MySDI: A Generic Architecture to Develop SDI Personalised Services

(How to Deliver the Right Information to the Right User?)

JOÃO FERREIRA, ALBERTO SILVA
ISEL / INESC ID Lisboa – Portugal** , **IST / INESC ID Lisboa – Portugal
Email: jferreir@acm.org, Alberto.Silva@acm.org

Keywords: Information System, Personalised Architecture, Information Retrieval, Agent, Profile, Filtering.

Abstract: We introduce in this paper a generic architecture to deal with the general problem: “How to Deliver the Right Information to the Right User?”. We discuss this issue through the proposal of our SDI (Selective Dissemination of Information) Personalised Architecture, called MySDI, which is based on the software agent paradigm as well as on information retrieval techniques. In order to clarify and validate this proposal we also introduce in this paper a prototype service, called MyGlobalNews, which should be a public service to provide personalised news.

1 Introduction

Nowadays the selective dissemination of information (SDI) becomes progressively an important research and industrial topic. Internet as well as the increasing convergence between telecommunications and computation and other information broadcast (e.g., digital TV) contributes decisively to that importance. This situation raises huge quantity of information, unstructured and multimedia information. Of course, news mechanisms and services should be provided in order to help end-users discover their right information and bring together producers and consumers of information. SDI can also be viewed as a personal intermediation service, like a librarian, and a filtering system can collect information from different sources. This information doesn’t have the value of librarian information, but can be performed automatic – an important feature due to the amount of information produced nowadays.

We define SDI as an asynchronous service composed by two complementary workflows. Firstly, identify and classify, from heterogeneous and continuously updated sources of information, the information relevant for users known by the respective system. Secondly, notify these users about those significant results and keep track of their reactions (e.g., if they manifest their dissatisfaction) in order to use this feedback to tune future results and refinements.

The problem of the SDI is related with the issue of information filtering and information retrieval, which have been largely discussed since December of 1992 when appeared in Communications of the ACM an issue dedicated to these topic [1]. Since then, several systems have been built based on common information filtering and retrieval techniques and today rather than simply removing unwanted information, SDI services gives users the ability to reorganise the information space. These subjects are still an actual item far way of 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 [4]. This last issue reflects the evolution of SDI systems to personalisation, and we can see evolution of previous systems build between 90 and 94. For examples we can point out, My yahoo [5] constructed based on Firefly [6] and Personalised Television [7] based on ClixSmart [29]. One overview of SDI systems can be consulted at <www.ee.umd.edu/medlab/filter>.

In this paper we will address the issue of SDI following two purposes. Firstly, by proposing a generic architecture to support the design and the construction of SDI personalised services. We call this architecture MySDI. Secondly, by presenting and discussing, based on the MySDI architecture, a specific service, which we call MyGlobalNews, which is a agent-based personalised news public service.

The MySDI 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 a possibility of given the same news in different languages having as source different countries.

This paper is organised in 5 sections. Section 2 ("Overview of Selective Dissemination of Information") gives a brief state of art of the SDI motivation, services and main problems. Section 3 ("The MySDI Architecture") presents the MySDI architecture, which is the main focus of this paper. Section 4 ("A MySDI Prototype: The MyGlobalNews Service") introduces the MyGlobalNews service, which is useful to clarify and proof the concept of the MySDI architecture. Finally, in Section 5 ("Conclusions and Future Work") we conclude and discuss the importance of our proposals and depict the future work.

2 Overview of Selective Dissemination of Information

The purpose of the SDI 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 grown of information created or exchanged by an order of magnitude.

Due to this overload of information in several fields, filtering systems appeared and the number has been increasing. We can divide these systems in two main categories [8, 9]:

- 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
  - Matching user profiles against explicit user judgements (annotations to documents).
  - 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.
  - 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 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 it has 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 called collaboration-based, filtering systems, involve also humans in the SDI 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. Those kinds of systems are known as recommend collaboration or social filter systems. Those 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 (such as, users who are not really collaborative but only interested in giving high rates to themselves or related friends).

During the last decade, SDI systems are being 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 personalised TV.

Table 1 identifies some of initial SDI systems with the focus on the relationship of user profiles and matching techniques. Sift [10] and Newsweeder [11] are two examples of content-based (automatic) SDI systems. The basic difference between them is the way profiles are defined (Newsweeder also uses an implicit method based on past user experience). On the other hand, Grouplens [12], Firefly [6] and ReferralWeb [13] are examples of collaboration-based (social) SDI systems.

