A New Collaborative Filtering Model based on Robust Graph Coloring for generation of communities

Lizbeth Gallardo-López ⋆, Pedro Lara-Velázquez ⋆⋆, Miguel Angel Gutiérrez-Andrade ⋆⋆⋆, Sergio G. de-los-Cobos-Silva †, and Beatriz González Beltrán ‡

Universidad Autónoma Metropolitana, Mexico

Abstract. Collaborative Filtering is one of the most important models that have been implemented to support the retrieval of relevant information for users of recommendation systems. This paper proposes a Collaborative Filtering model based on the Pearson correlation index and the Robust Graph Coloring model. According to the results obtained from three algorithms (Greedy, and two versions of GRASP) we can affirm that it is possible to implement this model in a collaborative filtering recommendation system that have thousands of users.

Key words: Collaborative Filtering, Data Mining, Metaheuristics, Robust Graph Coloring

1 Introduction.

The amount of information available on the Web grows exponentially every year. In contrast, the retrieval of information relevant for a user becomes more complicated. Nowadays there are two systems to support the user in this task: the information search systems and recommendation systems. Both systems are

⋆ Departamento de Sistemas, Universidad Autónoma Metropolitana Azcapotzalco, Av. San Pablo 180, Col Reynosa Tamaulipas, C. P. 02200, México, D. F., México, glizbeth@correo.azc.uam.mx

⋆⋆ Departamento de Sistemas, Universidad Autónoma Metropolitana Azcapotzalco, Av. San Pablo 180, Col Reynosa Tamaulipas, C. P. 02200, México, D. F., México, pedro_lara@correo.azc.uam.mx

⋆⋆⋆ Departamento de Ingeniería Eléctrica, División de Ciencias Básicas Ingeniería, Universidad Autónoma Metropolitana Iztapalapa, Av. San Rafael, Atlixco No. 186, Col. Vicentina, Del. Iztapalapa, 09340 México, D. F., México, gamma@xanum.uam.mx

† Departamento de Ingeniería Eléctrica, División de Ciencias Básicas Ingeniería, Universidad Autónoma Metropolitana Iztapalapa, Av. San Rafael, Atlixco No. 186, Col. Vicentina, Del. Iztapalapa, 09340 México, D. F., México, cobos@xanum.uam.mx

‡ Departamento de Sistemas, Universidad Autónoma Metropolitana Azcapotzalco, Av. San Pablo 180, Col Reynosa Tamaulipas, C. P. 02200, México, D. F., México, bgonzalez@correo.azc.uam.mx

designed to assist users in finding interesting items (documents, books, music, video, websites, etc.). However, the search systems require users to formulate a query in each search, contrary to the recommendation systems, which provide their users with a continuous flow of items, without having to express explicitly what to look for. To make this possible, the recommendation systems need to model the interests of users (profiles). There are three approaches to define a user profile: 1) Content-Based Filtering model: a set of weighted items, 2) Collaborative Filtering model: evaluating a set of items, 3) Hybrid Filtering model: uniting the two previous proposals.

Our research work seeks to develop a recommendation system, which aims to support users in an University in their teaching and research activities, where working groups share the same interests for information [6]. Specifically, in this article, we propose a new model that of Collaborative Filtering that combines the Pearson correlation index [12] and the Robust Graph Coloring model (RGC) [16].

The Pearson correlation index is used to calculate the degree of similarity between pairs of users and RGC is used to form communities. Commonly, the CGR model is used to optimize the assignment of resources within a group of users. However, we found that it can also be used to identify a set of users that share the same interests, information, etc. and therefore, belong to the same community with different degrees of affinity, for instance: compatible, indifferent and incompatible. Based on these communities the system can make recommendations on compatible items not known by the user, that other members of their community have evaluated positively.

This paper is structured as follows: section 2 will be a brief introduction to the Collaborative Filtering model, in Section 3 the RGC model is defined, Section 4 reviews the related work; in Section 5 a Collaborative Filtering model based on the Pearson correlation index and the RGC model is proposed, in section 6 some preliminary results are shown, and finally, the conclusions are presented.

