

# Towards an Information Filtering System in the Web Integrating Collaborative and Content Based Techniques

José de J. Pérez - Alcázar  
Maritza L. Calderón - Benavides  
Cristina N. González - Caro  
Advanced Computing Laboratory  
Universidad Autónoma de Bucaramanga  
Bucaramanga, Colombia  
{jperez, mcalderon, cgonzalc}@unab.edu.co

## Abstract

*The amount of information currently available, the different media and presentation formats joined with the little time availability of the researchers and people in general, make necessary the implementation of automated tools selecting and evaluating information, aiming at not only optimizing resources but also obtaining useful and personalized results that optimize the daily work of its users. A technique known as Information Filtering could be seen as a solution to this problem. Within an information filtering system, a user introduces a profile in the System which represents his/her information needs; then, the system works to display the relevant information. This article contains a sample of the research carried out by us in this important area, focusing the work towards two of its most representative techniques: "Content Based filtering" and "Collaborative filtering." These techniques have been studied from different points of view, allowing to create a solid framework which involves the necessary criteria for designing and creating a tool using the most outstanding characteristics of each technique. They provide a view to facilitate the work of people devoted to the search, depuration and distribution of information.*

## 1. Introduction

Due to the huge amount of data (images, text, video and so on) circulating nowadays in different media, there is a serious overload problem. People do have a lot of information, but they do not have any tools that allow them to classify it and find what they exactly need. An efficient way to carry out this task, is

the users not searching for resources, but previously selected and evaluated information coming to them. Techniques such as the Information Filtering might be viable solutions to this kind of problem.

Some of the most representative techniques of Information Filtering are Collaborative (CF)[13] and Content Based Filtering (CBF)[9]. They perform their predictions taking into account diverse characteristics, for instance: the CF recommends the user items based on the user's preferences and on other users with similar interests. On the other hand, the content based technique only exploits information that can be derived from the content of the objects to give the user an answer to their needs. Besides, CF and CBF have proved to be complementary techniques [10]. A perfect technique based on content will never find something new, limiting the range of applications by which it can be useful. The CF techniques are outstanding at identifying a novelty but only when humans feedback the system.

In considering the importance and utility of the Information Filtering (IF), we have decided to focus part of our efforts on research in this area. Several works oriented to the analysis of the aforementioned techniques have been carried out. Firstly, the basic aspects of each one of them have been analyzed up to the point of reaching the level of more complex details such as its joined behavior. All of this with a view to improve their management and reach the level of ability to design and implement a IF tool using the best characteristics of each one of them. The current poster is a sample of the research carried out and the results obtained.

The poster is divided in the following way: Section two shows some works carried out by us and some re-

sults obtained; section three presents the architecture proposed for the Information Filtering System to be implemented by us. Finally, in section four, some conclusions of our work are presented.

## 2. Works carried out

We have been working on a series of projects related to Information filtering. At the beginning, an information dissemination system for the Technical Information Center (CIT) for the Colombian Oil Institute (ICP) was developed. The objective of that prototype, besides facilitating and making the CIT's Information Disseminators job more efficient, was to acquire a mastering of the filtering basic concepts and to understand the work scheme of filtering techniques. As a next step to that job, three more pieces of work were carried out and a fourth one is being carried out: a wider comparison on the collaborative filtering algorithms existent in literature, an initial comparison of proposals of combined content-based and collaborative filtering techniques, the development of a prototype based on information dissemination agents, and a comparison of information retrieval models. The latter is very useful at the implementation of content based filtering systems. Next, a description of these pieces of work is shown:

- The first piece of work was oriented towards a study of a sample of collaborative filtering algorithms that although sharing a same goal, assume the problem of recommending the users items from different points of view. This work was conducted using binary data.

The algorithms selected for this evaluation were: memory based algorithm (MB) (Pearson Correlation [13]); Dependency Networks (DN) [6]; Aspect Model(MA) [7]; Support Vector Machines (SVM) [14]; On-line learning (Weighted Majority Prediction - WMP and Memory Weighted Majority Prediction MWMP [4]). Experiments were held on two data sets with different characteristics. The results obtained in this study indicated that, in general, Online Learning algorithms and SVM have a better performance than the other models on the data sets used. Such algorithms perform well on different amounts of available information and diverse data dispersion conditions. A description more detailed in this empirical study can be found in [5].