This initial SDI systems still works and was the basis for new current systems, (examples, ClixSmart [29] and My yahoo [5]), and other commercial system which information are not available (ZyFilter [30], Agentware i3 [31] and KE Media [32]).

3 The MySDI Architecture

MySDI is a generic and conceptual architecture to SDI personalised services. It is a generic and conceptual architecture because its intent is not to be “yet another real system”, but just to provide a better way to analyse, compare and discuss a multitude of recent and emerging SDI services and platforms. (Of course, MySDI can (and will be) also be used as a reference model to the development of component-based SDI services).

Figure 1 depicts the main concepts of MySDI architecture and their respective interactions through a UML diagram, namely:

- User space: a space to store and manage users and profiles (see Section 3.1).
- Information space: a space to store and manage documents (see Section 3.2).
- Classification space: a space to store and manage classification systems and thesauri (see Section 3.3).
- Filtering system: a system able to identify, from the information spaces, relevant information for the users registered in the user space (see Section 3.4).

The described component's structure will be integrated through a communication space supported by different protocols such as HTTP, LDAP, TCP/IP, SQL, electronic mail and other appropriate protocols. However, it is not relevant in this paper a discussion about these technology details.

User’s information with their long-term interests will be handled and stored in the user spaces as profiles. The filtering blocks are supposed to use those profiles and confront them against the information space. The relevant information found in that process will be send to the user through some delivery/notification mechanism. The role of the classification spaces is to assure pre-coordination for the information and users spaces, so making it easy to establish conceptual relationships.

3.1 User Space

Two different interfaces to interact with the system will be available for the user: electronic mail and Web. The Web interface will run in an interactive way and the mail interface following a batch process. Using these interfaces users can archive and check personal information, store and configure the frequency of messages and give feedback on documents received.

3.1.1 Profiles

One of the key points of our system is to build an efficient profile mechanism that represents the exactly user’s information needs. Users are classified following two categories: experience-user and novice-user.

- Experience-users will provide terms that define their interests (profile) and terms that are definitely not interesting for them (negative profile). All these terms should be normalised in a pre-coordination process, based on the classification systems. Nevertheless, these terms are always defined with a final acceptance of their users.
- Novice-users will use the existing classification space and community space to choose the subjects/communities that they can be interested on.

Profiles can be divided in two categories: positive and negative. Positive is used to retrieval documents which user are interested and negative profile is used to avoid relevant documents which user manifest his dissatisfaction. This profile can be built directly (user provide the information) or indirectly by the system, through the study of user behaviour. Retrieval techniques are applied on both profiles, one to show results (positive profile) and the other (negative profile) to remove relevant documents.

User profiles serve three main purposes, as shown in Table 2:

- Automatic notification: profiles are used to provide an automatic notification service, supported by electronic mail, through which users receive automatic notification of new events.
- Searching: profiles can be used to rank search results, for example highlighting documents that better match user’s interests (but ranking will never hide or restrict the access to other documents that also match the queries).
- Retrieval: the access to different kinds of documents or the access to special user information can depend of the user profile. This feature will require defining profile fields not controlled by the user but by an administrative authority. (This scenario will concern with privacy issues that we don’t handle in this paper).

Automatic notification can involve five kinds of events:

- New documents: any user whose profile matches the classification of a new document is informed about its arrived to the data base archive.
- Changes in stored documents: if a new version of a document is submitted (document title assumed to be the same), users (that, for example, had retrieved that document) will receive a notification.
- New document classification available on catalogue: any user whose profile matches a new document description of subject will be notified about.
- New users: when a new user is registered, users with similar profiles will be notified.
- Changes in user profiles: when a user profile changes, users matching the new profile will be notified.

As said above, user profiles can also be used to rank search results, giving more relevance to results that best match the profile of the user. For this task, it is also possible for the user to choose to identify herself with a standard profile created by the system (a default or community profile), instead of her own profile. In case of the "community profile" the system should identify groups of users with similar profiles, and some human authority should decide if that makes sense and eventually create a standard profile for that community.

