

Personalization Techniques in Electronic Publishing on the Web: Trends and Perspectives

Proceedings of the AH'2002 Workshop on
Personalization Techniques in Electronic
Publishing
Málaga, Spain, May 2002
Selected papers

Stefano Mizzaro and Carlo Tasso (Eds.)



Universidad de Málaga

Departamento de Lenguajes y Ciencias
de la Computación

Volume editors

Stefano Mizzaro
Dept. of Mathematics and Computer Science
University of Udine
Via delle Scienze, 206
Loc. Rizzi, 33100 Udine, Italy
mizzaro@dimi.uniud.it
<http://www.dimi.uniud.it/~mizzaro>

Carlo Tasso
Dept. of Mathematics and Computer Science
University of Udine
Via delle Scienze, 206
Loc. Rizzi, 33100 Udine, Italy
tasso@dimi.uniud.it
<http://www.dimi.uniud.it/~tasso>

© The authors

ISBN: 699-8193-5
Depósito Legal: MA-584-2002
Impreso en España - Printed in Spain

Table of Contents

Preface	7
Towards an Integrated Personalization Framework: A Taxonomy and Work Proposals Nuno Correia, Miguel Boavida	9
Modelling production of personalized information services and their delivery on multiple distribution channels Thomas Ritz.....	19
Adaptive Special Reports for On-line Newspapers Sébastien Iksal, Serge Garlatti.....	31
Cross-References in Web-Based Adaptive Hypermedia Hongjing Wu, Erik de Kort.....	45
Towards the tailoring of a ubiquitous interactive model applied to the natural and cultural heritage of the Montsec area Montserrat Sendín, Jesús Lorés, Jordi Solà.....	57
TORII – Access the Digital Research Community Marco Fabbrichesi.....	71
Okapi in TIPS: The Changing Context of Information Retrieval Murat Karamuftuoglu, Fabio Venuti	77
Personalization techniques in the TIPS Project: The Cognitive Filtering Module and the Information Retrieval Assistant Stefano Mizzaro, Carlo Tasso.....	89

Building Thesaurus from Manual Sources and Automatic Scanned Texts Jean-Pierre Chevallet	95
QCT and SF services in Torii: Human Evaluations of Documents Benefit to the Community Nathalie Denos	105
Toward conceptual indexing using automatic assignment of descriptors Arturo Montejo Ráez	115
Digital content sewed together within a library catalogue WebLib - The CERN Document Server Jens Vigen	121

Preface

Electronic publishing on the Web is a fast growing field which includes various heterogeneous systems for information access: traditional journals, magazines and newspapers accessible via Web, fully electronic journals, e-prints repositories, news, vertical and horizontal portals, and so on.

Both information producers and users of these electronic publishing systems experience typical problems, such as information overload, information oversupply, information waste, miss-delivery, miss-retrieval, and untimeliness. Moreover, the increasing amount of hypermedia content which characterizes the whole field of electronic publishing, as well as the mobile/wireless revolution, provide new chances, but also pose new challenges.

Personalization can increase the utility, user satisfaction, and user loyalty of electronic Web sites, by providing the user with accurate and effective services tailored to his/her specific needs, improving in such a way the quality of the transfer of information from publishers to readers.

Personalization in electronic publishing addresses: (i) the user need of receiving timely and accurate information relevant to his/her interests, (ii) the user need to be adequately supported during search of archive information, and (iii) the publisher need to proactively disseminate information only to interested users. An essential feature of personalization techniques for information access is the capability to autonomously and automatically learn user interests and preferences from the observation of user's behavior, i.e. adaptivity. This capability is based on various machine learning techniques and provides the mean to unobtrusively build user profiles.

The goal of the Workshop "Personalization Techniques in Electronic Publishing on the Web: trends and perspectives" is to review the current state of the art in the exploitation of personalization techniques within Web sites devoted to electronic publishing and to discuss major trends and open research problems for the future. Several issues and topics are relevant within the general theme of the workshop, such as, among others: personalization strategies adopted in electronic publishing, techniques for user support during search, ephemeral personalization, information filtering, integration of cognitive and collaborative filtering, innovative services in electronic publishing

portals, personalized information services for Mobile Internet, Web clipping services, evaluation criteria of electronic publishing portals and sites, new business models for electronic publishing.

The workshop includes also an account of the final achievements of the TIPS Project (Tools for Innovative Publishing in Science) within the 5th Framework Programme (contract no. IST-1999-10419), where specific personalization techniques for personalized filtering and search support have been experimented and included in tools exploited by researchers for their daily work.

We thank Nick Belkin, Nathalie Denos, Murat Karamuftuoglu, and Marco Fabbrichesi for the cooperation in the organization of the workshop. We wish also to thank Ricardo Conejo, Carlo Strapparava and all the local organizers in Malaga, who accepted and made possible the idea of organizing this Workshop within AH2002, the 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems

Stefano Mizzaro
Carlo Tasso

Towards an Integrated Personalization Framework: A Taxonomy and Work Proposals

Nuno Correia¹, Miguel Boavida²

¹ Computer Science Department, Faculty of Sciences and Technology,
New University of Lisbon, Portugal
nmc@di.fct.unl.pt

² Department of Systems and Informatics, School of Technology,
Polytechnic Institute of Setúbal, Portugal
mbv@di.fct.unl.pt

Abstract. The paper presents a survey of personalization approaches, ranging from personalized web sites to personalized multimedia video programs. One of the main objectives of this work is to find the common concepts that are shared by existing approaches. The proposals are classified according to different dimensions and properties, namely user interaction and content and presentation management. This taxonomy is used as input to our current work on personalization, described in the last part of the paper. The goal is to define a framework for personalization of interactive multimedia content, that will integrate the different dimensions that we are exploring: annotation, personalization of content and several forms of personalization of presentation, namely for interactive television.

