EVALUATION OF RECOMMENDER SYSTEMS THROUGH SIMULATED USERS

Miquel Montaner, Beatriz López and Josep Lluís de la Rosa
Institut d'Informàtica i Aplicacions - Universitat de Girona
Campus Montilivi, 17071 Girona, Spain
mmontane, blopez, peplluis@eia.udg.es

Key words: recommender systems, evaluation procedure, user simulation, profile discovering

Abstract: Recommender systems have proved really useful in order to handle with the information overload on the Internet. However, it is very difficult to evaluate such a personalised systems since this involves purely subjective assessments. Actually, only very few recommender systems developed over the Internet evaluate and discuss their results scientifically. The contribution of this paper is a methodology for evaluating recommender systems: the "profile discovering procedure". Based on a list of item evaluations previously provided by a real user, this methodology simulates the recommendation process of a recommender system over time. Besides, an extension of this methodology has been designed in order to simulate the collaboration among users. At the end of the simulations, the desired evaluation measures (precision and recall among others) are presented. This methodology and its extensions have been successfully used in the evaluation of different parameters and algorithms of a restaurant recommender system.

1 INTRODUCTION

Recommender systems help users to locate items, information sources and people related to their interest and preferences (Sanguesa et al., 2000). This involves the construction of user models and the ability to anticipate and predict user preferences. Many recommender systems have been developed from several years ago applied to very different domains (Montaner et al., 2003). Unfortunately, only a few of them evaluate and discuss their results scientifically. This situation is caused by the difficulty of acquiring results that can be used to compute evaluation measures. As a consequence, to date, it is very difficult to determine how well recommender systems work, since this involves purely subjective assessments. However, advances on recommender systems require the development of a comparative framework. Our work is in this line.

The contribution of this paper is a new methodology to evaluate recommender systems that we call "profile discovering". The aim of this method is to provide the different steps to design the simulation of the execution of a recommender system with several users over time. The most interesting characteristic of our methodology is that the tastes, interests and preferences of the users are not invented in the simulation process. The "profile discovering procedure" bases the results of the simulations on real information about users. In particular, each user to be simulated has to provide a list of evaluated items. Then, the method simulates the recommendation process of each user and when information about them is required, it is obtained from the lists. Therefore, the simulation can be seen as a progressive discovery of the lists of evaluated items (user profiles). After the simulations, the methodology provides a set of results from the point of view of different measures.

The main properties of this methodology are:

- The simulation process does not invent the user evaluations, they are extracted from real user profiles.
- The recommendation process considers the development of the user profile over time.
- Large-scale experiments are carried out quickly.
- Experiments are repeatable and perfectly controlled.

This paper also proposes an extensions for the "profile discovering procedure" that incorporates the collaboration among users into the recommendation process.

The "profile discovering procedure" and its extension allow us to test the performance of the parameters of the recommendation process in order to tune functions and algorithms. Moreover, it is a suitable instrument to compare different recommender systems, an inconceivable experiment so far.

The outline of this paper is as follows: the next section describes some evaluation techniques for recommender systems that have been used in the current state-of-the-art. Then, our proposal for evaluating recommender systems, namely what we have called "profile discovering" is presented in section 3. Following, its extension for performing multi-agent collaboration is detailed in section 4. Then, some experimental results are shown in section 5 and, finally, section 6 concludes this paper.

2 RELATED WORK

In the current state-of-the-art, recommender systems use one of the following approaches in order to acquire the results for evaluating the performance of their systems: a real environment, an evaluation environment, the logs of the system or a user simulator.

First, results obtained in a real environment with real users is the best way to evaluate a recommender system. Unfortunately, only a few commercial systems like Amazon.com (Amazon, 2003) can show real results based on their economic effect thanks to their information on real users.