3.2 Information Space

The information space is a heterogeneous space with formal and informal information. This space is distributed over several servers. The formal space is divided in classified and non-classified. Information will be organised and described in a catalogue available using, for instance, CORBA, RMI, HTTP, or LDAP interfaces. (Mechanisms to access the different sources of information aren't discussed here in this paper). Authors put their contribution in this space thorough and submission agent. In this process authors are invited to describe their contribution (Metadata fill) using terms available in the classification space.

3.3 Classification Space

We will use the existing classification systems that are stored in LDAP technology due to its hierarchical and distributed proprieties. However, other technologies could be used alternatively.

3.3.1 Normalisation

One of the functions of the classification space will be the normalisation. All terms introduced for document description (author given) or user profiles should be checked against the classification space in a pre-coordination phase. The system will look for all classification systems and search exact terms after their concatenation. Equivalent normalised terms will be proposed and if the user agrees they will be considered for the profile. The existence of classification systems in different languages will allow a cross-language SDI service, through the correspondence of same term in several languages. If the user requests in her profile a service in different languages, the translation of each term of her positive/negative profile will be done using the classification system. The automatic translation is possible by using the links available within the descriptions in different languages on these classification systems. Then all normalised terms are available in the classification space in different languages and with respective links.

3.3.2 Communities (Clustering Profiles)

Establishing communities is an important challenge proposed in this paper. The retrieval system should be adapted to treat user profiles as document representatives and provide clustering for community's identification. Clustering will be based on a distance function within an N dimensional space. The similarities measured by the distance will be evaluated based on experience and on the singularity of treated subjects.

Communities are only effective after a human authority decision. Every time a new member arrives, all partners of the community and in their neighbour communities shall be notify through automatic notification service. Feedback from the partners on judging this classification will contribute to a better community definition in a collaborative way.

Central profile is used as community representative and will be chosen as central profile the nearest to the geometric centre of the community. The user communities will be useful information to different organisations (publishers, editors, etc). Central profile can be used to define "special" media services, like personalised newspaper, digital TV and so on.

This community concept can be used as a free approach that can later influence the dynamic creation of information catalogues.

Complex communities can be divided into smaller groups inside a small sub-space. Community's boundaries will be an interesting and difficult problem to solve, where users can belong to one or more communities. All this measures are performed in an N dimensional Euclidean space where profiles are defined by vectors. The dimension of space is defined by classified terms available in document collection.

3.4 Filtering System

The Web interface will run in an interactive way and the electronic mail interface in a batch process. Using these interfaces, users can archive and check personal information, store and configure the frequency of messages and give feedback on documents received. SDI system should handle conveniently users as well as documents.

Regarding documents, filtering can provide three types of services, namely:

- Subject-based (will provide classified service, such as Yahoo);
- Document description based (content SDI system);
- Full document based (content SDI system).

This system can use standard information retrieval techniques in well-defined sub-spaces. The traditional query initiated by the user will be replaced by his profile and the system will perform the respective matching (profile versus document representatives). Only results above a threshold level will be considered. The output shown to the user should be the result from positive less negative profile. The maximum number of messages in the output should be pre-defined and make part of the user profile.

MySDI is an open architecture for SDI personalized services, with the mission of automatic delivering the right information to the user and automatic text classification. A special note should be referred to the MySDI modular structure (for instance, based on Java components and Web technologies) based on different blocks that allow the integration and test following different approaches.

4 A MySDI Prototype: The MyGlobalNews Service

The information society can be characterised by new paradigms in the production and dissemination of information. Those paradigms are raising new challenges to institutions traditionally placed between the producers and the receivers of that information, such as the editors, bookstores, newspapers, radio and TV stations, libraries, etc. In this context, we will concentrate on the digital newspaper paradigm (obviously, the issues discussed here can be easily extrapolated to other situations). The digital newspaper concept is something that is well diffused with an increasing number of sources available all over the Internet. We can classify those services in different categories, for example:

- Copy or partial copy of edition available in other media; this is the case of Diário de Notícias [14], Publico OnLine [15], TVI [16], Information Week On Line [17] or MyCNN [18].
- Edition available only in electronic format; this is the case of Diário Digital [19].
- Compilation of news from other sources available on web; such as IOL [20] or New York Times [20], NewsHound [21], or Sapo [22].

In spite these different approaches, none of the precedent examples provided a personalised service. However we can find some (a real reduced number of examples) personalised digital news at Pointcast [23], Crayon.net [24] and LaTimes [25].