1 Introduction

Personalization is a pervasive concept in several areas of interactive multimedia, namely web electronic publishing applications. This is caused by the need to adapt the content and presentation style to the preferences of a given user or set of users. There are different techniques for personalization, ranging from the simplest, that allow the user to personalize a web page, by defining its preferred colors or layouts, to the more complex, where the content is built on the fly, according to the profile of the user [10]. The work presented in this paper is an attempt to classify the numerous systems, models and tools for personalization as a way to identify the common characteristics among all these approaches. The main goal is to define a common framework for building personalized interactive multimedia systems that go beyond current web based electronic publishing paradigms. We are currently working on several aspects of personalization namely shared annotation spaces [1], personalization in interactive TV environments [2] and semi-automatic generation of content. The paper is organized as follows. The next section presents the dimensions and categories that were used to characterize the personalization techniques. Section 3 presents a classification of the systems according to the categories presented in the previous section. Section 4 describes our current work on different aspects of personalization, towards

an integrated personalization framework. Finally, the last section presents conclusions and directions for future work.

2 Personalization

Personalization in interactive systems can involve adapting the user interface or adapting the content to the needs or preferences of a specific user. In general, personalization of interactive systems considers these two aspects, sometimes simultaneously:

- Personalization of Presentation: allowing to personalize aspects of the user interface, including colors, position of interface items and fonts.
- Personalization of Content: where the content can be adapted to the needs and preferences of different users. For example, for a news on demand system the user would only receive the news about sports, based on a pre-defined profile. The commercials would also be about sports, but could consider the previous purchases of the user.

Considering these two aspects, we tried to characterize the different ways in which personalization is done in interactive systems. Next we present several categories used to classify the systems.

Content and Presentation Management: Comparing the normal, non-personalized, content presented to the user, with a personalized version of the same content, we define two main options: **enhancement** and **configuration**.

We define as enhancement all situations where the personalized content has additional material superimposed to the normal or base content. This kind of personalization is typically associated to continuous media, such as audio or video, where personalization may assume the form of video/audio enhancements or annotations. By contrast, configuration is present when the personalized content is adapted for presentation. This may happen when adapting the presentation layout by changing colors, fonts and positions. One example where configuration happens is when the content has to be adapted for presentation: for example color pictures have to be converted to black and white or HTML content has to be converted to WML (for WAP enabled mobile phones). Additional issues arise when a more complex semantic change has to be made. Examples of this are the generation of summaries from video streams [3,8]. In this case rules are necessary for adapting the content. These rules are based on the characteristics of the medium, such as the narrative and aesthetic properties, and the user preferences.

User Interaction: Personalization may take many forms, ranging from the explicit definition of a profile to an automatic process of user categorization based on behav-

ioral analysis. We define personalization according to the interaction that is required from the user as:

- **Explicit:** the user is invited to manually state her preferences, by answering questions that are directly related with the content that is being delivered.
- **Implicit:** the user interacts with the system, expressing her level of happiness regarding the content being delivered.

The two options above are also considered in [7]. Finally, we may have a non-interactive approach, were the user profile is collected automatically by data mining behavioral information taken from access logs or from subsidiary systems.

Group Personalization: Personalization cannot be seen only in terms of the individual user. Sometimes, we may assume that, specific content may be of interest to all the users from a given geographical area, age, from the same organization, or that share the same main interests.

2.1 Annotation as Personalization

In the next paragraphs the classification above is applied to annotation. Our previous work in annotation [1] and additional approaches provide input to this. This paper puts a special emphasis on annotation because it is a very useful and natural way to enhance electronic media, especially if we consider shared annotations. Also, annotation in digital form can change the traditional paradigms of author and reader in electronic publication, by giving an active role to the usually passive reader. Annotation is the most traditional form of personalization. Even before there were digital media solutions, annotations were done by adding personalized content to books and other types of printed materials. Annotated materials enhance the assessment experience and annotations become part of the content. These principles transposed to a digital media, have been addressed not only using textual materials but also in video and television. An annotation in a video can be almost everything: text, sound, a picture or even another video may be quoted as a note. Next, annotation is characterized in terms of the properties defined previously.

Content and Presentation Management: Annotated materials are by definition produced by adding new content to a previously existing source. As such, annotations are always enhancements to the original material. When annotations are already present in the system, a form of configuration may happen: the system filters the annotations that are present to the user, accordingly to its implicit or explicit profile.

User Interaction: Digital annotations invite the user to an active participation in the personalization process. An explicit form of user interaction will always be coherent to the active role that user has. Non-interactive rules may be difficult to implement in annotation systems, and only applicable to specific systems where the role of the user in the system is well established. In general, user interaction in annotations is explicit.

Group Personalization: Shared annotations are very important, enabling collaborative tasks in user communities. The new content added by an individual could be very important to others of the same social or professional group. Shared annotations, constrained or not by security rules, may be presented to groups of users with similar interests.