Second, evaluation environments are an alternative for some systems to be evaluated in the laboratory by letting a set of users interact with the system over a period of time. Usually, the results are not reliable enough because the users know the system or the purpose of the evaluation. An original approach was accomplished by NewT (Sheth, 1994); in addition to the numerical data collected in the evaluation sessions, a questionnaire was also distributed to the users to get feedback on the subjective aspects of the system. The main problem of the real and the evaluation environments is that repetition of the experiments, in order to evaluate different algorithms and parameters, is impossible.

Third, the analysis or validation of the logs obtained in a real or evaluation environment with real users is a common technique used to evaluate recommender systems. A frequently used technique is the "10-fold cross-validation technique" (Mladenic, 1996). It consists of predicting the relevance (e.g., ratings) of examples recorded in the logs and, then, comparing them with the real evaluations. These experiments are perfectly repeatable, provided that the tested parameters do not affect the evolution of the user profile and the recommendation process. For example, the log being validated would be very different if another recommendation algorithm had been tested. Therefore, since the majority of the parameters condition the recommendation process over time, generally, experiments cannot be repeated.

Finally, a few systems are evaluated through simulated users. Important issues such as learning rates and variability in learning behaviour across heterogeneous populations can be investigated using large collections of simulated users whose design was tailored to explore those issues. This enables large-scale experiments to be carried out quickly and also guarantees that experiments are repeatable and perfectly controlled. It also allows researchers to focus on and study the behaviour of each sub-component of the system, which would otherwise be impossible in an unconstrained environment. For instance, Holte and Yan conducted experiments using an automated user called Rover rather than human users (Holte and Yan, 1996). NewT (Sheth, 1994) and Casmir (Berney and Ferneley, 1999) also used a user simulator to evaluate the performance of systems. The main shortcoming of this technique is that, at present, it is impossible to simulate the real behaviour of a user. Users are far too complicated to predict, at every moment, their feelings, their emotions, their moods, their anxieties and, therefore, their actions.

3 "PROFILE DISCOVERING"

In order to solve all the shortcomings of the current techniques while benefitting from their advantages, we propose a method of results acquisition called *"the profile discovering procedure"* (see Figure 1). This technique can be seen as an hybrid approach between real or laboratory evaluation, log analysis and user simulation.

First of all, it is necessary to obtain as many item evaluations from real users as possible. It is desirable to obtain these user evaluations through a real or laboratory evaluation although it implies a relatively long period of time. However, it is also possible and faster to get the user evaluations through a questionnaire containing all the items which the users have to evaluate. We call the list of real item evaluations of a given user A, the A's real user profile (RUP).

Once the real user profiles are available, the simulation process; that is the profile discovering procedure starts. It consists on the following steps:

1. Generation of an initial user profile (UP) from the real user profile (RUP, $UP \subset RUP$).
2. Emulation of the real recommendation process: a new item (r) is recommended from the UP.
3. Validation of the recommendation:

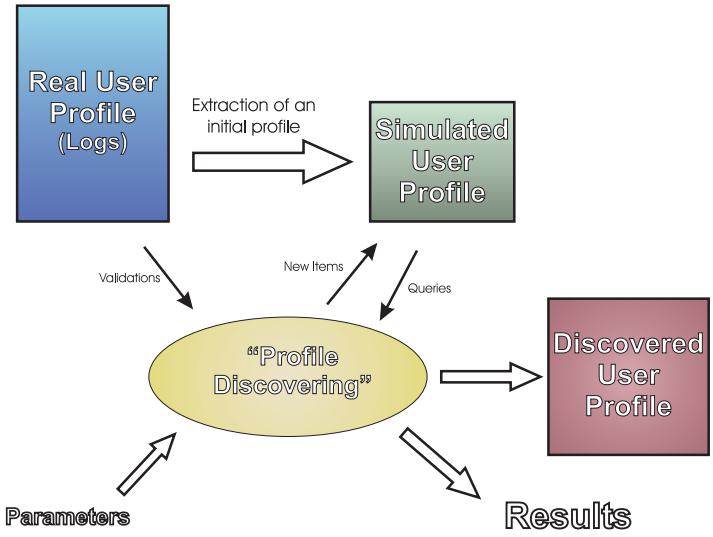

Figure 1: "Profile Discovering" Evaluation Procedure.

