EUROgatherer: a Personalised Gathering and Delivery Service on the Web

Giuseppe Amato, Umberto Straccia and Costantino Thanos
I.E.I. - C.N.R.
Via S. Maria, 56, I-56126 Pisa, ITALY
{amato,straccia,thanos}@iei.pi.cnr.it

ABSTRACT

In this paper an overview of EUROgatherer will be presented. It is an advanced personalised information gathering and delivery service on the Web.

EUROgatherer satisfies user's information needs in several ways: through (i) ad-hoc search for satisfying users' short term information needs; (ii) content-based information filtering; and (iii) time scheduled search, both for satisfying users' long term information needs. Techniques of relevance feedback and a complex user profile schema are used to learn and store the preferences of subscribed users.

Keywords: Information Search and Retrieval, Information filtering,, Distributed systems, User profiles and alert services, Web-based services

1. INTRODUCTION

The Web, and consequently the information contained in it, is growing rapidly. Every day a huge amount of "new" information is electronically published. This has made it increasingly difficult for individuals to control and effectively seek for information among the potentially infinite number of information sources (Web pages, online databases, News Groups, Publishers, Digital Libraries, News Agencies, etc.) available on the internet. Ironically, just as more and more information is getting on-line, it is getting increasingly difficult to find relevant information in a reasonable amount of time, unless one knows exactly *what* to get, *where* to get it from and *how* to get it.

Typical information sources on the Internet, like Search Engines, Digital Libraries and online database (e.g., Altavista [1], ACMDL [1], Medline [11], NCSTRL [12], just to mention some), provide a search and retrieval service to the web community at large.

A common characteristic of most of the traditional retrieval services is that *they do not provide any personalised support to individual users*, or poorly support it. Indeed, they are oriented towards a generic user. They answer queries crudely rather than, for instance, learning the long-term requirements idiosyncratic to a specific user.

Providing personalised information search and delivering services, as additional services to the uniform and generic information search offered today, is likely to be the first step to make *relevant* information available to people in the appropriate *form*, *amount* and level of *detail*, at the *right time* through the *right medium*, and with *minimal* user effort [4,13].

The EUROgatherer service, which is the result of a EU Telematics Information Engineering Programme founded project [6][1], was designed to cope with the requirement of personalised search according to the most wide meaning of it. Indeed, it is an advanced personalised information gathering and delivery service on the Web.

The topic of this paper is to give both a description of EUROgatherer in terms of provided functionalities as wells as in terms of it's architecture.

We will proceed as follows. Section 2 introduces those concepts which have to be taken into account in a quite general information seek scenario; Section 3 gives an overall description of the system; Section 4 describes the user's profile model that has been used; Section 5 makes a brief comparison with other existing services; Section 6 concludes.

2. A GENERIC INFORMATION SEEK SCENARIO

From a quite general point of view, we may depict the scenario of the retrieval and delivery of relevant information as in Figure 1. We can distinguish two main actors and three main tasks in it. The actors are the *user information needs* and the *information sources* . The tasks are *fetching information from the Web sources*, *selecting which is relevant to the user* and *delivering it*. These will be described in the following sections.

User information needs

With user information needs we mean which information a user is interested in. An example of user information needs may be

- "I'm interested in news concerning the latest trend about stock quotes of High-Tech companies."

It is worth noting that a user information need may be a *short term* user information need or a *long term* user information need, depending on the interets of a user. In the former case we refer to an ad-hoc, occasional user information need, e.g. an economist which would like to go to the sea at the weekend is looking for the weather forecast. Whereas, in the latter case we refer to an user information need which is of interest during a relevant time period, e.g. an economist would like to be informed every morning about the economic trend. We will call any long-term information need of a user a *topic of*

[1] See Appendix for list of partecipants.