3 Personalization Approaches and Taxonomy

This section presents several personalization approaches that are characterized according to the properties that we defined in the previous sections. The approaches that are presented were chosen to cover the more relevant personalization systems and techniques. The described systems include the possibility to choose multimedia content [4], the generation of personalized hypermedia information [5,6], the automatic construction of multimedia content [7], web site personalization [8,12], video enhancements [9], a personalized TV information system [10] and a personalized library for active learning [13]. Each of the systems is briefly described and then characterized in terms of the properties that were presented in the previous section.

Personal DJ [4]: The expected upcoming of new kinds of portable multimedia devices, like the audio wearable computer, will give a new strength to the research efforts in personalized audio. Half way between the user controlled CD player and the DJ mediated radio broadcast, there is the personalization of audio content. User interaction with radio like devices is normally processed using some sort of station/genre selection and in a system like Personal DJ, a fast forward button is used to advance to the next song. This feedback from the user, coupled with some poll information provides a first level of profile adaptation.

- Content and Presentation Management: Configuration. The personalized content is a subset of all the possible content, i.e. music, available for delivery.
- User Interaction: Implicit. This medium is excellent for implicit interaction. The user interaction (advance to the next song) may be used to infer the adequacy of the user profile to his level of "happiness" with the songs being selected.
- Group Personalization: Not described. A system like personal DJ seems to be targeted to individual preferences. Anyway, some sort of genre or mood diversity could be injected in the playlist by observing accepted songs played to users with a similar profile

HERA [5]: The authors defend the automatic generation of hypermedia over semi-structured data as a more effective way of delivering information to users as opposed to the approach normally taken in legacy systems based on strongly structured data. The presented system, HERA, accesses collections or libraries of digital semi-structured data and try to derive a hypermedia presentation that is adapted to the users situation. The prime goal is that the derived presentation "adds value" based on knowledge about the data, the user and the application itself. The generation process starts from XML data that represents the query result. This data is then taken by a

"presentation manager", the software responsible for the actual presentation in the user browser.

- Content and presentation management: Enhancement. The main purpose of the system is to provide additional information (enhancements) to an original query result. These additional information results are presented as optional links in the generated hypermedia presentation.
- User interaction: Non interactive. The system builds a presentation based on a query, and this query may be expressed implicitly or in an explicit form. The personalization is performed automatically not involving any form of user interaction.
- Group personalization: Not described.

PERSIVAL [6]: PERSIVAL is a system that uses personalization to improve the search capabilities of an healthcare information infrastructure. In healthcare settings, patients may need access to online information that can help them to understand their medical situation while physicians need information that is clinically relevant to an individual patient. A patient or physician query is augmented with important information taken from the patient record and then is processed by a multimedia search engine. The final result takes the form of a multimedia presentation, where textual summarization is produced with references to relevant articles and video data (segmented and presented as a storyboard).

- Content and presentation management: Enhancement. The basic query performed by a patient or physician is augmented and the result is an hypermedia presentation assembled by the system.
- User interaction: Explicit (medical record). This is a domain where explicit interaction works well. The medical record already exists, and contains all the required information the system needs to create a personalization profile.
- Group personalization: Yes. The group personalization features are implicitly present, when the system provides information that is relevant to a group of patients or specialists.

Personalized TV News programs [7]: Merialdo et al. discuss personalization in the generation of TV news programs. The construction of customized programs will be an important paradigm in the near future, as a result of the development of Digital Television. TV News is a good candidate for this kind of customization, because it is a very successful type of program and the current broadcast paradigm is very rigid. With personalized news, every user would expect to have a TV news program at a time of his choice, with duration and content that specifically matches this user interests.

- Content and presentation management: Configuration. Some elements are removed from the original presentation (content configuration) and the ordering of topics in the presentation (layout configuration) is automatically performed by the system.
- User interaction: Explicit and implicit. The duration of the program and main aspects of the user profile are explicitly stated by the user. Implicit personalization is used, enabling profile evolution based on user feedback.

- Group personalization: Not described.

PROTEUS [8]: PROTEUS is a web site personalizer system that observes the behavior of web visitors and automatically customizes and adapts the site for each mobile visitor. A personalizer may be associated with one web site or situated on a proxy server and adapt many sites. It can also exist on the browser device and serve only one visitor. A web site personalizer can make frequently visited destinations easier to find: highlight content that interests the visitor or elide uninteresting content and structure. The key information behind web site personalizers is that a great deal of information about visitors is readily available in the form of access logs (at the web site or at an intermediary web proxy)

- Content and presentation management: Configuration. The web site personalizer works at the web-site layout level, in order to improve the web site usability. The personalizer may change a link location or highlight content, based essentially in the analysis of access frequency.
- User interaction: Non interactive. Information about the user is obtained through data mining, namely by the analysis of the web server access logs.
- Group personalization: Not described.

Enhancements in Digital Sports Broadcasts [9]: In this system, the personalization is viewed as the possibility of enabling custom enhancements in television broadcasts. These enhancements are provided to the user by exploring additional data sources, supplied by the event organizer or by specialized data providers. The personalization of the digital broadcast, rely on the possibility given to the user to accept the enhancements available.

- Content and presentation management: Enhancement. The main broadcast is enhanced with graphical or textual information obtained from additional data sources.
- User interaction: Explicit. The user accepts or rejects the proposed enhancements, having full control over the "visual purity" of the event.
- Group personalization: Not described. The prototype assumes that the same enhancement data is sent to all viewers.