- If $r \in RUP$, then r is considered as a discovered item and is added to UP ($UP = UP \cup \{r\}$).
- Otherwise, r is rejected

4. Repeat 2 and 3 until the end of the simulation.

When the simulation process starts, it is desirable to initially know as much as possible from the user in order to provide satisfactory recommendations from the very beginning. Analysing the different initial profile generation techniques (see (Montaner et al., 2003)), namely: manual generation, empty approach, stereotyping and training set, we found different advantages and drawbacks. The training set approach depends totally on the profile learning technique: the user just gives a list of evaluated items and the learning technique generates the profile. There is nothing to annoy the users and the users easily define their preferences. Therefore, in our approach, the training set seems to be the best choice, although the others can also be used in our framework. Thus, the first step of the simulation consists in the extraction of an initial item set from the RUP in order to generate the initial UP.

Then, the simulator emulates the recommendation of new items over time. In particular, it executes a recommendation process cycle by cycle, where a cycle is a day in the real world and corresponds to steps 2 and 3 of the profile discovering procedure. During the simulated day, the recommendation algorithm recommends a group of items based on the information contained in the UP. All the functions, constraints and constants involved in the recommendation process are parameters of the simulator (for example, the time of the simulation, the recommendation algorithm or the learning parameters).

After each recommendation the simulator checks its success. In order to do that, the user's assessments are needed. Instead of inventing them like other simulators do, the profile discovering simulator looks up the real user profile containing the real evaluations of the user. Thus, if the item is contained in the RUP, the simulator can check the user's opinion and classify the recommendation as a success or a failure accordingly. This discovered item is learned in the UP and a new item is recommended.

Once the simulation has finished, the initial UP will have evolved in a more complete profile that is called the discovered user profile (DUP). Moreover, the method provides results on how many recommendations have been made or how many successful/unsuccessful items have been recommended. Thus, based on the DUP and the final simulation results, different metrics are evaluated such as precision, recall and diversity (see (Montaner, 2003) for more details about some measures analysed).

4 "PROFILE DISCOVERING WITH COLLABORATION"

An extended version of the profile discovering evaluation procedure has been designed in order to simulate the collaboration among users in a recommender system. User collaboration is a frequent technique used in open environments since it has been proved to improve recommendation results (Good et al., 1999).

If the current techniques proposed in the state-of-the-art do not allow the proper evaluation of single user, neither are they valid for evaluating a community of collaborating users. Thus, we propose the

"profile discovering evaluation procedure with collaboration". The main idea of this new technique is essentially the same as the profile discovering but it takes into account the opinions and recommendations of other users in the system. The simulation is also performed cycle by cycle. At every cycle new users enter into the simulation process and the recommender system recommends new items to each user based on the simulated user profile with the collaboration from the other users. Thus, the profile discovering evaluation procedure with collaboration consists of the following steps:

1. Initial Profile Generation: as in the profile discovering procedure, an initial UP is generated as from the RUP contained in the logs.
2. Contact List Generation: each user has a list of users with which to collaborate. We refer to these lists as *contact list* and the users contained in them as *friends*. There are several techniques to fill up such list: direct comparison among user profiles (the collaborative filtering approach), "playing agents procedure" (Montaner et al., 2002),... Thus, the simulator emulates the process where each user looks for friends with a technique that is a parameter of the simulator.
3. Recommendation Process: the simulator recommends new items to each user based on their UP and with the collaboration of their friends (see Figure 2). Furthermore, users can also receive collaborative recommendations by directly asking for interesting items to their friends. Finally, after each recommendation, the simulator checks its success based on the user's assessments contained in the RUP.
4. Profile Adaptation: besides classifying the recommendation as a success or a failure, the simulator has to adjudge on the collaboration of the other users. Such information is used in order to adapt the contact list of the recommender systems to the most recent outcomes. The parameters controlling the modification of the contact list are parameters of the system.
5. Repeat 2-4 (a cycle) until the end of the simulation.