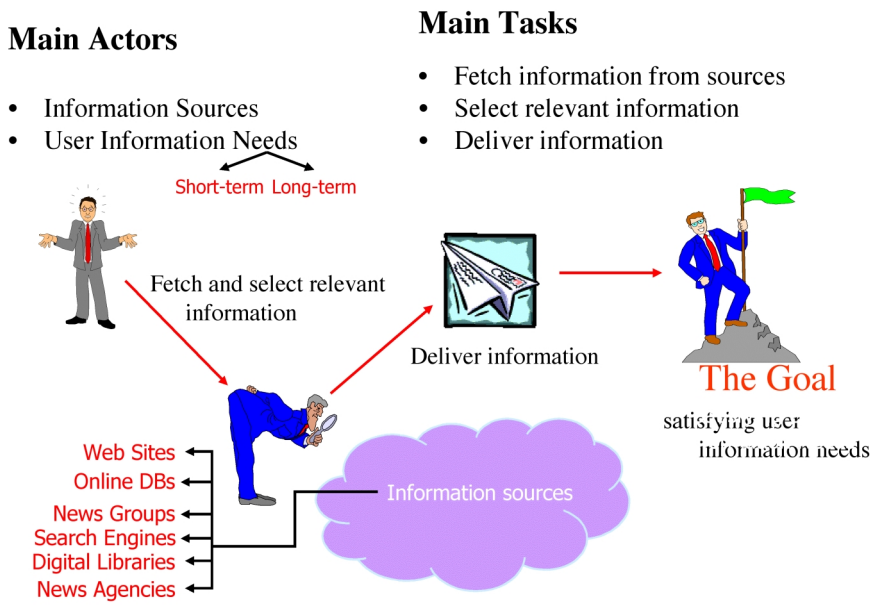

Figure 1: A general information seek scenario

interest. It is easy to verify that in fact our daily information seek process involves both short-term (occasional) interests as well as long-term interests. Of course, whether an information need is a short-term or a long-term interest depends on the user.

Information sources

With information sources we mean all the heterogeneous digital information providers distributed over the internet, which make available any kind of information which might be of interest to internet users. Examples of information sources are web sites, online databases, news groups, news agencies, search engines, digital libraries, etc. Essentially, they differ in *what kind of information they provide, what services they provide and which users they address to.*

Fetching, selecting and delivering relevant information

In order to satisfy the user, three tasks have to be executed: the first two tasks, typically the hardest ones, are (i) fetch information from a heterogeneous set of Web sources, (ii) select from the fetched information the one which is thought to be of interest to the users; and, once the information has been collected, (iii) deliver it to the users, according to their preferences. Delivery preferences should take into account both the *delivery medium* as well as *the delivery time.* Examples of delivery medium may be web (this is the usual case for which most of us are familiar with), e-mail, phone, fax (e.g., a user wants to receive stock quotes by phone) - to be delivered at a certain time, time interval or as soon as the information is made available, or surface mail (e.g., a user wants to receive the proceedings of a conference by surface mail).

3. AN OVERVIEW OF THE SYSTEM

In order to cope with all the parameters listed above, and since some needed services are already available on the Internet (in particular, some partners of the project consortium already provide some specific services), rather than building a system from scratch providing all the functionalities, the EUROgatherer service was designed as an *open architecture*: indeed, it is a federation of autonomous services (see Figure 2) distributed over the internet. Each service provides a set of functionalities. It may co-operate with other services and may work independently. The interoperability is guaranteed by relying mainly on a common

- inter-service communication protocol based on HTTP [17]; and
- a common user's profile schema. A user profile essentially describes user's information needs and user's delivery preferences.

The federation of services that form the EUROgatherer system offer the followings:

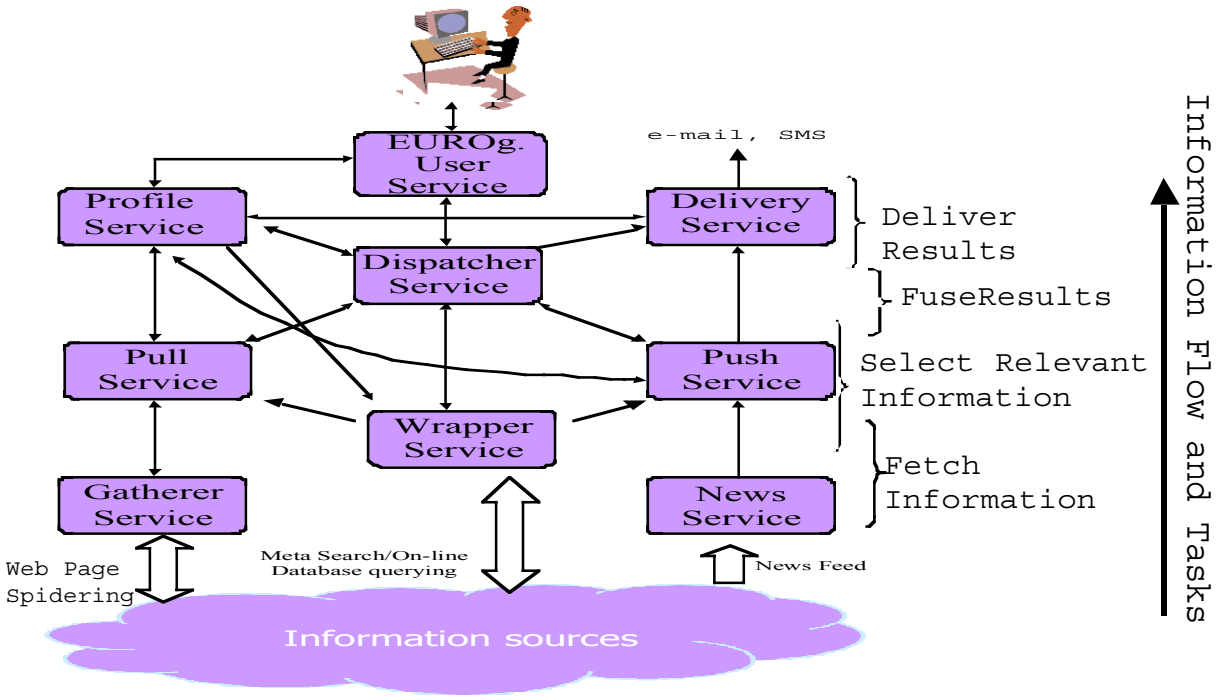

Figure 2: A federated view

(i) advanced user profiling with the use of relevance feedback for automatically learn the interests of the users;

(ii) ad-hoc querying, for satisfying a short term information need;

(iii) topic-on demand, which allows users, profile during a user session, to retrieve documents pertaining to a topic of interest defined in their profile;

(iv) scheduled topics, to deliver relevant document autonomously at a specific time (time interval or as soon as the systems gathers information relevant to a topic) according to the user preferences;

(v) transparent access to multiple information sources available over the Internet, like the Web, online databases and News from News Agencies;

(vi) customised delivery modalities, defined in terms of the delivery medium to be used (Web, e-mail, SMS) and time (see point (iv)).

Currently, the EUROgatherer service architecture is made out by the following services distributed over Europe. In Figure 2, from bottom up, there are three services, i.e., the Gatherer Service (GS), the News Service (NS, and the Wrapper Service (WS), which are responsible for fetching documents from the Web. Each of them is specialised for a certain information source.

(i) **Gatherer Service (GS):** This service is responsible for collecting HTML pages by spidering a set of predefined sites on the internet. The collected pages are sent to the Pull Service and the Push Service for indexing.

(ii) **News Service (NS):** This service is responsible for managing the incoming news transmitted by News Agencies. The news are sent to the Push service for indexing.

(iii) **Wrapper Service (WS):** This service is responsible for retrieving HTML pages by querying several online Web databases and search engines [5]. The collected pages are sent to the Pull service and the Push Service for indexing. The WS uses the topics of interest defined by the users for automatically select good candidates of information sources to be queried (see [9]).

Therefore, through the GS, NS and the WS Eurogatherer is able to cover almost all information sources available on the Web. The collected Web pages are then indexed and classified by the Pull Service and the Push Service, which are indeed search engines. Both, given a topic of interest, select from their own document base those documents which are thought to be of interest to the user. But, they a work in two complementary modes, i.e., pull modality and push modality, respectively:

(iv) **Pull Service (PullS):** This is a typical search engine which given a query (expressed by a topic of interest), *matches the query against the set of documents of its internal database.* The PullS is specialised for managing documents spidered from the Web.