PTV [10]: PTV (Personalized Television Listings Service) is an Internet service that provides personalized TV listings content to over 20.000 users in Ireland and Great Britain. The system is based in a content personalization engine named ClixSmart, developed in the department of Computer Science at University College, Dublin. ClixSmart performs two essential tasks: it monitors the online activity of users (from a given website) and automatically constructs profiles for these users. The user profile information is used to personalize a target website by filtering information content for the target user. The ClixSmart personalization manager employs different content filtering strategies: (1) content based filtering and (2) collaborative filtering. A content-based filtering approach seeks to recommend similar items to the items a user liked in the past, while collaborative filtering recommend items that a similar user also liked.

- Content and presentation management: Configuration. The television listing produced for a user is content adapted to its known interests.
- User interaction: Implicit. The system monitors online activity of users and asks them to rate their recommendations when trying to gather information that can help to build or evolve the user profile. At the moment of registration the system tries to sketch the main guidelines of the users profile through some sort of explicit interaction. The explicit user interaction is mainly a bootstrap procedure so user interaction is categorized as implicit.
- Group personalization: Not described. The collaborative filtering approach may include some sort of group recommendation. Recommending an item based on user similarity can be complemented with group recommendation.

Active Web Museum [12]: The Active WebMuseum is a user-adapting website, that uses the collection of paintings from the WebMuseum, Paris. In an ideal world a visitor of a museum would enter the museum and then find in the first corridor exactly those items, which he would find most interesting. This approach, that is impossible to implement in a real museum, becomes feasible when a museum's art collection is presented through the web. The web museum uses content-based filtering and collaborative filtering as the main techniques for generating personalized content. The content categorization of a painting is difficult and time consuming. The Web Museum uses automatic content categorization of the digital images, based on color, texture and caption information. Preferences are obtained by inviting users give symbolic ratings to paintings: excellent, good, neutral, bad, and terrible.

- Content and presentation management: Configuration. The museum is essentially the same for all users, the main corridor being a personal view, i.e., it contains the paintings that a given user would find more interesting.
- User interaction: Implicit. The user give ratings to a painting whenever she wants, or when asks for a detailed view of a painting. The user profile is not expressed in an explicit form, but may be inferred from the ratings.
- Group personalization: Not described.

Active Learning in Digital Libraries [13]: Active learning is the ability of learners to carry out learning activities in such a way that they will be able to construct knowledge from information sources effectively and efficiently. PIE (Personalized Information Environment) is a framework that provides a set of integrated tools based on individual users requirements and interests with respect to access to library materials. PIE deals with material personalization and collection personalization. Material personalization corresponds to facilities for learners to use library materials according to their individual requirements. Collection personalization captures the learners learning context and interest from the material personalization in order to provide a personalized view of the organization of the digital library.

- Content and presentation management: Mostly enhancement. PIE is composed by several tools, most of them working as annotation managers or performing some sort of personalized query enhancement.
- User interaction: Non interactive. The personalization data is obtained through the analysis of the annotated documents that form the personal library of the user.

The user profile is built based on the documents constructed by the user, using "shallow copy" tools for the different multimedia materials

- Group personalization: Not described.

4 Current Work

Currently we are doing work on several personalized information systems, that allow to explore the different dimensions and properties that were characterized above. The most challenging one is to generate new content on the fly, without explicit user interaction. We are working towards this goal but still on intermediate levels. Current paradigms for web based electronic publishing will also evolve in order to accommodate these new dimensions. The traditional roles of the reader and author will be blurred allowing a more active participation from everyone. Current developments are summarized in the next paragraphs.

Video annotation tools: Our previous work reported in [1] allows to personalize video based documents. The tool, AntV, provides an interface for adding annotations of different types (text, audio, video) to a given video stream. Each user can have its private set of annotations and can choose if it wants to publish those annotations or not. Our first AntV prototype worked as a standalone application, but more recently we have upgraded it to work and display results on the Web. We are using SMIL (Synchronized Multimedia Integration Language) as the output for the hypermedia documents that combine video and annotations. The personalization characteristics in this system, result from the fact that the user can add its own materials, in the form of annotations and it can generate its own "new" hypermedia documents from the original video and annotations made by different users.

Personalization in interactive TV environments: In interactive TV environments the need for personalization is even bigger, given that the TV set is usually shared by different users in the same house. We have made a first experiment on interactive TV personalization [2]. The prototype, named MyTV, is a design for the customization of interactive TV services. It is based on pre-defined templates that can be customized by the user. Although it is a very simple experiment it allows to define the design characteristics that are more important for a given user in an interactive TV environment.

Semi-automatic content generation: Generation of content based on the user preferences or choices, allows building customized and personalized information spaces. We are exploring this concept, in two different ways: (1) as a result of the video annotation process, as described above, the original video is combined with the annotation materials resulting in a new video or in a hypermedia document; (2) as a result of applying the user preferences, to multimedia queries. The user specifies the type of preferred content, in terms of topics, durations, media types and the system creates a document that conforms, as much as possible, to that specification.

Personalization based on user profiles: Based on standard XML/XSLT technologies for generating content for different types of users and devices, including mobile devices and interactive television, we are building a platform for interfacing user profile information. This platform will be able to handle user information that was gathered automatically or that was introduced by the user. It will work as a gateway between the content personalization system and the various ways to gather information about a user or a set of users.

