ISBN: 972-99353-6-X © 2005 IADIS

PERSONALIZED FILTERING SYSTEMS BASED ON THE MULTI-METHODS COMBINATION

João Ferreira
ISEL
jferreira@deetc.isel.ipl.pt

Alberto Rodrigues da Silva
INESC-ID, IST
alberto.silva@acm.org

José Delgado
IST
Jose.Delgado@tagus.ist.utl.pt

ABSTRACT

We propose a modular platform to support the development of personalized filtering systems. According our proposal, filtering systems can be constructed through the integration of different modules and changes on specific parameters. We also introduce a hybrid approximation to improve filtering performance based on the combination of content and collaborative filtering, which suppress weakness of each traditional approach.

KEYWORDS

Fusion, Filtering Information, Profile, Combination, hybrid approximation

1. INTRODUCTION

Nowadays personalized filtering information is becoming progressively an important research and industrial topic. Internet as well as the increasing convergence between telecommunications, computation and media industry (e.g., digital TV) contributes decisively to that importance. This situation raises huge quantity of information, either structural or unstructured and multimedia data. Of course, new mechanisms and services should be provided in order to help end-users to discover the information they need and bring together producers and consumers of information. Filtering can also be viewed as a personal intermediation service, like a librarian, and a filtering system can collect information from different sources. We define filtering as an asynchronous service composed by two complementary workflows. Firstly, identification and classification of heterogeneous sources of information! the relevant information to users known by the respective system. Secondly, end-users notification about significant results and keeping track of their reactions in order to use this feedback to tune future results and refinements.

Information Filtering, has been largely discussed since December of 1992, when an issue dedicated to this topic appeared in Communications of the ACM [1]. Since then, several systems have been built based on common information filtering and retrieval techniques and today rather than simply removing unwanted information, filtering services give users the ability to reorganise the information space. These themes are still research, far way from being solved. This subject comes again in Communications of the ACM, in March 1997, in an issue dedicated to *Recommended Systems* [2] and in August 2000 with the topic *Personalisation* [3]. This last issue reflects the evolution of filtering systems to personalisation, and we can see the evolution of previous systems build between 90 and 94. For example, we can point out My yahoo [5], based on Firefly `www.firefly.net` and Personalised Television [6]. An overview of filtering systems can be found at `www.ee.umd.edu/medlab/filter` and a survey on hybrid approximation [7]. Fusion and combination of methods will have the same meaning in this paper.

In this paper, we address the issue of personalized filtering information by pursuing two objectives. First, we propose a generic architecture to support the design and the construction of personalised services based on the combination of individual methods. Second, we present and discuss the results of this architecture on the application MyTv, which is an application to advise TV programs from Portuguese TvCabo `www.tvcabo.pt` to registered users and MyEnterpriseNews.

The filtering architecture should handle conveniently the use and adoption of the following subjects:

– Classification systems for information spaces: techniques and strategies for normalisation of information

spaces (documents and users) through classification systems.
- User profiles: techniques and strategies for the definition, creation and maintenance of user profiles.
- Collections of documents: techniques and strategies to help defining sub-spaces in collections of documents (thus defining clusters of documents to optimise the tasks of classification and search).
- User communities: techniques and strategies for the definition of communities.
- Multilingual issues: techniques and strategies for cross-language information search and retrieval based on the help of multilingual classification systems. In this case we will explore the possibility of given the same news in different languages having as source different countries.

2. OVERVIEW OF FILTERING SYSTEMS

The purpose of the Filtering systems is to help end-users finding what they want in a large set of information; this is a problem with huge relevance in the information society. Currently, newspaper editors select which articles their readers will potentially read. Similarly, book publishers decide which books to print. Electronic information removes these barriers, allowing an easy and cheap access to information and these results in a growth of information created or exchanged by an order of magnitude.

Due to this overload of information in several fields, filtering systems appeared and their number has been increasing. We can divide these systems in two main categories [6,8]:
- Content-based filtering systems: they present an automatic approach based on matching user profiles against document representatives (word or sentences).
- Collaboration-based filtering systems: they present a social approach based on; (1) Matching user profiles against explicit user judgements (annotations to documents); (2) Matching user profile against other user profiles. In this way we can use other users' judgements that have a similar profile. Match the profile against other profiles and to choose the information in the nearest one; (3) Matching user profile against standard profile of communities. Match the profile against community standard profile and to use the nearest standard to get information (social filtering).

Matching techniques can be Boolean, vector space, probabilistic or connectionist networks. User profiles can be created using an explicit method and these profiles can be improved using machine-learning techniques based on explicit feedback or user observation. Currently, user profiles are one of the richest areas of research, especially in the implicit approach. There are several experiences, for example, using the time spent reading, analysis of users' bookmarks, server log files, etc.