(v) **Push Service (PushS):** The PushS acts as a filtering engine and is specialised for managing a continuous flow of incoming documents, as it is the case for news. The PushS, *matches an incoming document from the NS against all topics of interest.* If a document is considered being relevant to a topic (the matching is above a certain threshold), the document is assigned to the topic. Once, a query (topic of interest) is submitted to the PushS, the previous assigned documents are returned.

In summary, the PullS matches a topic against documents, whereas the PushS matches a document against topics (see also [4]). This has an important consequence. In fact, the PushS provides a functionality that cannot be provided by the PullS. In fact, since the PushS matches a document against the topics of interest of the users, the PushS is able to alert a user about the presence of a relevant document as soon as the document is made available.

(vi) **Profile Service (PS):** This service is responsible for the management (storage, maintenance and retrieval) of the users' profiles.

(vii) **Delivery Service (DelS):** This service is responsible for notification and delivering of the results to the users according to their delivery preferences. Currently, the notification modes are e-mail and SMS cellular phone messages.

Finally, there are the EUROgatherer User Service and the Dispatcher Service.

(viii) **EUROgatherer User Service (EGUS):** This service is the main entry point to EUROgatherer. Through an usual Web browser, the users can subscribe to the EUROgatherer services, can log in and can interact wit the service. The functionalities offered to each user can be broadly classified as

User profile management: these functionalities allow the users to create, modify and inspect their personal profiles stored in the PS.

Query: the user may submit queries corresponding to short-term information needs (called ad-hoc query) and inspect their topics of interest, i.e., long-term information needs (called topic on-demand).

User Feedback: when results are shown to the users, they can express their feedback, i.e., specifying which retrieved document is relevant and which is not. This information is used by other components of the service architecture to learn a better description of the users' information needs.

Once, an ad-hoc query or a topic on-demand request is submitted by the user through the interface, the request is sent to the Dispatcher Service.

(ix) **Dispatcher Service (DispS):** This service is responsible for dispatching information requests to the search services, mainly PullS and the PushS. The two lists of ranked documents produced by the PullS and the PushS are submitted to a data fusion process (according to [8]). Once an unique result list has been generated, it is sent either to the DelS or to the EGUS. Additionally, the DispS has a scheduler which allows at user specific time to submit a topic of interest to the PullS and PushS (scheduled-topic).

It is worth noting that there is no limitation in the number of the typology of services that can be added. In particular, several DelSs, PullSs, PushSs, WSs, GSs, NSs may be further considered, as well as new service of different kind may be added (e.g., multilingual access).

4. USER PROFILE MODEL

A prerequisite for developing a personalised search and delivery service is to rely on *user profiles*, i.e., a representation of the topics of interests of any individual user and his delivery preferences.

The model that we defined to represent user profiles relies on the P3P data schema [14], but extends it significantly. In fact, it adopts all data sets and data types proposed by the P3P specification, but in addition a new group of data sets and data types to represent information needed by the EUROgatherer service architecture have been defined. The overall profile schema is sketched out in Figure 3.

The P3P data sets are used to represent the user personal information such as its name, its address, etc.

The EUROgatherer specific data sets represent information concerning:

(i) delivery preferences (Delivery data set)
(ii) the description of user's information need (Content data set)
(iii) The sources (e.g. web sites) where documents should be (or should not be) searched into (Source data set)
(iv) Relevance feedback actions provided by the user (RF data set).

Topics of interest are represented in the profile by means of the EGTopic data set. Each user profile may contain a set of topics of interest. Each topic has a name and a unique identifier (different users may have topics with the same name; however the topic identifier is unique). For each topic, preferences corresponding to (i), (ii), (iii), and (iv) can be expressed.

