Alerting Services in a Digital Library Environment

Annika Hinze
Institute of Computer Science
Freie Universität Berlin, Germany
hinze@inf.fu-berlin.de

Abstract

The classical paradigm of finding information in the WWW by initiating retrieval and browsing becomes more and more ineffective. Other techniques have to be considered. Automatic delivery of contents to the user according to their needs and filtered by her profile of interests is required. Current implementations of such *Alerting Services* at content providers side have several drawbacks. In my research project I evaluate methods and techniques for Alerting Services with special respect to the area of digital libraries. I intend to provide a framework that supports design decisions in building alerting services depending on the infrastructure and desired system parameters.

1   Introduction

Imagine one morning you just arrive at your office and switch on your computer to have a look at the recent news in your special field of research. Little pictures for each topic tell you that some interesting documents arrived. Behind one icon you find for instance the new announcements for congresses, behind another one some interesting papers, tables of contents of your favorite scientific journals, announcement of new books in your field.

This description is not completely science fiction. There already are techniques to inform users of interesting events instead of forcing them to actively search for the information they need. Since the quantity of scientific publications doubles every 10 - 15 years [14] there is urgent need for the use of alerting systems. Such systems keep the users informed only about the document and events they are interested in. A variety of techniques and systems exists but every software developer seems to have to find their own solution (e.g. systems base on simple mailing-lists, CDF Channel, Netcaster Channel for a more a detailed survey see Section 2.1).

Personalization in Alerting Services ensures that each client only gets notified about events they are interested in. The interest of a client is defined in a profile. The service matches events (e.g. new documents) against the profiles and informs the interested users.

Therefore the subject of Alerting Services is also closely related to several areas such as events services and profile definition in Information Retrieval (IR).

In my research project I evaluate existing techniques for notification and formats as well as methods for information delivery. The digital library project at the Free University of Berlin, DARWIN [20], will serve as a testbed for the Alerting Service evaluation implementations. This work will be done in close relation to the sub-project Alerting Services of the BMBF project Global Info [23]. This close contact to publishing houses will be used to take the information providers demands into consideration.

The following set of problems has been specified:

1. It exists *no common language* for the delivery of notifications, various standards for data delivery exist but none of them is suitable for the application space of digital libraries.

2. There is a variety of techniques and models for different application levels, that have *no common set of descriptive terms and definitions* and are therefore incomparable. This is easy to see in the term Alerting Service aka notification service aka selective dissemination on information (SDI) and so on.

3. Several implementations for data delivery services exist. For the implementation of an Alerting Service the programmer can choose between a variety of techniques but there is no *common set of criteria* to valuate existing implementations and techniques and to support these design decisions.

The remainder of this paper is structured as follows: Section 2 provides an overview of related research topics and existing dissemination technology. Section 3 outlines the derived open question and my research plan and the state of my work so far. I conclude in Section 4 with the direction for future work and discussion.

2 State of the Art and Related Work

The application domain I consider in my project encloses such applications as stock tickers, common weather report channel, notification about new items in digital libraries or traveler information systems. The range of services for data dissemination reaches from awareness services to event based infrastructure for distributed control.

An awareness service ensures the actuality of repository replicas by sending updates (e.g. a library catalogue is updated according to new or deleted books in the publishers repository). Alerting Services for digital libraries have different requirements, e.g. often they deliver notifications about events on provider side, not the data itself. Additionally alerting service has greater number of clients to serve, personalization is highly required.

Due to the given testbed in the DARWIN project, I will focus on the Digital library environment.