Content-based filtering systems have had success only in very simple collections. The main problem is that they have to deal with the issue of automatic creation of representatives of documents (or surrogates), a complex task even for well-defined areas.

Due to human subjectivity and to achieve better results, several systems, which we call *collaboration-based, filtering systems*, also involve humans in the filtering process. For example, in some cases, user reactions to the documents are recorded (such as ranking, notes, etc.) and later used to help other users. This kind of systems is known as recommend collaboration or social filter systems. These systems are in general more successful than the automatic ones, but unable to provide information in documents that have never been read. Another weakness is the problem of finding the correct tools to keep out (or to minimise the effect) of disruptive users (e.g., users who are not really collaborative but only interested in giving high rates to themselves or to related friends).During the last decade, filtering systems have been applied mainly to: Usenet news, mailing lists, technical reports, WWW spiders, music and movies/video. Now these systems are being applied to emergent applications, such as digital libraries, personalised electronic newspapers and TV.

3. PLATFORM ARCHITECTURE

In this section, we describe our proposed platform. The main advantages of a platform are: (1) less work for testing methods and algorithms; (2) a common infrastructure provides better results comparison. The platform is implemented, as shown in figure 1 and 2, in a modular way:
- Collaborative filtering: Relevant information is identified from judgements made (of information read) in

a scale of 1 to 5, from the five most similar judgments (all users) and also community central profile. Results are merged by SM (Similarity Merge) combination formulae (1).Relevant documents are stored for the final merge with content based filtering.

- Content based filtering, we identified 3 main approaches: (1) Matching from user profile with documents; (2) Matching from community profile (identified from cluster analyses) [8] with documents. All results are combined through SM formulae.
- Link analyses from both relevant documents identified in the process above (1,2) and collaborative filtering.

The last step is the combination through ROWRS formulae (3) of content based with collaborative filtering. This combination suppresses problems of individual approaches: contend based have difficulties to deal with context and meaning and collaborative with new items (approach only identified items already read) and wrong rates (made by users in a trend that benefits the same items). Also links analyses of all document identified as relevant contribute for more.

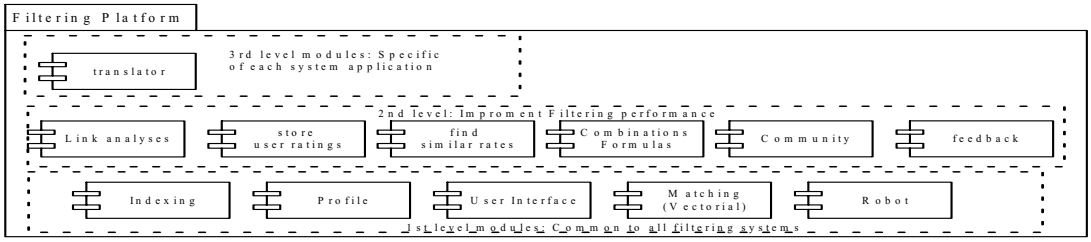

Figure 1. Main Platform modules.

SM formula, combines and computes the combination score of a document by the sum of normalized component scores boosted by the retrieval overlap different (combination score) **(1)**; where: olp = number of systems that retrieved a given document, m(i) = number of systems in a method to which system i belongs; NSi=(Si–Smin)/(Smax–Smin), **(2)** [9], NSi = normalized score of a document by system I, Si is the retrieval score of a given document and Smax and Smin are the maximum and minimum document scores by system i:

ROWRS formula uses rank-based scores (e.g. 1/rank) (this allow combinations of systems with different metrics), and takes in account the overlap of results and rank at which a document is retrieved in its computation of weights: **(3)**, where: w_{ikj} = weight of system *i* in overlap partition *k* at rank *j* and RS_i = rank-based score of a document by system *i*.

Link analyses are performed by modifying the HITS formulas: **(4)** and **(5)**, where *auth_wt(q,p)* is 1/*m* for page *q*, whose host has *m* documents pointing to *p*, and *hub_wt(p,q)* is 1/*n* for page *q*, which is pointed by *n* documents from the host of *p*.

Community profile are created by clustering user profiles using **(6)**, (profile), where f_k is the number of times the term *k* appears in the query, and idf_k is the inverse document frequency of term *k*. The denominator is a document length normalization factor, which compensates the length variation in queries. Cluster is performed by $p^{\mathrm{T}} p_i = \sum_{k=1}^{t} p_k p_{ik}$ **(7)**.