The *Delivery data set* stores information about the delivery modality to be used when a certain document is judged to be relevant to the corresponding topic. In particular it contains preferences about the device to be used (i.e. e-mail, SMS), and the destination (e.g. the phone number, the e-mail address). The Deliver data set contains also information about the time when the user wants to be notified (e.g. every morning at 9.00 am, or as soon as possible). In addition some information to keep track

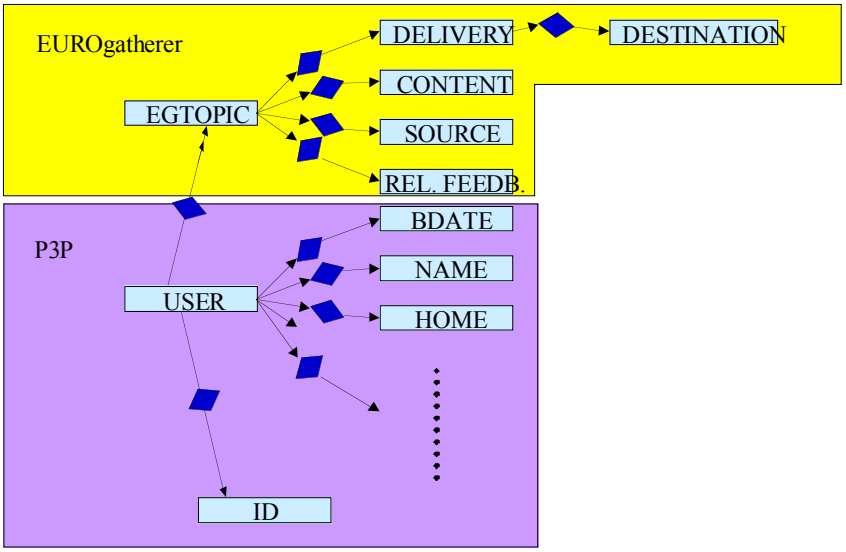

Figure 3: User's profile schema

of which documents have been seen by the user is provided (Res-Management data set).

The *Content data set* contains the description of the information content to be of interest. A topic may be described by a predefined category chosen from the EUROghaterer category hierarchy and a set of user specified keywords used to refine the category. For instance, the topic of interest of a user indicated by it with "Formula one" can be defined using the category "sport" and adding keywords like "Formula one, Ferrari, McLaren, etc.". It should be noted that the EUROgatherer service architecture currently relies on a predefined and common category hierarchy, where to each category an identifier is associated. The system may be easily extended working without the common category hierarchy.

The *Source data set* allows a user to specify two optional lists of URLs. The first one is the allow-list that contains the list of URL that should be used to search for interesting documents for the corresponding topic. The second one is the deny-list that contains the list of URL that should not be considered to search for documents relevant to the corresponding topic. For instance, for the topic "Java" one may want to receive documents originated in the SUN Web site and to reject documents originated in the Microsoft Web site.

The *RF data set* records for each topic of interest available user feedback data. It contains an attribute "Action" which maintains all actions performed by the user on a proposed document and a "Keywords" attribute that contains weighted keywords to be added to the ones specified by the user. EUROgatherer services, analysing the list of user actions, may update the "Keywords" attribute to refine the user profile.

5. RELATED SERVICES

To date, several search engines offers some form of personalised search. In this section we will describe the kind of personalisation offered in world wide and well known search engines like Altavista [1], Yahoo [18], Excite [7] and Infoseek [10]. Certainly, there are many others, but we believe that the above are good candidates in showing what is going on about personalised content-based search.

Curiously, apart the differences concerning the different ways a user can personalise the layout of the to be displayed information, all the above search engines have a common characteristics:

1. a user may define different topics of interest
2. each topic is defined in terms of a category selected among the ones of the categorisation scheme of the search engine
3. for each topic, the result-on demand functionality is provided, i.e., by clicking on the topic name, a query is build from the topic and submitted to the search engine and the (pulled) result is displayed (this corresponds to the activity of the Pull Service in EUROgatherer).

It is immediate to verify that w.r.t. EUROgatherer

1. typically, no additional keywords are allowed to be specified for further topic refinement
2. no relevance feedback is managed (mark documents as "don't show me it next time", "relevant", "not relevant", "seen")
3. no scheduled delivery of documents relevant to a topic is provided. An exception is Yahoo which allows to specify some refresh rate (15, 20, 25, 30, … Min) of the Web page showing the content of a topic.
4. no alerting service is provided (Push Service's functionality). As a consequence, the "delivery relevant information as soon as it has been gathered" functionality cannot be provided. An exception is Altavista which provides a week alerting service of the type "alert when you've got mail or your favourite stock moves" (a PC program should run on the client, for a continuos connection to the server)

5. information delivery is based on the Web browser only.