2.1 Bibliographic Alerting Services

Some non profit services like Ariadne [17] or TEL-HP [25] offer keyword-based profiling: clients are notified as soon as documents matching the profile arrived. Profiling services have also been offered by bibliographic databases, some with sophisticated profile definition languages, like CompuScience [18]. Some publishing houses offer Alerting Services to push new content to the subscribers. This is offered for example by Springer with Springer Link Alert [28] and Elsevier with Contents Direct [21]. Currently users can only subscribe to a selection of journals and will get the Table of Contents (TOC) with embedded links to full-texts by email. Elsevier additionally offers a notification service based on web channels (ScienceChannel [22]). The user can choose a topic from a keyword list and subscribe to the journals controlled by the keyword. Notification about changed channel content comes by email or desktop icon. With SwetScan Swets offers an awareness service for libraries [29]. It provides the electronic delivery of a catalog update. Subscribed libraries can define their interest profile with bibliographic data. The profiles in the ISI Alerting Service [24] offered by the Institute of Scientific Information are defined in a similar way.
This large number of different offers has some unpleasant side effects:

- The user probably misses some important information providers, because they have to register at each providers service individually.

- The situation forces the user to train oneself with the different services of the providers. One must define special interest in the way required by the provider's interface.

- Most of the tools just send many email full of irrelevant information for the user. The tools do not use modern technical possibilities. Nobody wants to deal with an endless stream of email.

- The available notification functions are bound to the offer of individual publishing houses and distributors and information from several providers is not combined.

- Personalization is very rare. If implemented it forces the user to give personal data (profile) to many different persons or companies.

- In the mentioned services profiles cannot be matched against the full-text documents. This would allow more sophisticated profiles and more precise matches on the document sets.

This problems lead to the installation of a mediating Alerting Service. There the user can subscribe and get access to the resources, the various notification schemes of the different provider services are unified for the user and the clients privacy is protected. The overall architecture of such an Alerting Service is shown in Figure 1.

There one can find cooperating and non-cooperating providers. Cooperating providers notify the Alerting Service of the occurrence of events at their repository,
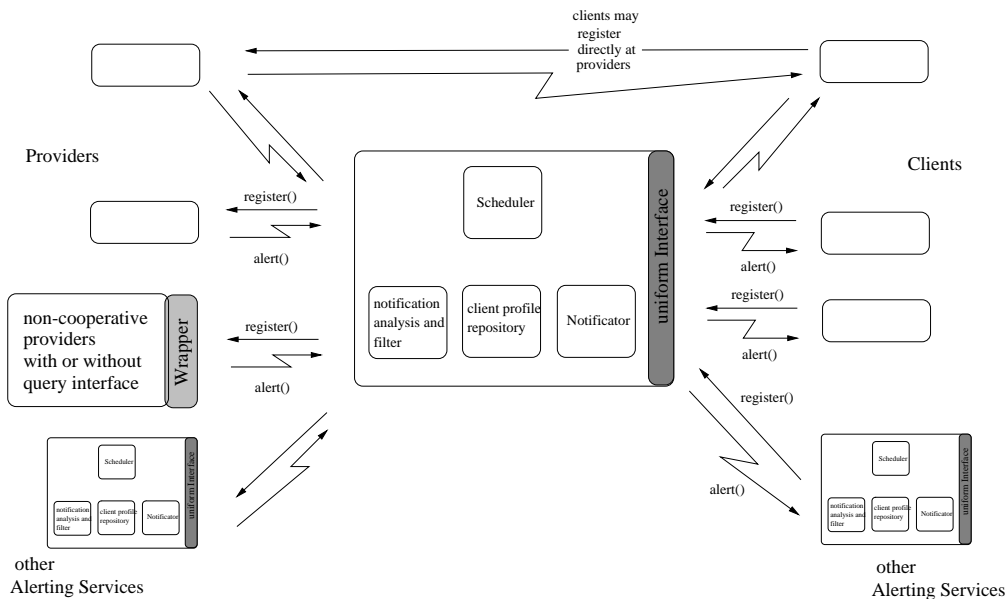
Figure 1: Architecture of a mediating Alerting Service

non-cooperating providers do not offer any notification and therefore wrappers have to be implemented to query the providers repository.