$CS = (\sum NS_i) * \frac{olp}{m(i)}$ **(1)**	$h(p) = \sum_{p \to q} a(q) \times hub\_wt(p,q)$ **(4)**	$p_k = \frac{(\log(f_k)+1) * idf_k}{\sqrt{\sum_{j=1}^{t} \left[(\log(f_j)+1) * idf_j \right]^2}}$ **(6)**
CS= $\sum(w_{ikj} * RS_i)$ **(3)**	$a(p) = \sum_{q \to p} h(q) \times auth\_wt(q,p)$ **(5)**	

In order to validate the requirements and features of our platform we developed two prototype applications: (1) MyTV; (2) MyEnterpriseNews. Both prototypes use common platform modules with addition of specific modules. Main filtering modules are responsible for indexing information as well as for storing and handling profiles.

The MyTV prototype: This system personalizes TV information, sending relevant information about Tv programs to users by integrating a range of different information-filtering strategies, content based, link and collaborative filtering, with user-profiling techniques. As information source we use Portuguese TvCabo `www.tvcabo.pt` (around 60 channels). Larbin <**larbin**.sourceforge.net/> crawler was configured to take information from TvCabo website daily. Wordtrans `wordtrans.sourceforge.net` makes translation when it

is necessary. We have around 30 registered users (campus users) and results of 3 months of the system run in a closed environment due to the delay in the development of www interfaces. All users evaluate the results of the system based on guide precision of 3 approaches: collaborative (Co), content based (Cb) and combined results (Cr). Every time a user receives a recommendation we receive also an evaluation sheet. Results are analyzed in a monthly baisis and the results are presented in figure 2. The combination of results (Cr) improves the final results and we noticed that collaborative filtering (Co) helps in system orientation (identifies main subjects) and content based filtering refines the information. Co works better with an high number of users. We notice that link analysis in this situation doesn't provide a great number of links to follow, but in other subject areas it could have greater potential. Negotiations with Portuguese TvCabo are in process in order to put online a system integrated with TvCabo website.

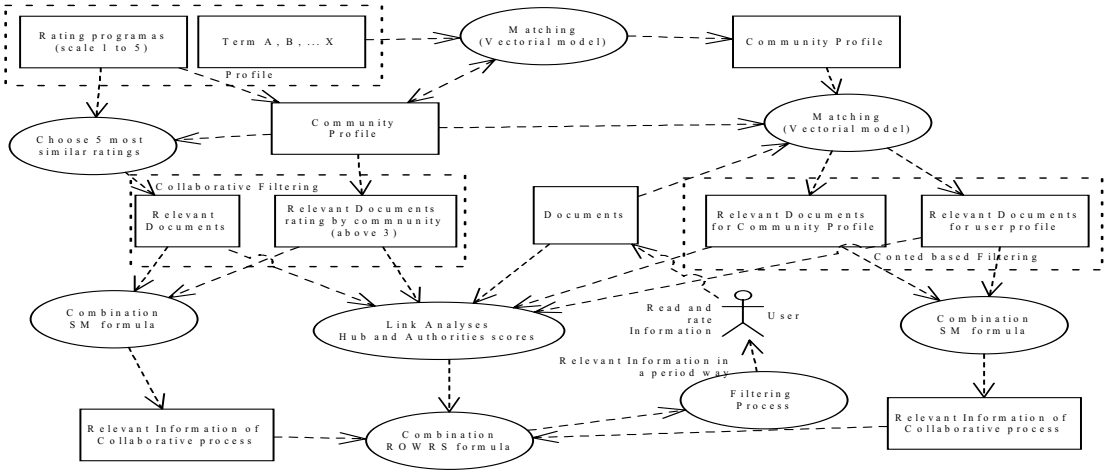

Figure 2. Main components of Personalized Filtering System.

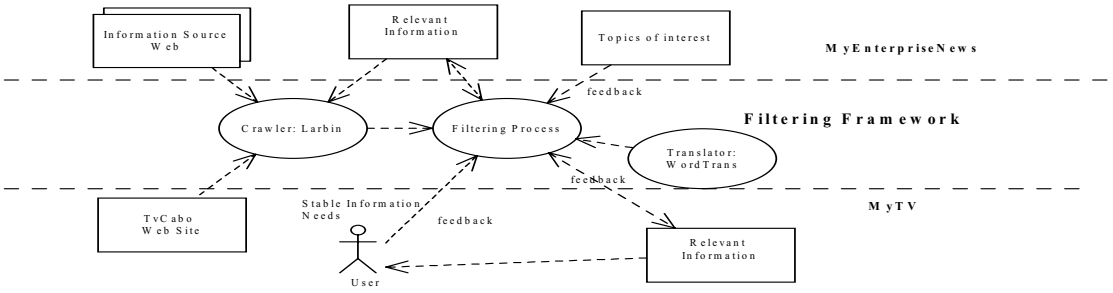

Figure 3. MyTv and MyEntrepriseNews systems build form Filtering Platform.