Finally, when the simulation duration is exhausted, several metrics are analysed and the results are presented.

5 EXPERIMENTAL RESULTS

The proposed methodology was implemented in order to evaluate GenialChef[1], a restaurants recommender system developed within the IRES Project[2].

The contributions of GenialChef are a CBR recommendation algorithm with a forgetting mechanism and a mechanism of collaboration based on trust (Montaner, 2003). All these contributions were deeply tested with the methodologies proposed in this paper. In particular, the CBR recommendation algorithm, the initial profile generation technique and the different parameters regarding the forgetting mechanism were evaluated by means of the "profile discovering procedure" and several "cross-validations through profile discovering". Then, the method to generate the contact lists, the different parameters concerning how and when to collaborate and the functions and parameters to adapt the contact lists to the outcomes were evaluated with the "profile discovering with collaboration". A snapshot of simulator interface with the different parameters used in the simulations are shown in Figure 3.

One of the most interesting experiments that we performed with this user simulator is the comparison of the information filtering methods used in the recommendation process of GenialChef (Montaner, 2003) with the ones provided in the state-of-the-art. In particular, the opinion-based filtering method and the collaborative filtering method through trust are compared to the typical information filtering methods: content-based filtering and collaborative filtering. Thanks to the simulator, the same set of user profiles where submitted to the different methods, getting comparable results. Figure 4 shows the precision of the recommender system when different combinations of information filtering methods are applied. Y-axis represents the precision of the system and x-axis represents how much tolerant users are when adding new friends to their contact lists, ranging from 0.4 (almost all the other users in the system are considered as friends) to 1.0 (nobody is considered as friend). Executing all the filtering methods upon the same user simulator, we can guarantee that the results are comparable and assure that the information filtering methods proposed (Simulation5) improve the performance of the typical ones.

6 CONCLUSIONS

This paper is focussed on the evaluation of recommender systems. Due to the lack of evaluation procedures for such a personalised systems, we have carried out an important work on how these systems can be evaluated scientifically. The main purpose is that this evaluation procedure be as similar as possible to an evaluation performed with real users. Our

[1]GenialChef was awarded the prize for the best university project at the E-TECH 2003.

[2]The IRES Project was awarded the special prize at the AgentCities Agent Technology Competition (2003).

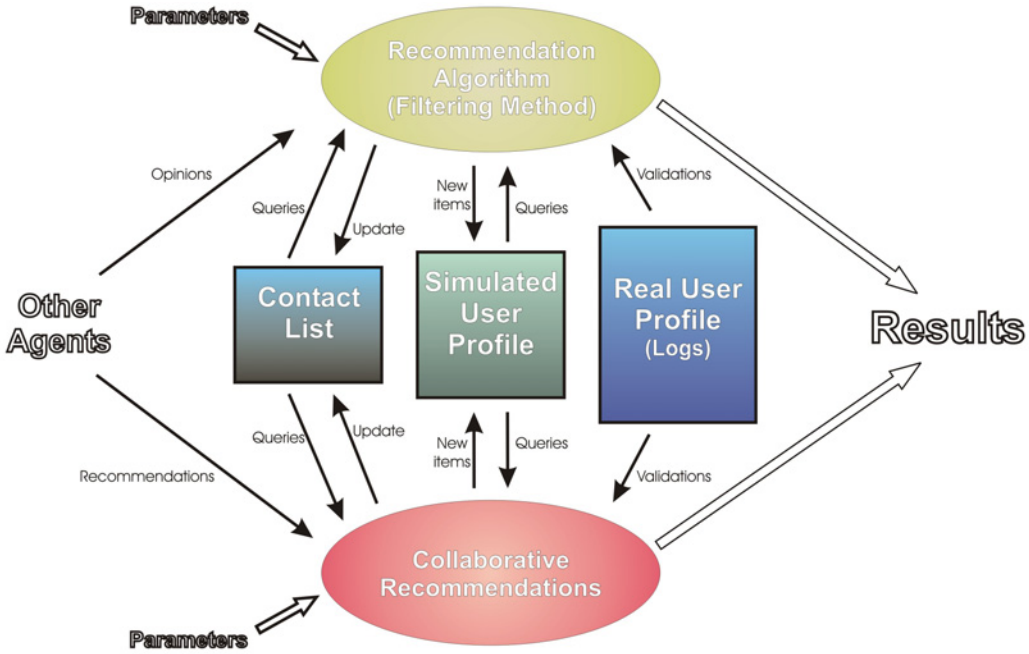

Figure 2: Recommendation Process in Profile Discovering with Collaboration.

