of customers (M) and millions of distinct catalog items (N), a CF algorithm with the complexity of O(n) is already too large. As well, many systems need to react immediately to online requirements and make recommendations for all users regardless of their purchases and ratings history, which demands a high scalability of a CF system.
3) Synonymy: Synonymy refers to the tendency of a number of the same or very similar items to have different names or entries. Most recommender systems are unable to discover this latent association and thus treat these products differently. For example, the seemingly different items "children movie" and "children film" are actual the same item, but memory-based CF systems would find no match between them to compute similarity. Indeed, the degree of variability in descriptive term usage is greater than commonly suspected.
4) Gray Sheep: Gray sheep refers to the users whose opinions do not consistently agree or disagree with any group of people and thus do not benefit from collaborative filtering.
5) Shilling Attacks: In cases where anyone can provide recommendations, people may give tons of positive recommendations for their own materials and negative recommendations for their competitors. It is desirable for CF systems to introduce precautions that discourage this kind of phenomenon.

IV. Collaborative Filtering Techniques

4.1 Memory-Based Collaborative Filtering Techniques

Memory-based CF [2], [11] algorithms use the entire or a sample of the user-item database to generate a prediction. Every user is part of a group of people with similar interests. By identifying the so-called neighbours of a new user (or active user), a prediction of preferences on new items for him or her can be produced. The neighbourhood-based CF algorithm, a prevalent memory-based CF algorithm, uses the following steps:

1) Similarity Computation 2)Prediction and Recommendation Computation 3) Top-N recommendation

- Weighted Sum of Others' Ratings: To make a prediction for the active user, a, on a certain item, i, we can take a weighted average of all the ratings on that item according to the following formula:

$$P_{a,i} = \bar{r}_a + \left(\sum_{u \in U} \left(r_{u,i} - \bar{r}_u \right) \cdot w_{a,u} \right) / \left(\sum_{u \in U} | w_{a,u} | \right) \quad (1)$$

Where, r_a and r_u are the average ratings for the user a and user u on all other rated items, and $w_{a,u}$ is the weight between the user a and user u. The summations are over all the users $u \in U$ who have rated the item i.

- Simple Weighted Average: For item-based prediction, we can use the simple weighted average to predict the rating, $P_{u,i}$, for user u on item i

$$P_{u,i} = \left(\sum_{n \in N} r_{u,n} w_{i,n} \right) / \left(\sum_{n \in N} |w_{i,n}| \right) \quad (2)$$

Where, the summations are over all other rated items $n \in N$ for user u, $w_{i,n}$ is the weight between items i and n, $r_{u,n}$ is the rating for user u on item n.

4.2 Model-Based Collaborative Filtering Techniques

Model based collaborative filtering is a two stage process for recommendations in the first stage model is learned offline in the second stage a recommendation is generated for a new user based on the learned model. The design and development of models such as machine learning, data mining algorithms can allow the system to learn to recognize complex patterns based on the training data, and then make intelligent predictions for the collaborative filtering tasks for test data or real-world data, based on the learned models.

4.3 Hybrid Collaborative Filtering Techniques

Hybrid CF systems combine CF [8],[6] with other recommendation techniques (typically with content-based systems) to make predictions or recommendations. Hybrid CF recommenders are combined by adding content-based characteristics to CF models, adding CF characteristics to content-based models, combining CF with content-based or other systems, or combining different CF algorithms.

V. Comparative Analysis

TABLE II
COMPARISON OF COLLABORATIVE FILTERING TECHNIQUES

Parameters		**Algorithms**		
		Memory based	**Model Based**	**Hybrid recommender**
Scalability		Low	High	Very High
Accuracy		Low	High	Very High
Memory consumption		Low	High	Low
Time complexity	Offline	-	O(m)	O(m)
	Online	O(mn)	O(mn)	O(mn)

VI. Proposed Approach

The User-Item dataset is not smooth and the distribution of rating is not uniform; the minority of items has much more rating chances, while the majority of the items have much less rating chances. Due to this traditional collaborative filtering suffers from poor accuracy.

To solve this problem, we first normalize the database and then use that normalized database for further calculations. Here, we add data normalization is the first step and then subsequent steps of traditional collaborative filtering. All these steps are given below. Let us take Table III matrix data.