2 Collaborative Filtering.

A recommender system is intended to assist users in finding items (documents, books, music, video, websites, etc.), that could be interesting for them. To achieve this, a recommender system needs to know the interests of current users (profiles), as well as monitor their evolution over time. Recommender systems employ three types of models of information filtering [1]:

1. *Content-Based Filtering model.* The items are indexed by topic (in the form of keywords). The profile is given by a set of issues weighted by the same user, which characterize their interests. The model proposes to make a correspondence between the indices of the items and the user's profile, in order to suggest the most relevant items.
2. *Collaborative Filtering model.* Involves a community of users through the principle of mutual help: each member of the community receive recommendations of items that other members have deemed interesting. This is

possible because these members previously have provided their opinions (assessments) on the items. The user profile is formed with their own evaluations of items. Thus, the system may recommend items from the correspondence between user profiles.
3. *Hybrid Filtering model.* Consists in the articulation of the Collaborative Filtering and Content-Based filtering. Both approaches come together synergistically, i.e. Hybrid Filtering creates an outcome that takes advantage of both models and maximize their qualities.

While our ultimate goal is to develop a recommender system that incorporates a Hybrid model of information filtering, this article is focused in proposing a new model for Collaborative Filtering. Collaborative Filtering simulates the natural recommendation process between two people who have similar tastes on any item. To achieve this, Collaborative Filtering finds a community of people (users), who from the discovery of an item decides to provide freely their opinion. This opinion is determined by an assessment scale, for instance: 1 - Not important 2 - Somewhat important 3 - moderate 4 - good 5 - very good. The group of items becomes the user profile. The model allows to determine the degree of similarity between pairs of users, from the correspondence of their profiles; and finally, the system gives the recommendation(s) for each user. These can be obtained using algorithms that are commonly classified into two types [3]:

1. *Algorithms based on memory.* These algorithms employ heuristic techniques to make predictions based on assessments of the entire collection of items evaluated by all users. The value of unknown $r_{c,s}$ for user c and item s is calculated based on the affinity of the evaluations of other users on the same item.
2. *Algorithms based on the model.* These algorithms employ the collection of assessments as the basis for a model, which is used to make predictions about assessments. Probabilistic models are used in calculating the probability that the user c provide an assessment particular $r_{c,s}$ for the item s.

A comprehensive state of the art in filtering algorithms can be found in [1]. Specifically, our model of Collaborative Filtering is based on memory and uses the Pearson correlation index as a measure of affinity between users and the RGC model to determine user communities of different sizes and different degrees of affinity, and thus obtain a set of recommended items. The following section describes formally RGC.

3 Robust Graph Coloring.

The problem of resources optimization, where a set of users compete for a limited set of resources, has inspired several models for its solution [5]. For example, the Minimum Coloring model (MC) and the Robust Graph Coloring model (RGC) [16]. In the MC model the assignment of a resource is indicated only as prohibited

or permitted. In contrast, the RGC model indicates degrees of feasibility to share a particular resource.

The RGC model defines two graphs: the original and complementary. The original graph $G = (V, E)$, consists of a set of vertices V, representing users and a set of edges $E = \{i, j\}$, representing the incompatibilities among users i and j. The complementary graph $\bar{G}$ is made up by the same set of vertices V and all edges not included in the set E which is denoted as:

$$\{i, j\} \in \bar{E} \Leftrightarrow \{i, j\} \notin E$$

Therefore, $\bar{G} = (V, \bar{E})$.

The set $\bar{E}$ is penalized by a value of incompatibility $P_{i,j}$ between two users. $P_{i,j}$ is a measure of how undesirable is that two vertices have the same color. While the share of the same color is more undesirable between i and j, the value of the penalty will be greater. A $C(i)$ color (resource) is assigned to each vertex i, and C_k is the set of all colors assigned to the vertices of the model.

The function of rigidity $R(C_k)$, is defined as the sum of penalties of up edges whose end vertices have the same color:

$$R(C_k) = \sum_{\{i,j\} \in \bar{E}, C(i)=C(j)} p_{ij}$$

In this way, the goal of RGC is to get a C_k that is valid in G and has a minimal value of $R(C_k)$ in $\bar{G}$.