5 Conclusions and Future Work

The work described in this paper helps to characterize the personalization techniques that are currently in use for interactive publishing and distribution systems. This characterization is a first step to integrate the personalization applications that we are developing. The different systems that were surveyed convey distinct aspects of personalization, but have many similarities in the ways that profile information is gathered or personalization is achieved, by configuring content or presentation. Implicit personalization is, in general, the preferred personalization technique although sometimes it is difficult to implement. This happens mainly in the bootstrap process when there is not enough data available. Regarding content and presentation management both techniques (enhancement and configuration) are relevant, depending on the application. Enhancement is mostly used in learning systems and it is a very powerful tool for augmenting existing materials. Configuration is also essential in content presentation systems where the amount of information is immense and it must be tailored and selected to fit the needs of a given user. Group personalization (or shared personalization settings) is not supported by many systems, but it is a helpful technique for managing large numbers of users with similar interests.

Future work on personalization techniques includes the extension of some of the systems that are described in Section 4. This work will be oriented towards the identification of common, reusable software modules that will be shared by the different applications. The objective is to build a software framework that will include automatic and manual personalization tools, personalization of content and personalization of presentation. Personalization of presentation will use current XML based techniques that are commonly used for building electronic publishing systems on the web, combined with modules and interfaces for getting user profile and statistical information. Regarding personalization of content we will continue to work towards the semi-automatic generation of content, based on user preferences and user input, such as annotation. Our ultimate goal is to have a flexible content generation system that provides the information that the user wants, when and where it is needed.

References

1. Correia N., Chambel T.: Active Video Watching Using Annotation, ACM Multimedia'99, Orlando, Florida, USA, (1999)
2. Correia N., Peres M.: Design of a Personalization Service for an Interactive TV Environment, Submitted, (2002)
3. Correia, N., Martins, J., Oliveira, I., Guimarães, N.: WeatherDigest: An Experiment on Media Conversion, Proceedings of SPIE'95 International Symposium on Information, Communications and Computer Technology, Applications and Systems, Photonics East'95, Philadelphia, PA, USA, (1995)
4. Field, A., Hartel, P., Mooij W.: Personal DJ, an Architecture for Personalized Content Delivery, WWW10, Hong Kong, (2001)
5. Houben, G., De Bra, P.: Automatic Hypermedia Generation for Ad-hoc Queries on Semi Structured Data, ACM Digital Libraries, San Antonio, Texas, USA, (2001)
6. McKeown, K., et al.: PERSIVAL, a System for Personalized search and Summarization over Multimedia Healthcare Information, JCDL'01, Roanoke, Virginia, USA, (2001)
7. Merialdo, B., Lee, K., Luparello, D. Roudaire, J.: Automatic Construction of Personalized TV News Programs, ACM Multimedia' 99, Orlando, FL, USA, (1999)
8. Anderson, C., Domingos, P., Weld, D.: Personalizing Web Sites for Mobile Users, WWW10, Hong Kong, (2001)
9. Rafey, R., Gibbs, S., Hoach, M., Van Gong, H, Wang, S.: Enabling Custom Enhancements in Digital Sports Broadcasts, WEB3D 2001, Paderbon, Germany, (2001)
10. Smyth, B., Cotter, P.: A Personalized Television Listings Service, Communications of the ACM - vol 43, n 8, (2000)
11. Oracle: The Art of Personalization, An Oracle White Paper, (2001)
12. Kohrs, A., Merialdo, B.: Improving Collaborative Filtering with Multimedia Indexing Techniques to create User-Adapting Web Sites, ACM Multimedia' 99, Orlando, FL, USA, (1999)
13. Jayawardana, C., Hewagamage, K., Hirakawa, M.: Personalization Tools for Active Learning in Digital Libraries, MC Journal: The Journal of Academic Librarianship, 8 (1), (2001)

Modelling production of personalized information services and their delivery on multiple distribution channels

Dipl.-Inform. Thomas Ritz

Fraunhofer IAO
Nobelstr. 12
D-70569 Stuttgart
Thomas.Ritz@iao.fhg.de

Abstract. Personalized information services have become a typical information product offered online on different media, but mainly the web. After a short definition of personalized information services we propose a modelling formalism, defining the building blocks for the production of personalized services and is applicable as “glue” between business processes modelling and software engineering. Within this paper we will instance this model for a very simple personalization method and show how the former elaborated model could be mapped on a three-tier software architecture. Finally it is shown that these architecture eases the personalized delivery on different output channels (f.i. HTML, WML and VoiceXML), based on one single content repository.

Introduction to Personalized Information Service

Need for personalization

Personalized delivery of content is gaining importance, which is reflected in ongoing research. A huge number of such services are not primarily designed as information product like a newspaper. Mostly the services are designed to cope with 1:1 marketing aspects [1, 2]. Nevertheless the added value for publisher as well as for consumers of such information services is proved [3] and could be traded as part of value-added-publishing strategies [4]. As common subset of definitions which could be found in literature [5,6,7], we could state that personalized information services are service which deliver the right information at the right time on a media preferred by the consumer [8]. Information is considered in the following as multimedia information (text, audio and video) as proposed by Smyth [9]. Furthermore, we assume publishers in the right position to produce and deliver such services.