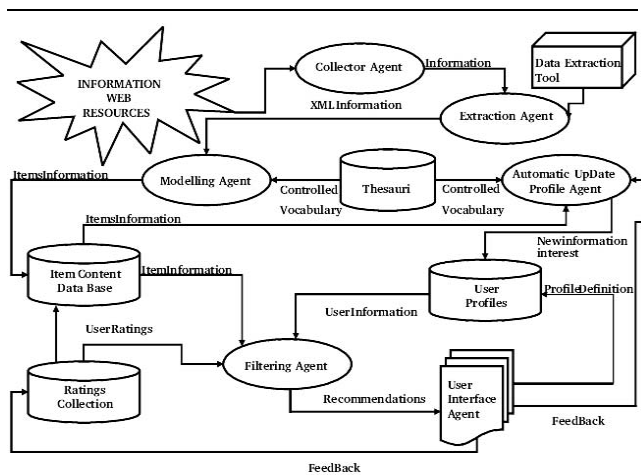
After carrying out a piece of work focused on the analysis of diverse algorithms for the collaborative filtering on binary data, we carried out a more

complex study on this topic. To do that, three data sets for two different information domains (movies and jokes) were selected. These data sets use diverse rating scales of their items. The same algorithms evaluated in the first work were considered here. In general terms, the results of this work show that the memory based algorithms are a good option due to the fact that its results are more precise and reliable in comparison with the other methods.

- As it was previously stated, collaborative and content-based filtering are techniques that show to be complementary. On the experimental level, We have carried out some pieces of work combining these techniques taking into account diverse aspects of its interaction with a view to determining its characteristics, performance and execution. We used the following models: Pure collaborative filtering [13]; pure content-based filtering [15]; Melville et al. improved collaborative filtering [11]; Polcicova et al. combined proposed model [12]; and the model proposed by Claypool et al.[3].

As a result of this study some aspects about the performance of the hybrid models were confirmed, and some others providing new leads for further studies were also established. The incorporation of the content-based technique in order to complement the collaborative filtering model, significantly improves predictions, and also makes possible recommendations when dealing with a new user or with highly different interests from the rest of the community. Integration models implemented offered better results than the content-based or pure collaborative filtering systems. The model proposed by Polcicova et al. [12] throws better results than the other models. The results of this work, described in detail, can be found at [8].

- Although the purpose of this work was to obtain more experience in the use of methodologies of analysis and design of multi-agent systems, an additional pretty important result was the implementation of a prototype of selective information dissemination (information filtering) based on agents on the Web [1]. This system includes a module for the extraction of information based on the data contained in the web pages of the predetermined digital libraries. The system also combines content-based and collaborative filtering techniques. The combination of extraction techniques with filtering techniques allows a continuous addition of digital libraries (web sites) to our system in an automatic or semi-automatic way;



**Figure 1. Architecture of the proposed system**

that way we add a scalability characteristic into our system.

Currently, there is a project in a development stage whose main objective is to perform a detailed study of several existing information retrieval models [2] (useful in content-based filtering), with the aim of analyzing their operation, computational complexity and the precision of the results offered by the selected models.

### 3. Proposal of an Information Filtering system in the Web

The purpose of this system is to take the most important characteristics from the results obtained with works already made, and implement them in a multi-agent system. The purpose of this system is to disseminate existing information in different digital libraries and electronic magazines available on the Web.

The architecture proposed for our system is presented in Figure 1. The Collector Agent is in charge of collecting information contained within different Web resources such as Digital Libraries, Journals, Magazines, etc. registered by the system administrator. This download is made periodically depending on the publishing or domains (periodicity of update). For the collector agents to download items, it is necessary to generate the extraction agents associated to the registered web sites. These extraction agents allow to convert the information from HTML to XML, and they are generated by the Data Extraction Tool. Once the items are registered in XML, their fields are used by the Modelling Agent to be modelled through VSM (Vector Space Model)[2] and stored after in the Item Con-

tent Data Base. The Modelling Agent uses a controlled vocabulary (Thesaurus) which normalizes the system's terms. Users can define their profile explicitly through the User Interface Agent, as well as receiving its recommendations and do feedback (ratings) about them. The purpose of using such feedbacks is two-fold. The first one is to serve as a base for the Automatic Update Profile Agent to update the profile; those items valued as positive are analyzed in content and compared to the profile in order to determine new interests and control the existing ones (some topics included in the user profile can no longer be of his interest). The second purpose is to be stored in the Ratings Collection to be used by the filtering process, in a later stage.