It remains the choice of the appropriate technique for the notification between provider and Alerting Service and between the service and the users. Another interesting problem is the consideration of non-cooperative providers, i.e. publishers that do not maintain an Alerting Service. The service could actively search for information according to the user profiles, e.g. by analyzing page structure with external programs (e.g. HyperNavigator [6]), via search interface to the providers database or by use of agent technology [8].

Conceivable are mobile autonomous agents, that search the different providers for new documents, or inform about available documents to certain topics (similar approaches are found in the Profile Project [16], Amalthaea [12] and the Next Generation Digital Library Project [13]).

The transport of the user profiles to the documents at provider's side is expected to be faster then vice versa. In this way the matching of the document against the profile can happen on the provider side. Unneeded data transport will be avoided. The concept of an intelligent mobile agent as an information gatherer is well known in the context of Information Retrieval. Here the primary task for agents is the retrieval of the referenced documents.

To evaluate the implementation-techniques it is necessary to define first the special constraints in the digital library environment. An example of such a constraint is the high number of clients with intersecting profiles that has to be considered in profile checking.

2.2 Evaluation of Delivery Models

Several systems for evaluation of different delivery models and profiling schemes have been designed recently.

The Salamander system [10] is a wide-area network data dissemination substrate that supports push-based applications. It provides a variety of delivery semantics from basic data delivery up to collaborative group communication. One part of the architecture maintains a channel subscription service. The publisher provides data objects with text-based attribute lists. These attribute lists are matched by Salamander against persistent queries defined as profiles by the subscribers.

SIFT [15], the Stanford Information Filtering Tool allows users to subscribe to an information dissemination service by submitting their profiles. The documents transmitted by SIFT are stored in a central repository. So publishers have to forward their full-text documents to the system. This system is not suitable for publishers of electronic documents such as journals. Additionally non-cooperative providers are not supported within this system.

DBIS (Dissemination based Information System) proposes the infrastructure for notification systems. It gives a general classification of architectures of dissemination-based systems [7]. An implementation is Broadcast Disks [1], a system for broadcasting documents via satellite.

Other systems to mention are Yeast [9] - an event/action system- and Siena [3] - an event notification service for wide area networks.
The systems mentioned above have been used to exploit delivery methods and infrastructures for notification systems. They implement various delivery models such as pull, push or adaptive methods, but the evaluation is restricted to the method and does not consider the underlying techniques or design restrictions due to a given infrastructure.

2.3 Technologies for Notification

The following introduces some of the techniques that can be used to implement an Alerting Service.

The *Corba Event Service* [19] allows applications to communicate with one another no matter where they are located or who has designed them. The object request broker (ORB) is the middle-ware that establishes the client-server relationships between objects. Using an ORB, a client can transparently invoke on a method of a server object, which can be on the same machine or across a network. The event service decouples this process and allows several servers to communicate asynchronously with the clients. The suppliers may push or pull their objects.

Active databases could also be used as a platform to implement an Alerting Service. The main drawback of current active database technology is that event-triggered actions are either performed inside the database management system itself or as an invocation on a registered client software. In the digital library the information provider holds all information concerning the documents while the user profiles are kept elsewhere. Matching newly entered documents with user profiles will be hard or impossible.

Internet Channels. CDF is an open specification for the definition of web-channels. It has been submitted as proposal [4] to W3C define a open industry standard for data definition of content to be pushed across the Internet.The CDF technology does not explicitly support the personalization of channels, a channel always supports only one provider each. The push-technology has been implemented as pseudo-push, the client-software performs a scheduled pull. For a short valuation see Section 3.

So far there has been no comparing evaluation of existing technologies that could serve as platform for Alerting Services. A first step has been done in our working-group by comparing channel technologies [5].

2.4 Event Notification Systems

An Alerting Service can be build on top of a low-level event notification service as used for distributed control mechanisms or device monitoring. Conceivable techniques are CORBA Event and Notification Service, Java Distributed Events and Messaging Service and others. The underlying model of these services differs from the conditions found in a digital library environment. As an example the events considered in the models belong to objects of interest, that have limited numbers of states. An event can then be defined as a state transition. In the context of digital libraries the events are state transitions of the repository that contains the objects of interest (e.g. the journals and articles).