proposal, the "profile discovering procedure", is a methodology that simulates users based on a list of item evaluations provided by real users. Therefore, the evaluation is only based on real information and does not invent what users think about items.

Besides, a extension have been designed as a complement to this methodology in order to simulate the collaboration among users.

Therefore, the methodology proposed and its extension allow researchers to carry out large-scale and perfectly controlled experiments quickly in order to test their recommender systems and, what is also very important, compare their whole systems with others.

The next step in our work is to improve our methodology in order to incorporate information about the context of the users and their emotional factors like in (Martínez, 2003). We believe that such information can provide simulated users with a behaviour more similar to the users of the real world.

REFERENCES

Amazon (2003). http://www.amazon.com.

Berney, B. and Ferneley, E. (1999). CASMIR: Information retrieval based on collaborative user profiling. In *Proceedings of PAAM'99*, pp. 41-56.

Good, N., Schafer, J., Konstan, et al. (1999). Combining collaborative filtering with personal agents for better recommendations. In *Proceedings of AAAI*, vol 35, pp. 439-446. AAAI Press.

Holte, R. C. and Yan, N. Y. (1996). Inferring what a user is not interested in. *AAAI Spring Symp. on ML in Information Access*. Stanford.

Martínez, J. (2003). *Agent Based Tool to Support the Configuration of Work Teams*. PhD Thesis Project, UPC.

Mladenic, D. (1996). Personal WebWatcher: Implementation and design. *TR IJS-DP-7472, Department of Intelligent Systems, J.Stefan Institute, Slovenia*.

Montaner, M. (2003). Collaborative recommender agents based on CBR and trust. *PhD Thesis in Computer Engineering*. Universitat de Girona.

Montaner, M., López, B., and de la Rosa, J. L. (2002). Opinion-based filtering through trust. In *Proceedings of CIA'02. LNAI 2446.*, pp. 164-178.

Montaner, M., López, B., and de la Rosa, J. L. (2003). A taxonomy of recommender agents on the Internet. *Artificial Intelligence Review*, volume 19:4, pp. 285-330. Kluwer Academic Publishers.

Sanguesa, R., Cortés, U., and Faltings, B. (2000). W9. Workshop on recommender systems. In *Autonomous Agents, 2000. Barcelona, Spain*.

Sheth, B. (1994). A learning approach to personalized information filtering. *M.S. Thesis, Massachusetts Institute of Technology*.