One property of the RGC model is that requires an equal or greater value than the minimum number of colors required to get a valid solution in G, since the objective is to optimize the rigidity and not minimize the number of colors used. However, the model requires the specific number of colors to be used in the instance, which represents the number of resources available.

The Collaborative Filtering model proposed in this work must be heuristic, because of the computational complexity of the RGC model [16]. In the next section a review of the related work will be made.

4 Related work.

Collaborative Filtering algorithms based on memory determine the similarity (affinity) between two users in order to build communities and ultimately make recommendations to the items most relevant to each user.

The two most popular approaches for determining similarity between two users are the correlation-based approach [12] and the cosine-based approach [3]. Also, there are studies that have proposed extensions to the techniques based on correlation based on the cosine. In particular, we refer the work of Breese [3], which proposes a method called the default voting (rating). Chen [4] proposes a graphical similarity between items and from them creates an array of affinity between users similar to that generated by the Pearson correlation index.

To solve the Collaborative Filtering models, heuristic solution techniques are used; we can mention Kpodjedo [9], which proposed a taboo object-oriented algorithm, to identify critical classes (in object-oriented paradigm) in which a tester should focus their testing. On the other hand, Al-Shamri [2] proposes a fuzzy model to segment the population. This model uses an algorithm metaheuristics (genetic) which considers a modified Euclidean distance function instead of a correlation matrix.

There are some works where a Collaborative Filtering model is proposed like Wang [14], Huang [8] and Chen [4]. A related model to ours is [13]; this model seeks to define a distance between users, and for each of them finds their closest peers. However this model does not create communities that optimizes the affinity between the members, so this model can not generate recommendations with different levels of affinity for the members. Another related work is [15], where a clustering-based model, similar to standard coloring (where only the original graph is considered) is proposed, the model considers one community for the user, this caracteristic does not allow to measure how related to other communities is the user.

5 Collaborative Filtering Model Based on Robust Graph Coloring.

We propose a collaborative filtering model based on memory to build a recommender system. The system will obtain a prediction of the evaluation $r_{u,i}$, of the user u on item i, based on assessments made by other users. The general idea of the model is to use the Pearson correlation index as a measure of the affinity (or dislike) between two users, then use this index to identify communities, i.e. groups of users with similar tastes, and finally make recommendations for the most relevant items to each user. The model of RGC will find the communities with higher affinity between its members, minimizing incompatibilities (rigidity) of the total users, and forbiding the inclusion of users with antagonistic tastes in the same community. Our premise is that the information in the Pearson matrix correctly reflects the tastes of users. In this way, the Collaborative Filtering model comprises the following steps:

1. We define the coefficient of incompatibility between two users as the complement of the Pearson correlation coefficient:

$$p_{i,j} = \frac{1}{2} - \frac{\sum_{I \in I_u \cap I_w} (r_{u,i} - \bar{r}_u)(r_{w,i} - \bar{r}_w)}{2\sigma_u \sigma_w}$$

   Where:
   - $r_{u,i}$ is the assessment given by the user u to item i.
   - $r_{w,i}$ is the assessment given by the user v to the item i.
   - $\bar{r}_u$ and $\bar{r}_w$ are the average of the assessments issued by users u and w respectively.

- σ_u, σ_w is the standard deviation of the assessments of u and w respectively.
- I_u, I_w is the set of items evaluated by the user u and w respectively.

When this number is very small, indicates that users u and w have a high affinity; when the number is very large, indicates that users u and w have opposite affinity. Other measures of similarity can be used; however, as an instance of aplicability in this work a Pearson correlation index is detailed.

2. In the CR model, each vertex represents a user. Given two users u and w the value $p_{u,w}$ represent the difference in taste between the two users. If the value of $p_{u,w}$ is very close to 0 we had "a match made in heaven", with a value of 0.5 we got "indifference", and for a value close to 1 we got "a match made in hell". For this reason, for every value equal or bigger than 0.5 we made a prohibition that both users can belong to the same community, introducing an edge in E.
   In contrast, all the edges not included in E belong to the complementary graph $\bar{G} = (V, \bar{E})$. The goal is to find a valid coloring C_k whose rigidity $R(C_k)$ is minimal. In this way, the vertices with a small coefficient of incompatibility will have the same color (same community) and the vertices with an incompatibility bigger than 0.5, the vertices will be painted with different colors (different communities).
3. We define γ the average number of users per community related, as follows:

$$\gamma = \frac{v}{k}$$

   Where v is the number of users and k is the number of communities. Since a user can determine values of γ to obtain recommendations from other users with different degrees of affinity. These degrees of affinity, in turn, determine levels of quality in the recommendation.
   If we choose 5 degrees of affinity, for instance $\gamma = 3, 6, 10, 20$ and 30, we obtain the following recommendations qualities: the items from the first 3 users are considered as "highly recommended" while the recommendations from the following 3 user $(6-3)$ will be considered "very good recommendations", the items from the following 4 user $(10-6)$ will be considered "moderately recommended, the following items from the following 10 users $(20-10)$ will be considered "slightly recommended" and the items in the last 10 $(30-20)$ will be considered "poorly recommendable". For 5 degrees of affinity, the algorithm must be run 5 times, with different values of γ. This caracteristic is an improvement of the aproach of [15] which only allows one community for each user.
   One application of this model is a recommendation system for a university community, which is under development. The values of γ may be defined by the system administrator, or may be determined by the user, who will be based on their experience in using the system to fix this. One way to make the recommendations will be updated based on evaluations of users daily, and send recommendations to mail users. At this time, we are focused on experiencing different algorithms to solve instances of the RGC model.

In the medium term, we will deliver a prototype of this recommendation system using as an instance the database of MovieLens. The following section presents the results of existing algorithms to solve the RGC model.

6 Experimental results.

At present, there are good algorithms to solve this in a reasonable time, e.g. Guo [7] between others. We generated some algorithms that solve several instances of the order of a thousand vertices, in reasonably short time. These instances were executed on a Pentium 4 processor at 1.2GHz. The following comparison table, shows the results obtained by executing three algorithms: *Greedy*, *GRASP*, and *GRASP Improved* of instances very similar to those generated in a matrix of incompatibilities between users, just like the ones used in a Collaborative Filtering instance. A detailed description of the algorithms is made in [10] and [11]. The uncertainty values indicate the rigidity for a population of v users with k communities. The figure after the slash indicates the average time of execution of each instance in seconds.

Instance		Type of solution algorithm ($\gamma = 3$)		
v	k	**Greedy** $R(C_k)/secs$	**GRASP-1** $R(C_k)/secs$	**GRASP-2** $R(C_k)/secs$
100	34	$24.4 \pm 1.7/0.02$	$16.8 \pm 1.4/0.10$	$14.5 \pm 0.9/0.12$
200	67	$41.6 \pm 2.0/0.17$	$24.8 \pm 1.5/0.95$	$22.9 \pm 1.3/0.96$
300	100	$60.1 \pm 1.8/0.58$	$32.3 \pm 2.4/4.2$	$28.5 \pm 1.2/3.60$
400	133	$82.3 \pm 2.3/1.1$	$36.8 \pm 1.8/10.5$	$35.2 \pm 1.3/8.40$
500	167	$105.1 \pm 2.6/2.6$	$49.1 \pm 1.5/22.2$	$37.5 \pm 1.8/16.1$
600	200	$128.7 \pm 2.6/4.6$	$45.1 \pm 1.6/38.3$	$43.1 \pm 1.4/28.2$
800	267	$175.4 \pm 3/10.7$	$52.3 \pm 1.9/96.1$	$50.8 \pm 1.6/71.9$
900	300	$199.3 \pm 3.2/15.2$	$54.3 \pm 2.0/143.3$	$53.4 \pm 2.6/102$
1000	333	$223.3 \pm 3.8/22$	$60.7 \pm 2.3/199$	$58.1 \pm 2.2/143$

Table 1. Experimental results

According to these data, for 1000 users, to generate the recommendations will requiere $143 * 5 = 715\ minutes = 12\ minutes$, which is a very reasonable running time for a daily maintenance system.

Conclusions.

This article presented a Collaborative Filtering model that combines the Pearson correlation index and the model Robust Graph Coloring. The Pearson correlation index is a measure that allows to calculate the degree of affinity or dislike between users. The result is expressed in a matrix that feeds the RGC model to form communities.