MyGlobalNews is a prototype of our personalised newspaper service that has being developed based on the conceptual MySDI architecture, presented in this paper. MyGlobalNews is a Java agent based application developed on top of the AgentSpace platform.

AgentSpace [3, 26] is a Java mobile agent platform implement on top the Java Virtual Machine and the Objectspace's Voyager product [27]. AgentSpace platform allows the support, development and management of dynamic and distributed application based on the agent paradigm [3, 26, 33]. AgentSpace platform has being developed in the context of the Technical University of Lisbon (IST/INESC) since 1997. For more information, the reader is invited to consult the AgentSpace web site (http://berlin.inesc.pt/agentspace/).

In MyGlobalNews registered users receive the news they want from one or several sources based on parameters chosen by them in their respective profiles. User can chose the periodicity, the format and the subject interested on their messages and also subject that they don't want to see at all. System allows also seeing other user with similar interest, providing user communities to the information producer and broadcast information to standard user communities identified by the system. News in different languages is also available and classified system allows multilanguage cross retrieval through the existence of classified system in different languages (see 3.3.1). Figure 2 depicts an architectural overview of the MyGlobalNews's main components. Its main agents are the following:

- Matching Agent: This agent identifies relevant documents based on the matching of documents and user space representative. Agent use vectorial retrieval model for matching technique.
- Robot Agent: Collects news from the different information sources (IS). Provide submition forms to the authors. Information collected is storage in a predefined format (D), in a hierarchical structure where is fundamental to fill the fields: source of information, title, sub-title, date and author. Provide if necessary classification terms to the producer of information describe their contribution.
- Categorisation Agent: Picks the information available in the database (D) and in automatic way tries to associate with the predefined categories existing in the classified system. System authority is informed about categorization performed and he can change this automatic classification, if necessary. Alerts also system authority about new categories that should be considered.
- System Agent: Deletes old news, notify user about new information relevant. Only system administrator can change parameters on this agent.
- User Agent: Allow the register of users and handle the profile. All information related is storage in user database (U). The users define user parameters on through this agent. Collects also momentary user information needs and pass that needs to matching agent. This agent also represents the results of system (Retrieval, Filtering, Broadcast) to the user. Provide also feedback mechanisms.
- Community Agent: Cluster user profiles and submit for approval to system authority. Results are stored in the database (CS).
- Classification Agent: Agent to load and interact with classification system previous build. Information is storage in classification space database (CS). Provide visual interface of classification system and allow easy change performed only by system authority.
- Index Agent: Agent responsible for the automatic index of documents and provide a short description for user summary messages. Uses terms available in the appropriate classification system.

On the other hand, we have identified the following actors involved in the MyGlobalNews business model:

- End-user (or just "user"): a person that should receive personalised news based on a previously configured user profile.
- Author: person or system that produces information somewhere accessible in the Internet.
- System authority: person responsible by the administration of the classification space.
- System administrator: person responsible by the managing of end-users accounts, user communities, as well as by keeping track global configuration parameters and system statistics.
- System is available on line at <speedy.inesc.pt:8080/login.htm>. At moment, we are still solving same faced problems and make system improvements. Experimental results are reduced yet due to small number of registered users. Soon we intend to launch a new system version, with an intelligent web robot able to search and identify new web information sources and pick in automatic way information from these identified sites.

5 Conclusions and Future Work

We argued that personalisation is a very important requirement for new and emergent SDI services. Consequently, we proposed in this paper a generic and conceptual architecture to design SDI personalised services, which we called MySDI. Based on it, we showed a new approach to deal with these services based on the software agent paradigm. This approach can avoid the proliferation of different SDI systems build from scratch. With this approach, anyone can use or build specialised agent that can be used in different domains and with different requirements. In the future one system can easily be build from this agent space by the integration of different specialised agents. Potential benefits are apparent:

- Shorter application developments due to common guidelines and existing modules, which can be re-used very easily.
- A decrease in development costs due to shortened development time.
- Development costs for different platforms is significant reduced due to the use of Java.
- An increase of wide spread use and user acceptance due to software availability and similar user interfaces.
- A maximum exploitation of all available resources on the client side (local execution) and on the server side (native execution).

Nevertheless, there are still several problems and challenges. One of them is that 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 term. To minimise this problem is important to build user interfaces with easy and efficient dialogues and also use techniques to expand the terms provided by the users.