Definition of personalized information services

As stated before, lots of services are labelled personalized service. A definition based on common characteristics will result in a generic definition like

“Personalization is a service provided based on a user profile”. Which individual services generated using this profiles is still uncertain. The definition as follows (c.f. [3]) tries to integrate the dimensions found in definitions cited before:

- process
- user interest (textual)
- demand in schedule
- demand in media

Def. 1: A Personalized Information Service is a service towards a customer comprising

1. *filtering* of information out of former *gathered* and *qualified* information regarding users textual *interest*
2. *presentation* of this information using a user defined *time schedule* and *media* appropriate with recent user environment.

This brings us back a definition saying “personalized information services are services which deliver the right information at the right time on a media preferred by the consumer” [8].

Production of Personalized Information Services applying the mass customization paradigm

The production of information services

As for other products, the production process became part of competition in the media market [10]. Most of the efforts spent in research on production of media and information services focussed on “media independent publishing” strategies [11], which are quite often referred to as Cross-Media-Publishing ([12] or [13]). Hand in Hand with this developments it became standard to handle media elements in different formats and media during the production process as text, audio and video material [14]. To stick to this paradigm of cross-media publishing, markuplanguages as SGML and nowadays XML became common to handle and integrate the different media elements and to define a semantic structure for the information [15]. As the distribution media for information products lost focus in the production chain, more attention was spent to focus on market demands and target groups on information level [10]. The editorial work got more attention [14] but the discussion still sticks to the production of mass-media [16].

Modelling mass customized information services

We start setting up a theoretical framework by modelling the product to be produced. Same as for other customized or individual products, we assume modularity to be a crucial factor for efficient production [17]. Therefore we assume a building block structure [18] as shown in Fig. 1.

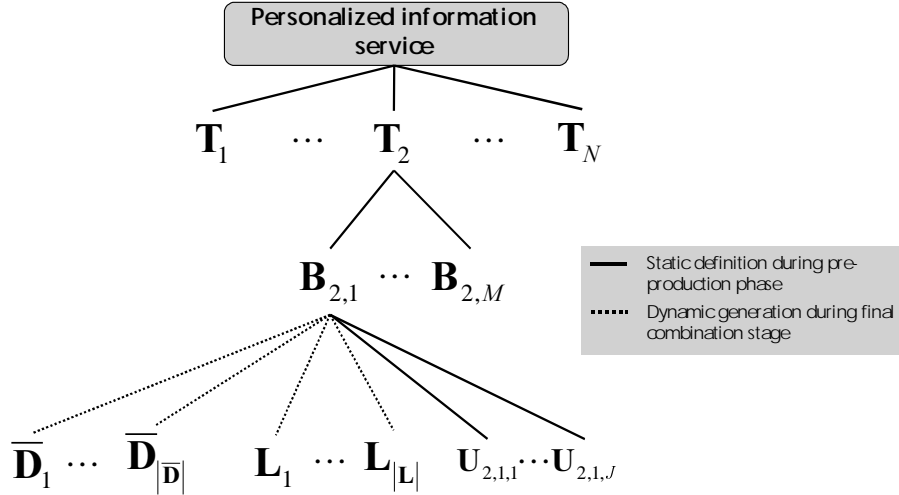


Fig. 1.: Building blocks for personalized information service

Depending on the user profile, the product shipped to the consumer consists of templates ($\mathbf{T}_i \in \mathbf{T}$) which aggregate building blocks ($\mathbf{B}_{i,j} \in \mathbf{B}$). \mathbf{T} and \mathbf{B} are sets of all available templates and building blocks. A building block represents a functional cluster (f.i. show teasers). So a building block may select a set of documents from the content repository ($\overline{\mathbf{D}}$) and renders it. Please note that this assignment of documents is not fixed. The dotted line should indicate that the building block is dynamically populated with content, taking the user profile under consideration. Further to the documents, a building block may contain links (\mathbf{L}). Links are relations between instances of templates. Finally a building block may contain user interfaces for functions as login etc.. The example shown in Fig. 2 illustrates the construction of a sample home page with these building blocks.

Knowing about the structure helps to investigate the production process for personalized information services. We start with the personalization itself. It could be easily modelled starting with a generic information retrieval model as proposed e.g. by Fuhr [19] (for further formalisms see f.i. [20]). This model shown in Fig.3 shows that Documents D are in real world in certain relations R to user queries Q . In order to process them electronically, both real world entities have to be mapped to a digital representation (i.e. for vector space retrieval or an ontology based classification). This is done by the functions α . The resulting digital representations could be optimized by β at runtime to a representation appropriate for the particular information retrieval concept applied (i.e. handling of sparse matrices). Finally, the information retrieval is done by applying the information retrieval function $\Theta: \mathbf{D} \times \mathbf{Q} \rightarrow \mathbf{IR}$, which result in a rank of the documents with regard to a users query.