One part of my work is to develop a clear definition of the similarities and inequalities of these models in order to make them comparable.

2.5 Languages and Standards

In the area of *digital libraries* nothing can be found so far on standards for Alerting Services. Important standards like Z39.50 [11] do not even mention this type of service.

For *low level methods* (as RMI or distributed control) several protocols have been proposed (such as General Event Notification Architecture (GENA) Base and Event Notification Protocol (ENP)). Due to their restricted expressiveness most of them are not suitable in the digital library context.

The various approaches can be found in standardization for *Internet services*: XML can be seen as a prime candidate for a notification delivery language (but there are several drawbacks to consider, for a critical evaluation see [2]). P3P [27] (based on XML) and OPS [26] are protocols for support of simple personalization and filtering of services.

The discussion about the appropriate language and protocol for notification via Internet has just begun. It is necessary to contribute with experiences in an application field to influence the upcoming decisions. A goal is to develop demands and proposal for a common notification protocol and language based on my research and the work in cooperation with publishing houses at the Global Info project.

For the choice of the profile definition language we will profit by experiences in Information Retrieval. Contrary to Retrieval in Alerting Services the user defines long-term information needs and the document set will change constantly. In a first approach the following set of possible profile features has been defined. We distinguish four types of profiles: (a) a set of identifiers for documents, journals in particular, (b) a list of keywords which can be either selected arbitrarily or from a thesaurus given by the alerting service, (c) a set of predefined groups of objects, such as scientific subjects, and (d) a query in a full-fledged retrieval language, that could even support relevance feedback or documents as queries (find all documents like this one). Additional data can be specified in the profile: document parts, the (keyword)profile should be matched against (e.g. full-text, abstract, metadata); time (scheduled or on event), way (e.g. e-mail, desctop-icon) and content of notification (e.g. hyperlink to full-text, abstract or complete document).

The profile language has to cover these features, they also give a measure for the power of the language.

3 Plan and State of my Work

The research for my research project can be divided in five components:

1. a) A survey of existing notification services, techniques, architectures and models.
   b) Definition of special requirements in the digital library environment.

2. Classification of the techniques for implementing Alerting Services in order to find criteria that can be used to support the design decisions. A classification consists of a number of terms that do not provide a judgment, whereas criteria give a value of quality.

   Terms of classification are: possible delivery methods (push, pull), synchrony or asynchrony of the notification, location of profile data, state-full vs state-less stores and clients.

   Candidate terms for criteria are: profile definition possibility: from simple keyword profile up to full fledged query language,required bandwidth, filter possibilities, performance, scalability in the number of documents and participants.

3. Comparative implementation of Alerting Services based on different technologies. Techniques to implement and to evaluate are such as CDF, CORBA, mobile agents.

4. Evaluation of the implemented techniques: Measures based on the found criteria will be defined and a testbed with a defined set of documents and measurable test condition has to be developed. The implemented services will be included in the DARWIN project and can therefore be tested in an existing environment with real data.

5. Valuation of the results. Dependent on the results a design framework will be developed. That framework will support developers to choose the appropriate technique for the implementing of an Alerting Services under given constraints. If none of the existing techniques fulfills the requirements of a digital library environment a possible recombination of the technical structures in a new way will be analyzed. This valuation will influence the proposal for an notification language and protocol.

I started the research one year ago with the literature research in the area of Alerting Services, it intended to explore the existing technologies and implemented systems. At the same time the DARWIN project started and first contacts to publishing houses (within the Global Info project) have been established. My project is planed to be finished within the next three years.

My work has now moved forward to step two. In order to unify the various methods and techniques found in the literature research I am currently working on common model for notification services. For the items and methods used in the diverse models several intersecting notations can be found. These notations and definitions have to be brought into one cooperation model. Often the existing models describe various abstraction levels of the same technique, but they have not been unified so far.