TABLE III
EXAMPLE OF RATING MATRIX

Users	**Names of the film**			
	Seeta Aur Geeta(SAG)	**Titanic (T)**	**Gadar (G)**	**Ravan (R)**
	I_1	I_2	I_3	I_4
U_1	4	?	5	5
U_2	4	2	1	
U_3	3		2	4
U_4	4	4		
U_5	2	1	3	5

A. **Step 1: Data Normalization**

At first we should normalize the database. The User-Item dataset is not smooth. The normalization is introduced as follows,

$r^{norm}_{u,i}=(r_{u,i}\text{-mean}_i)/ \text{var}_i$

where $r_{u,i}$ is the rating of user u on item i. mean_i denotes the mean of rating values on item i. var_i is the variance of rating values on item i . In this paper $r_{u,i}$ means the real rating user u on item i, and norm $r_{u,i}$ means the normalized rating of user u on item i. Data is normalized using above stated formula as shown in following Table IV.

TABLE IV
NORMALIZED DATABASE

Users	**Normalized Rating**			
	Seeta Aur Geeta(SAG)	**Titanic (T)**	**Gadar (G)**	**Ravan (R)**
	I_1	**I_2**	**I_3**	**I_4**
U_1	0.94	?	1.03	1.52
U_2	0.94	0.21	-0.8	
U_3	-0.63		-0.34	-3.03
U_4	0.94	1.08		
U_5	-2.19	-0.86	0.11	1.52

B. Step 2: Similarity and Prediction calculation

In this step we can use the traditional collaborative filtering similarity and prediction formula to calculate prediction of the user on particular item.

C. Step 3: Top-N neighbour selection

In this step we can analyze the user-item matrix to discover relations between different users or items and select the Top-N neighbors, use them to compute the recommendations.

So due to this normalization step user has more possibilities to get correct product recommendation.

VII. Conclusion

A recommender system is presented as a Web-based application that is known for its usage on e-commerce personalized Web sites, with the purpose of helping customers in the decision making and product selection process by providing a list of recommended items. Recommender systems employ information filtering algorithms to predict items. The most successful algorithm in this field is hybrid Collaborative Filtering. Even though this algorithm is the best, it suffers from poor accuracy and high running time. The proposed personalized recommendation approach have normalization step, which will improve the accuracy of traditional collaborative filtering technique.

REFERENCES

[1] Anil Sharma, Suresh Kumar, Manjeet Singh M.Tech (IS) Scholar, CSE, A.I.T. - G, New Delhi, India "*Semantic Web Mining For Intelligent Web Personalization*" in proceeding of Journal of Global Research in Computer Science Volume 2, No. 6, June 2011.

[2] Suresh Joseph" *A Imputed Neighborhood based Collaborative Filtering System for Web Personalization*" in proceeding of International Journal of Computer Applications (0975 – 8887)Volume 19– No.8, April 2011

[3] Boddu Raja Sarath Kumar1, Barre John Ratnam2 & Maddali Surendra Prasad Babu3 *"Improvement Of Personalized Recommendation Algorithm Based On Hybrid Collaborative Filtering"* International Journal of Computer Science & CommunicationVol. 1, No. 2, July-December 2010.

[4] Zeeshan Khawar Malik, Colin Fyfe "*Review Of Web Personalization"* in proceeding of Journal Of Emerging Technologies in Web Intelligence, Vol. 4, No. 3, August 2012.

[5] Boddu Raja Sarath Kumar *"An Implementation of Content Boosted Collaborative Filtering Algorithm"* in proceeding of International Journal of Engineering Science and Technology Vol. 3 No. 4 Apr 2011

[6] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, *"Grouplens: an open architecture for collaborative filtering of netnews,"* in proceedings of the ACM Conference on Computer Supported Cooperative Work, pp. 175–186, New York, NY, USA, 1994.

[7] T. Hofmann, *"Latent Semantic Models For Collaborative Filtering,"* ACM Transactions on Information Systems, vol. 22, no. 1, pp. 89–115, 2004.

[8] L. H. Ungar and D. P. Foster "*Clustering Methods For Collaborative Filtering,"* in Proceedings of the Workshop on Recommendation Systems, AAAI Press, 1998.

[9] G. Shani, D. Heckerman, and R. I. Brafman, *"An MDP-Based Recommender System,"* n proceeding journal of Machine Learning Research, vol. 6, pp. 1265–1295, 2005.