Personalised SDI brings 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 from users comparing with traditional SDI services, and that fact constitutes a really challenge for the development of personalised services.

Finally, we know the value of SDI services comes usually from their number of registered users. For that, we intend to develop, based on the MySDI architecture, a public and robust version of the MyGlobalNews service. In order that, we have started some preliminary contacts with some important Portuguese newspapers companies as well as with Portuguese "TV Cabo", which is the bigger company, in Portugal, operating the channel and interactive TV. Based on these real world projects we will have the opportunities to improve and validate our research proposals.

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] A. Silva, M. Silva, J. Delgado. AgentSpace: An Implementation of a Next-Generation Mobile Agent System. (Mobile Agents'98) Lecture Notes in Computer Science 1477, Springer Verlag, September 1998.

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

[5] Manber, U., Patel, A. and Robison, J. Experience with Personalization on Yahoo! Communication of ACM, Aug.2000, Vol.43,N8(107-111).

[6] Firefly Web Site: - Jan-2001.<www.firefly.net>

[7] Smyth, B. and Cotter, Paul. A Personalized Television Listings Service. Communication of ACM, August 2000, Vol. 43, N 8 (107-111).

[8] Ferreira, J; Borbinha.L., J.; Jorge,J; Delgado,J.,Nov-1997. Collaborative Filtering for a Community Digital Library using LDAP. Published and presented at 5th. Delos workshop,10-12 in Budapeste.

[9] Goldberg, D., Nichols, D. Oki,B.M. and Terry, D., 1992 Using collaborative Filtering to weave an Information Tapestry. Communication of ACM, December 1992, Vol. 35, N 12 (61-70)

[10] Yan, T.W. and Garcia-Molina, H. Distributed selective dissemination of Information. In Proceedings of the third International Conference on Parallel and Distributed Information Systems,89-98. IEEE Computer.

[11] Lang, K. , 1995. Newsweeder: Learning to filter netnews. Technical Report, School of computer Science, Carnegie Mellon University.

[12] Konstan J.A et al.,1997. GroupLens: Applying Collaborative Filtering to Usenet News. Communication of ACM, March 1997, Vol. 40, N 3 (77-87).

[13] Kautz, H., Selman, B. and Shah M. ReferralWeb: Combining Social Networks and Collaborative Filtering. Communication of ACM, March 1997, Vol. 40, N 3 (63-65).

[14] Diário de Notícias, Jan-2001. <www.dn.pt/ >

[15] Jornal Público, Jan-2001. < www.publico.pt/ >

[16] Televisão Independente. TVI OnLine, January 2001. < http://www.tvi.pt/ >

[17] Information Week. Information Week On Line, January 2001. <http://www.informationweek.com>

[18] MyCNN. MyCNN On Line, October 2000. < http://www.mycnn.com/ >

[19] Diário Digital. Diario.Digital, January 2001. < http://www.diariodigital.pt/ >

[20] IOL. Iol.pt, June 2000. < http://www.iol.pt/ >

[21] New York Times. The New York Times on the Web, January 2000. < http://www.nytimes.com/ >

[21] NewsHound. Welcome to NewsHound – Your Best Friend for Information, October 2000. <http://www.hotcoco.com/newshound/>

[22] Sapo. Sapo - Portugal Online!, October 2000.
< http://www.sapo.pt/ >

[23] Pointcast.–Jan-2001.<http://www.infogate.com>

[24] Crayon.net-Jan-2001.<http://www.crayon.net>

[25] LaTimes-Jan-2001.<http://www.latimes.com>

[26] AgentSpace, AgentSpace – A Next-Generation to a Mobile Agent System, October 2000. <http://berlin.inesc.pt/agentspace/>

[27] ObjectSpace Web Site, Voyager, December 1999.<http://www.objectspace.com/products/voyager/>

[28] J. Ferreira, 1999. Arquitectura para um serviço de disseminação selectiva de informação – JET.

[29] ClixSmart-Jan-2001 www.changingworlds.com.

[30] ZyFilter–Jan-2001. www.zylab.nl/p5/cases.

[31] Agentware i3 - January 2001.
http://www.agentware.com/main/server/index.html

[32] KeMedia - January 2001.
http://www.ke.com.au/ke/products/media.html.

[33] A. Silva, 1999,Agentes de Software na Internet – A próxima Geração de Aplicações para a Internet, Centro Atlântico, Lisboa.