The model of Collaborative Filtering was formalised and some instances of test were presented, which were solved using three generic algorithms: Greedy, and two versions of GRASP. Based on the experimental results, the algorithms provide good results with solution times in the order of minutes on a Pentium 4 processor at 1.2GHz. A study with the database of MovieLens, as an instance for the model proposed in this work, is being developed.

In the medium term, this model will be implemented in a collaborative filtering recommendation system for a university community. Given the characteristics of the proposed filtering model, an implementation with a daily update for a population of thousands of users is feasable.

References

1. Adomavicius, G. Tuzhiln, A.: Toward the Next Generation of Recommender Systems: A Survey of the State of the Art and Possible Extensions. IEEE Transactions of Knowledge and Data Engineering. Vol 17, No. 6, June 2005.
2. Al-Shamri, M. Y. H., Bharadwaj, K. K.: Fuzzy Genetic Approach to Recommender Systems Based on a Novel hybrid User Model. Expert systems with Applications 35 (2008) (pp. 1386-1399).
3. Breese, J.S., Heckerman, D. and Kadie, C.: Empirical Analysis of Predictive Algorithms for Collaborative Filtering, proceedings 14th Conference Uncertanty in Artificial Intelligence, july 1998.
4. Chen, Y.L., Cheng, L.C.: A novel collaborative filtering approach for recommending ranked items. Expert systems with Applications 34 (2008) (pp. 2396-2405).
5. Diestel, R. Graph Theory. Springer Verlag. New York, 2000.
6. Gallardo-López, L., González-Beltrán, B. Modelo Conceptual de un Sistema de Filtrado de Información para apoyar a una Comunidad Universitaria. Actas del XX Congreso Nacional y VI Congreso Internacional de Informática y Computación. Chihuahua, Chihuahua. Octubre, 2007.
7. Guo, S. Kong, Y. Lim A, and Wang, F. A New Neighborhood Based on Improvement Graph for Robust Graph Coloring Problem. AI 2003: Advances in Artificial Intelligence. Proceedings 16th Australian Conference on AI, Perth, Australia, Springer Berlin, (pp 126-136).
8. Huang, Z., Chen, H., Zeng, D.: Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. ACM Transactions on Information Systems, vol. 22 Number 1 (116-142), 2004.
9. Kpodjedo, S., Ricca, F., Galiner, P. Antoniol G.: Not all clases are created equal: Toward a Reccomendation system for focusing Testing. RSSE08 Atlanta, Georgia, USA. 2008.
10. Lara-Velázquez, P., Un algoritmo Evolutivo para resolver el problema de Coloración Robusta. PhD. Thesis. Universidad Nacional Autónoma de México, Facultad de Ingeniería. México, 2005.
11. Lara-Velázquez, P. Gutiérrez-Andrade, M. A., Ramírez-Rodríguez, J., López-Bracho, R., Un algoritmo Evolutivo para resolver el problema de Coloración Robusta. Revista de Matemática: Teoría y Aplicaciones. Vol. 12, No. 1-2, (pp. 112-119), 2006.
12. Resnick, P. Iacovou, N. Suchak, M. Bergstrom, P. Riedl, J.: GroupLens: An Open Architecture for Collaborative Filtering of Netnews. Proceedings of the 1994 Conference on Computer Supported Collaborative Work. ACM Press, New York (pp. 175-186).

13. Sarwar, B. M. Karypis, G. Konstan, J. A. Riedl, J. T. Application of Dimensionality Reduction in Recommender System – A Case Study WebKDD-2000 Workshop.
14. Wang, J., Vries (de), A.P., Reinders, M.J.T.: Unifying User-based and Item-based Collaborative Filtering Approaches by Similarity Fusion, Proceedings of SIGIR2006, august 6-11 2006 Seattle Washington USA (501-508).
15. Xue, G., Lin C., Yang Q., Xi W., Zeng H., Yu, Y., and Chen Z. Scalable Collaborative Filtering Using Cluster-based Smoothing. SIGIR 2005 August 15-19, Salvador, Brazil. 2005
16. Yañez, J., Ramirez, J.: The robust Coloring Problem. European Journal of Operational Research. Vol. 148, No. 3, 2003, (pp. 546-558)