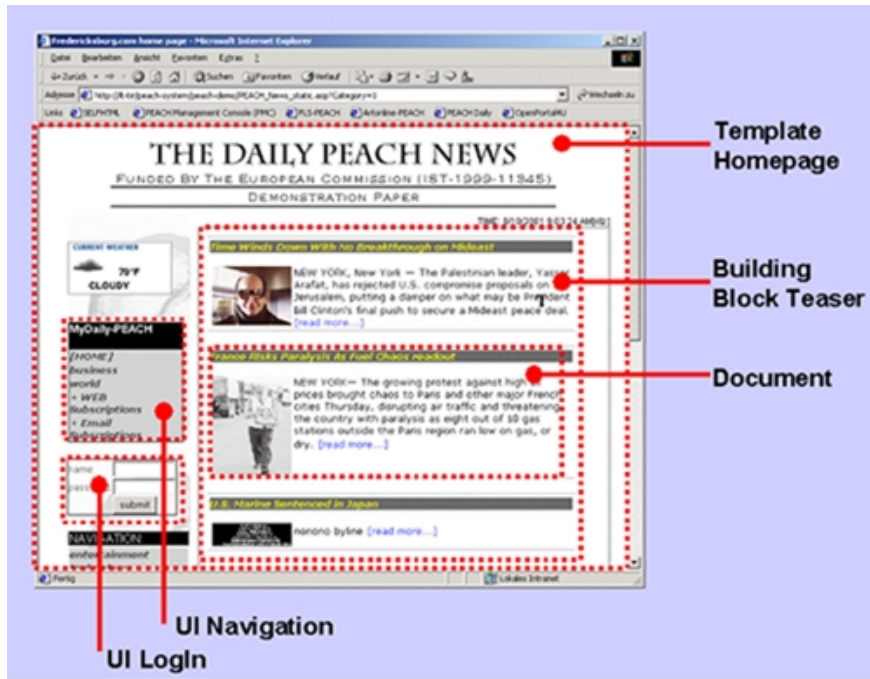


Fig. 2.: Sample of Buildingblock Structure

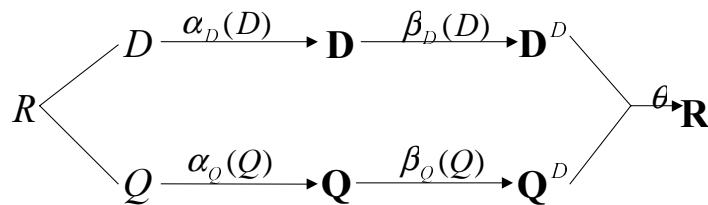


Fig. 3.: Information Retrieval Modell [19].

Personalization could be seen in fact as such an information retrieval process. The user query have to be seen as a set of user queries defining the users profile. But as seen, users preferences could contain much more than textual interest [3]. Therefore we define a user query as a data structure and a couple of methods to tailor this data structure:

The first component (C_p) of the vector is containing the description of preferences in contents (e.g. tennis). The second component (M_p) contains a media preference for this (i.e. WAP), while the last component (S_p) defines a schedule. So

a user could profile his interest, e.g. for “business news every morning on the PDA”. To give the user the chance to tailor its profile, we introduced two functions to change the digital representation of the user profile. The first ($\tilde{\alpha}_Q(Q_P)$) is intended for manually changing a profile. The second one ($\hat{\alpha}_{Q_P}(\mathbf{D})$) is aimed to tailor the user profile with document representations (i.e. TF vectors). This can be used to tailor the user profile applying relevance feedback methodologies [20, 21].

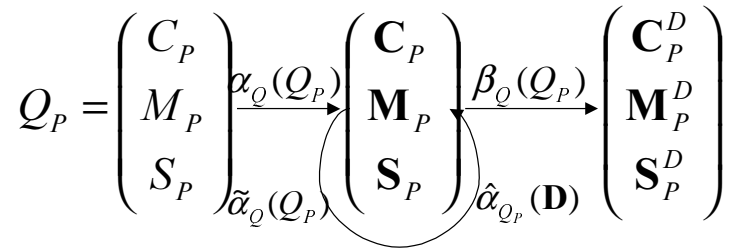


Fig. 4. Extended IR Modell for modelling user profiles

Further to this, the information retrieval model results in a ranking of contents. For personalized information services we have to result in an information product as modelled in Fig. 2.. So we extended the IR Modell again by the building blocks of a personalized information service (c.f. Fig. 1.) and resulting in a the model shown in Fig. 5.

In this model the user requests information from a certain media at a given schedule. The presentation function $\wp(P, M, S)$ renders the information product making use of the building block structure defined before. So a template appropriate with the requested media is called which contains several building blocks. This building blocks make use of the information retrieval function and retrieve the relevant documents (as in the standard IR Modell). While \mathbf{D} is the digital representation for information retrieval (computed by α), the digital representation of the documents $\bar{\mathbf{D}}$ contains all material of an article elaborated during the editorial process. To tailor it for the special need in the building block, a filter \mathbf{F} is employed by the function $\phi(\bar{\mathbf{D}}, \mathbf{F})$. This filters extract f.i. only the headline and the first paragraph for a teaser presentation (In fact $\bar{\mathbf{D}}$ and \mathbf{F} are realized as XML and XSLT documents in the later described application).

For the editorial processing, we introduce the editor function $\varepsilon_D(D, Y)$, which is used to markup a document in the real word according to a given document type definition \mathbf{Y} , which defines the document structure. Due to the fact that \mathbf{Y} , \mathbf{F} , \mathbf{B} and \mathbf{T} are not static, we defined design functions δ for this purpose. The user interfaces to the user profile functions ($\alpha_Q(Q_P)$, $\tilde{\alpha}_Q(Q_P)$, $\hat{\alpha}_{Q_P}(\mathbf{D})$ and $\lambda(P)$) are collected in the set \mathbf{U}_M for easy notation purpose.