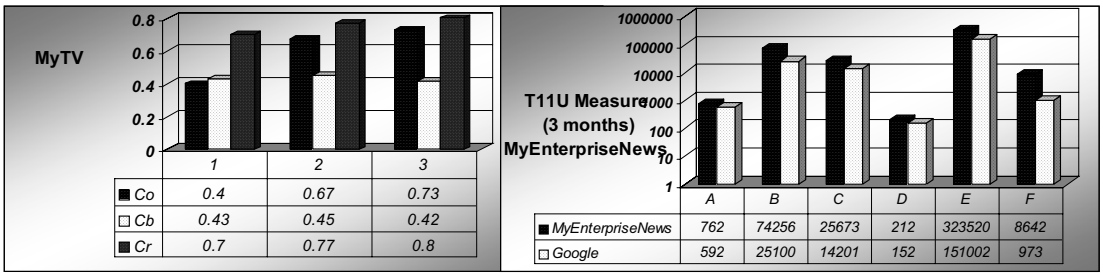

Figure 4. Evaluation results MyTv(left) and EnterpriseNews prototype (right).

The MyEnterpriseNews prototype: The main goal of MyEnterpriseNews system is to find out and provide information related with a company on the Web (figure 3). We tested the system with 6 enterprise cases in a period of 3 months; A – Motor Company (AutoEuropa) and a car (Sharan); B – INESC (investigation institute) on .pt domain; C – ISEL (university) under .pt domain; D - CEDP (Company); E – Sporting (futebol team); F – Information Retrieval under .pt domain. We use linear utility measure T11U=2*(nº Relv. Doc. Retrieval)-(nº non-relevant doc. Retrieval). We compare our results with Google results with an advanced query to restrict results to the last 3 months. In this case, collaboration is reduced to zero and the system benefits a lot from the combination of link analysis, vectorial matching and feedback (automatic). In case A profile was increase to MPV, B and C to most common investigators, E profile was expanded with name of players, coach and president. In this case we introduce negative profile to avoid wrong information like Sporting Braga. Results of figure 4, shows big improvements over Google results. Also, link analysis identified relevant URLs that are used to guide crawler search in an optimization process. We check that the chosen source of information has an high influence in the for system performance.

4. CONCLUSIONS AND FUTURE WORK

We argued that personalisation is a very important requirement for new and emergent filtering services. We proposed in this paper a generic but concrete platform to support filtering personalised services, which we have applied into two case studies: MyTv and MyEnterpriseNews. To improve results, we proposed a new hybrid approach based on the combination of content and collaborative results. At the same time we build filtering information system using a platform with share modules. This approach can avoid the proliferation of different filtering systems build from scratch. With this approach, everyone can use or build specialised filtering systems that can be used in different domains and with specific requirements. In the future, a system can easily be built from a standard platform. Apparent potential benefits are: (1) Faster application developments due to the usage of common guidelines and existing modules, which can be re-used very easily; (2) a decrease in development costs due to shortened development time and collaboration; (3) an increase of wide spread use and user acceptance due to software availability and similar user interfaces.

Nevertheless, there are still several problems and challenges. One of those is the fact that typical users are lazy and easily give up. Usually they provide few terms in their profile specification, and often are not accurate and also don't choose the right terms. To minimise this problem, it is important to build user interfaces with easy and efficient dialogues and also use techniques to expand the terms provided by them.

Personalised filtering services bring advantages for users because they see only what they want; don't lose time to find the information; and generate less traffic in the communication over the Internet. However, they require more work and interaction compared to traditional filtering services, fact that constitutes a real challenge for the development of personalised services.

Finally, we know that the value of filtering services usually comes from the size of registered users. For that, we intend to develop, based on the MyTv, a public and robust service. From this service we intend to explore this prototype towards diffusion of information to well defined communities.

REFERENCES

[1] Communication of ACM, Information Retrieval Special Issue. December 1992, Vol. 35, N 12.

[2] Communication of ACM, Recommended systems Special Issue. March 1997, Vol. 40, N 3.

[3] Communication of ACM, Personalization Special Issue. August 2000, Vol. 43, N 8.

[4] Manber, U., et al. Experience with Personalization on Yahoo!. Communication of ACM, 2000, Vol 43, N 8.

[5] Smyth, B. et al. A Personalized Television Listings Service. Communication of ACM, 2000, Vol 43, N 8.

[6] Ferreira, J et al.,1997. Collaborative Filtering for a Community Digital Library using LDAP. Published and presented at 5th. Delos workshop, 10-12 November 1997 in Budapeste.

[7] Burke R. Hybrid Recommender Systems: Survey and Experiments,Vol.12,Issue 4,Nov.2002,(331-370).

[8] Goldberg, D. et al, 1992. Using collaborative Filtering to weave an Information Tapestry. ACM, 1992, V 35, N 12 .

[9] Lee J. H. Analyses of multiple evidence combination. Proceeding of ACM SIGIR, 1997 (267-276).