Parallel to the theoretical research the implementational part has started. A notification service using the Channel Definition Formate has been implemented for the DARWIN project and I am currently working at an additional personalization feature implemented as Java servlet.

The results of the valuation of the CDF-technique are:

1. CDF supports only cooperative providers (that implement their own notification service)

2. Only one provider per channel is possible, otherwise a mediator is needed.

3. No personalization/filtering is possible, because CDF does not support the definition of additional keywords or metadata for the channel-items.

4. Without additional administration old channel-content remains, and therefore the notification contains always the whole content of the data base (incremental growth).

5. The (desktop) notification gives no information about the number or importance of an event.

6. Position dependent access: the desktop notification is browser-dependent and works therefore only for the workstation the user has done the subscription. The email notifications gives wider access, useful only with non-personalized channels.

7. CDF provides only a small number of tags, there is no way of extending the notification or the channel with additional information (e.g. for filtering or additional metadata).

Simple CDF Channels are easy to implement for providers but offers only a very restricted range of possibilities for more sophisticated use (e.g. personal profiles). Since it exists only one implementation of client software (in Internet Explorer 4.0) the technique is hard to evaluate without interfering with the implementational realities. Therefore a next step will be the implementation of an independent CDF Client.

4   Conclusion

The results of my research project will provide support for the design of an alerting service for Internet applications with special emphasis on digital libraries.

In the area of notification services several non-cooperating models (mostly for event notification handling) already exist. The contribution of my work is a common model for notification services. It gives the opportunity to compare the various techniques and to classify them.

There are already criteria and classifications for Internet based services such as scalability and delivery methods, but there are not commonly used or commonly defined (See the use of the term *push technology*: Depending on the abstraction the methods are called push or client-pull or scheduled-pull). My thesis will provide a set of classifications and criteria to valuate implemented Alerting Services. This criteria base on one hand on the theoretical model and on the other on the demands of the providers and users.

With the help of the developed criteria various techniques will be made comparable. The development of a design tool that supports implementational decisions is aspired.

References

[1] S. Acharya, M. Franklin, and S. Zdonik. Balancing Push and Pull for Data Broadcast. In P. Buneman and S. Jajodia, editors, *SIGMOD 1995, Proceedings ACM SIGMOD International Conference on Management of Data*, volume 23 of *SIGMOD Record*, San Jose, California, May 22-25 1997. available at `http://www.cs.umd.edu/projects/bdisk/pushpull.ps`.

[2] Soren Brandt and Anders Kristensen. Web push as an internet notification service. Talk held at 'W3C Workshop on Push Technology', Boston, Massachusetts, September 1997, available at `http://keryxsoft.hpl.hp.com/doc/ins.html`, september 1997.

[3] Antonio Carzaniga. *Architectures for an Event Notification Service Scalable to Wide-area Networks*. PhD thesis, Politecnico di Milano, Milano, Italy, December 1998.

[4] Castedo Ellermann. Channel Definition Format (CDF). Technical report, W3C, Microsoft, 1997. submitted to the W3C on 09 March 97.

[5] D. Faensen, A. Hinze, and H. Schweppe. Alerting in a digital library environment – do channels meet the requirements? Technical Report Number tr-b-98-08, Free University Berlin, 1998. available at `http://www.inf.fu-berlin.de/inst/pubs/tr-b-98-08.abstract.html`.

[6] L.C. Faulstich and M. Spiliopoulou. Building hypernavigation wrappers for querying publisher sites. In C. Peters and C. Thanos, editors, *Research and Advanced Technology for Digital Libraries. Second European Conference, ECDL '98, Heraklion, Greek*, volume 1513 of *Lecture Notes in Computer Science*. Springer, 1998.