To do recommendations through the Filtering Agent, we took into account the experiences obtained in previous research. Having in mind that the models involving collaborative filtering either as a exclusive technique or as a combination, require user ratings, the system make the recommendations by content when initiating the use of the system since there is total absence of ratings; in this way, people will have knowledge of items and will initiate a feedback on them. The algorithm of content based filtering will also be used for those items which do not have any rating, such as new documents; in this way one of the main problems of collaborative filtering can be solved (absence of ratings), and at the same time the user is assured a constant process of recommendation. When there are available ratings in the system (a minimum number of ratings per item is defined), the combination technique described by Polcicova et al. [12] is used, because it offers good results and it is a concrete way to use both models (content based and collaborative) in order to obtain a prediction. Then, the Filtering Agent estimates the necessary correlation indices and makes recommendations to the users, using the technique described before.

### 4. Conclusions and future works

In this poster we presented studies made by us with the objective of developing an Information Filtering System in the Web. Thus, the system's architecture is the product of studies on the most known and updated techniques existing in literature. These studies have been done through an extensive literature review and empirical evaluations of those techniques. This system is not a finished work, but it is intended to be in a constant evolution process.

As a result of this work, we have obtained a series of contributions such as: a wider comparative evalua-

tion of the collaborative filtering algorithms; a study and comparison of the integration techniques of collaborative and content-based filtering algorithms; and the use of data extraction tools to improve the system's scalability.

## 5. Acknowledgments

We express our gratitude to Juan Carlos García Ojeda, Juan Carlos García Díaz and Olga Monroy for their constant support. This work was partially funded by the Colombian Research Council (COLCIENCIAS-BID).

## References

- [1] A. Arenas, J. C. Garcia, and J. J. PerezAlcazar. On combining organisational modelling and graphical languages for the development of multiagent systems. *Integrated Computer-Aided Engineering Journal (To be published)*, 2003.
- [2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.
- [3] M. Claypool, A. Gokhale, T. Miranda, P. Murnikov, D. Netes, and M. Sartin. Combining content-based and collaborative filters in an online newspaper. In *Proceedings of ACM SIGIR Workshop on Recommender Systems*, 1999.
- [4] J. Delgado and N. Ishii. Memory-based weighted-majority prediction for recommender systems. In *Proceedings of the ACM SIGIR-99, Recommender Systems Workshop, August 1999, UC Berkeley*, pages 1–5, 1999.
- [5] C. N. González-Caro, M. L. Calderón-Benavides, J. de J. Pérez-Alcázar, J. C. G. Díaz, and J. Delgado. Towards a more comprehensive comparison of collaborative filtering algorithms. In *Proceedings of the 9th International Symposium on String Processing and Information Retrieval (SPIRE'02)*, Lecture Notes in Computer Science, pages 248–253. Springer, 2002.
- [6] D. Heckerman, D. Chickering, C. Meek, R. Rounthwaite, and C. Kadie. Dependency networks for inference, collaborative filtering and data visualization. *Journal of Machine Learning Research*, 1:49–75, 2000.
- [7] T. Hofmann and J. Puzicha. Latent class model from collaborative filtering. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*. Morgan Kaufman Eds., 1999.
- [8] O. Monroy. *Análisis de la Combinación de Modelos de Filtrado de Información*. Master thesis, Master in Computer Science, ITESM-UNAB-CUTB, June 2003.
- [9] D. W. Oard and G. Marchionini. A conceptual framework for text filtering process. Technical Report CS-TR-3643, University of Maryland, College Park, 1996.
- [10] D. Penneck, E. Horvitz, S. Lawrence, and C. L. Giles. Collaborative filtering by personality diagnosis: A hybrid memory- and model-based approach. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, UAI 2000*, pages 473–480, Stanford, CA, 2000.
- [11] P. Melville, R. Mooney, and R. Nagarajan. Content-boosted collaborative filtering. In *The Proceedings of the SIGIR-2001 Workshop on Recommender Systems*, New Orleans, LA, 2001.
- [12] G. Polcicova, R. Slovak, and P. Navrat. Combining content-based and collaborative filtering. In *ADBIS-DASFAA Symposium 2000*, pages 118–127, Prague, Czech Republic, 2000.
- [13] P. Resnick, N. Iacovou, M. Suchak, P. Bergstorm, and J. Riedl. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, pages 175–186, Chapel Hill, North Carolina, 1994. ACM.
- [14] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, NY, 1995.
- [15] T. Yan and H. Garcia-Molina. SIFT—A tool for wide-area information dissemination. In *Proc. 1995 USENIX Technical Conference*, pages 177–186, New Orleans, 1995.