6. CONCLUSIONS

Undoubtedly, to rely on a personalised gathering and delivering service is an emerging need of most user's seeking on the Internet for information. We have presented the EUROgatherer service architecture which would help such users. Indeed the EUROgatherer service architecture (i) allows advanced user profiling; (ii) accesses multiple information sources available over the Internet; (iii) has customised deliver modalities defined in terms of the delivery medium (Web, e-mail, SMS) and defined in terms of the delivery time (at a specific time, timed interval, as soon as possible).

Moreover, besides offering all the above features, from a architectural point of view, we point out that the EUROgatherer service architecture is made out of a pool of standalone, and autonomous services, each offering specific functionalities. This allows it to be easily extended with other services, compliant to the common user profile model and the inter-communication protocol based on http making it a through appealing.

Though EUROgatherer, as outcome of an EU project, is a prototype, it offers several interesting functionalities. Most of them and the experience gained through the project have already be transferred by the commercial partners of the project into their products.

7. APPENDIX

The following is the complete list of the partners involved in the EUROgatherer project:

1. I.E.I. - C.N.R. (Istituto di Elaborazione della Informazione, Consiglio Nazionale delle Ricerche). Pisa, ITALY. Research Institute.
2. Italia OnLine. Pisa, ITALY. Internet Service Provider.
3. XEROX Research Center Europe. Grenoble, France.
4. Eurospider Information Technology AG. Zurich, Switzerland. Commercialise an in home build information retrieval system and Internet services
5. XarXa Cinet SL. Barcelona, Spain. Internet Service Provider.
6. University of Dortmund. Dortmund, Germany.
7. Dublin City University. Dublin, Ireland.

8. REFERENCES

1. Amato, G. and Straccia, U. User Profile Modeling and its Application to Digital Libraries. Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries, LNCS 1696, pp. 184-197, 1999.
2. ACM Digital Library Home Page: http://www.acm.org/dl.
3. Altavista Search Engine Home Page: http://www.altavista.com
4. Nicholas J. Belkin and W. Bruce Croft. Information Filtering and Information Retrieval: Two Sides of the Same Coin? In Communication of the ACM, December 1992, Vol 35, No. 12, pp. 29-38.
5. Boris Chidlovskii, Uwe M. Borghoff and Pierre-Yves Chevalier. Towards Sophisticated Wrapping of Web-based Information Repositories. In Proc. 5 th Int'l RIAO Conf., Montreal, Canada, June 25-27, 1997, pp. 123-135.
6. EUROgatherer. Telematics Information Engineering Project Number 8011". Home Page: http://pc-erato2.iei.pi.cnr.it/eurogatherer/. Access to the Service is also provided.
7. Excite Search Engine Home Page: http://www.excite.com
8. J. Lee. Analyses of Multiple Evidence Combination. In *Proceedings of the 20th Annual International ACM/SIGIR Conference*, 267 - 276, Philadelphia, USA, 1997.
9. Norbert Fuhr. A decision-theoretic approach to database selection in networked IR. ACM Transactions on Information Systems, 17, 1997.
10. Infoseek Search Engine Home Page: http://www.infoseek.com
11. Medline Home Page: http://igm.nlm.nih.gov.
12. Networked Computer Science Technical Reference Library Home Page: http://www.ncstrl.org.
13. Peter W. Foltz and Susan T. Dumais. Personalized Information Delivery: An Analysis of Information Filtering Methods. In Communication of the ACM, December 1992, Vol 35, No. 12, pp. 51-60.
14. P3P Home Page: http://www.w3.org/P3P/
15. Gerard Salton and J. Michael McGill. Introduction to Modern Information Retrieval. Addison Wesley Publ. Co., Reading, Massachussetts, 1989
16. SSL description: http://home.netscape.com/products/security/ssl/protocol.html
17. HTTP protocol: http://www.w3.org/Protocols/
18. Yahoo Search Engine Home Page: http://www.yahoo.com