[7] M. Franklin and S. Zdonik. A framework for scalable dissemination-based systems. In *Proceedings of the 1997 ACM SIGPLAN Conference on Object-Oriented Programming Systems, Languages & Applications (OOPSLA '97)*, volume 32 of *SIGPLAN Notices*, Atlanta, Georgia, Oct 5-9 1997. available at `http://www.cs.umd.edu/projects/bdisk/oopsla97.ps`.

[8] Craig A. Knoblock and Jos Luis Ambite. *Agents for Information Gathering*. AAAI/MIT Press, 1997.

[9] Balachander Krishnamurthy and David S. Rosenblum. Yeast: A General Purpose Event-Action System. *Transactions on Software Engineering*, 21(10), Oct.1995.

[10] G.R. Malan, F. Jahanian, and S. Subramanian. Salamander: a push-based Distribution Substrate for Internet Applications. In *USENIX Symposium on Internet Technologies and Systems, Monterey, California, December 8-11, 1997*, volume 32, 1997. available at `http://www.eecs.umich.edu/~rmalan/publications/mjsUsits97.ps.gz`.

[11] W. E. Moen. *The ANSI/NISO Z39.50 Protocol: Information Retrieval in the Information Infrastructure*. NISO Press, Bethesda, MD, 1995. available at `http://www.cni.org/pub/NISO/docs/Z39.50-brochure/50.brochure.toc.html`.

[12] Alexandros Moukas. Amalthaea: Information Discovery and Filtering using a Multiagent Evolving Ecosystem. In *Proceedings of the Conference on Practical Application of Intelligent Agents & Multi-Agent Technology*, 1996. available at `http://moux.www.media.mit.edu/people/moux/`.

[13] Hiroshi Mukaiyama. Technical aspects of next generation digital library project. In *ISDL 1997 International Symposium on Research, Development and Practice in Digital Libraries*, Tsukuba Science City, Japan, November 18 - 21 1997.

[14] A.M. Odlyzko. Tragic loss or good riddance? The impending demise of traditional scholary journals. *International Journal of Human-Computer Studies*, 42:71–122, 1995.

[15] T.W.Yan and H.García-Molina. SIFT - a Tool for Wide-Area Information Dissemination. In *USENIX 1995 Technical Conference on UNIX and Advanced Computing Systems, Conference Proceedings*, pages 177–186, NewOrleans, Louisiana, Jan 16-20 1995. USENIX Association, Berkeley, CA, USA.

[16] B.C.M. Wondergem, P. van Bommel, T.W.C. Huibers, and Th. van der Weide. Towards an Agent-Based Retrieval Engine. In J. Furner and D.J. Harper, editors, *Proceedings of the 19th BCS-IRSG Colloquium*, pages 126–144, Aberdeen, Scotland, April 1997. Robert Gordon University.

[17] Ariadne. `http://ariadne.inf.fu-berlin.de:8000`.

[18] Compuscience. `http://www.zblmath.fiz-karlsruhe.de/cs/CS-cs.html`.

[19] Object Management Group: Corba Event Specifications. `http://www.omg.org/corba/csindx.htm`.

[20] DARWIN: The Digital Library Project at the FU Berlin. `http://darwin.inf.fu-berlin.de`.

[21] Elsevier Contents Direct. `http://www.elsevier.nl`.

[22] Elsevier Science Channel. `http://www.sciencechannel.nl`.

[23] Global Electronic and Multimedial Information Systems for Natural Science and Engineering Innovation Programme of the BMBF. `http://www.global-info.org/`.

[24] ISI: Alerting Services, formerly research Alert Direct. `http://www.isinet.com/prodserv/rad/radp.htm`.

[25] The London Parallel Applications Centre: The london & south-east centre for high performance computing. `http://www.lpac.ac.uk/SEL-HPC/`.

[26] Open Profiling Standard. `http://developer.netscape.com/ops/ops.html`.

[27] Platform for Privacy Preferences. `http://www.w3.org/TR/WD-P3P10-syntax/`.

[28] Springer Link Alert. `http://link.springer.de/alert`.

[29] SwetsScan. `http://www.swets.nl/sscan96.html`.