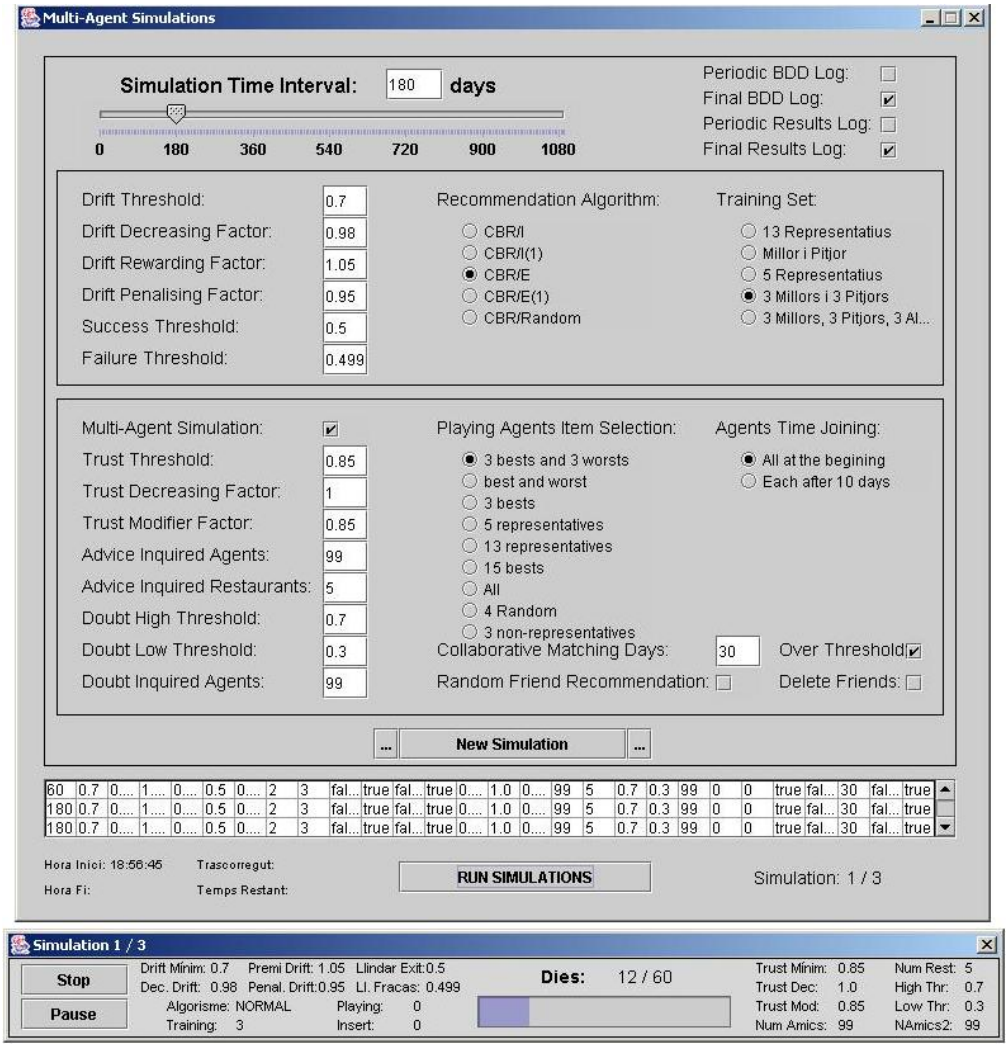

Figure 3: Interface of the Simulator.

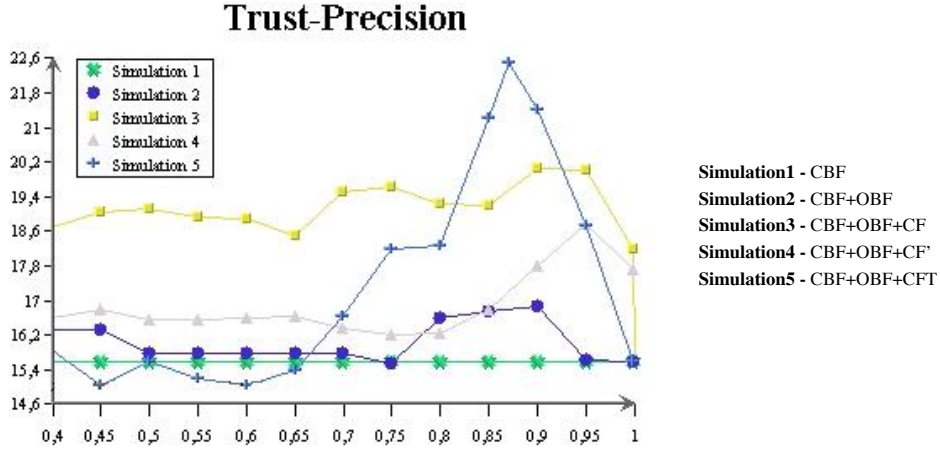

Figure 4: Precision of the system with different IF methods.