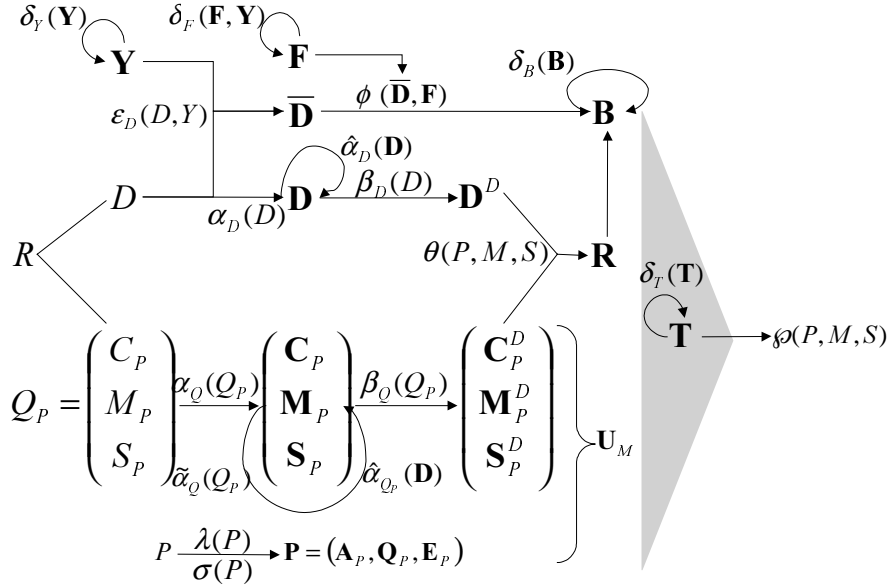


Fig. 5. Model of production and provision of personalized information services

Now we modelled all components needed to produce and provide personalized information services. The use of this model is manifold. We can use the model for requirements engineering by assigning model elements to production and workflow processes (c.f. [22]). Secondly, we can use the model for software engineering purpose by implementing the components. Our model eases this, cause dependencies of model components became transparent. But the greatest use is the mediation between requirements engineering and software engineering.

Employing the model

To show an instance of the former introduced formal model, we will employ the model for a very simple personalization application. The user checks on a multiple choice form, whether he is interested in certain categories or not. We assume that the contents were manually classified according the same classification scheme within the newsroom.

Let $\mathbf{KT} = (K, R)$ be a tree. K are the leaves of this tree and reflecting the categories. The edges $R = \{(k_i, k_j) \mid k_i, k_j \in K; 1 \leq i, j \leq |K|\}$ allow to order the leaves hierarchically. The manual assignments of documents to leaves form a relation $Z_D^M \subset K \times D$ between documents and leaves (vgl. [23]). This relations represents the digital representation of the Documents \mathbf{D} and the mapping function α_D is the

manual assignment. As mentioned before the user is asked to fill a multiple choice form in order to assign himself to parts of the tree **KT**. Thus a relation $Z_Q^M \subset K \times Q$ is established and defines C_p , the component of the user profile describing the users textual interest. The retrieval function θ could now be easily computed by evaluating the product of the relations, thus $\theta = Z_Q \circ Z_D$. Table 1 shows the use of model components in a multi-tiered-architecture for the buiding block which is used to select and present teasers to the consumer.



	$b \in \mathbf{B}$	Building Block HTML Teaser
Presentation	$\forall d \in \theta :$ $\phi(d, f) f \in F$	<p>Nascar after attack ...</p>  <p>Last week's terrorist attacks in New York and Washington will have an effect on the mood surrounding the MBNA Cal Ripken Jr. 400. And though the increased military presence in and around Dover Air Force Base will bring the reality of our world home that much more clearly, the direct impact on fans will be fairly minimal. That is, though there necessarily will be heightened security (see Question 2) throughout the weekend, few fans will bear any real consequences. [read more...]</p> <hr/> <p>Clemens First Pitcher to Go 20-1 as Yankees Roll On</p>  <p>Roger Clemens was not at his sharpest, striking out only one, in becoming the first pitcher to have a 20-1 record. The phone calls from family and friends began at noon today and came to Roger Clemens in rapid succession; good luck, best wishes. But no one mentioned how Clemens had the chance to become the first pitcher to win 20 of his first 21 decisions. They just kept reminding Clemens that he was a representative of New York. [read more...]</p>
Logic	$\theta =$ $Z_Q \circ Z_D$	Implemented as method, wich calls a SQL statement.
Data	$\bar{\mathbf{D}} \mathbf{D} \begin{pmatrix} \mathbf{C}_p \\ \mathbf{M}_p \\ \mathbf{S}_p \end{pmatrix} \mathbf{F}$	Relational DB ; $\bar{\mathbf{D}}$ XML documents; \mathbf{F} XSLT documents

Table 1. : Definition of a building block

Within the European funded research project PEACH, several other Information Retrieval and classification methodologies were applied, as semantic classification and full text retrieval. The former presented model allowed us to compare these methods mutually. Further more we placed elements of the former mentioned personalization model in business process models (c.f. [22]). As a result you can see that f.i. manual classification causes bigger efforts in the editorial process and in the

end limits the user to a small number of categories. In comparison automatic/semiautomatic classification causes rather no efforts for the editors, but is not easy to set up and to understand by the end users. Fulltext search causes again no additional efforts, the users are not mapped on a comparable classification scheme which makes recommendation and collaboration more difficult.

Implementations

As shown in the example the model could be easily transferred to a 3-tier software architecture [24, 25].

This approach ensures that the business logic components could be reused for several media, by handling the particular presentation in the presentation layer. The data layer comprises the atomic information. The business objects are designed to deal with basic functionality which can be easily embedded in the presentation layer and which encapsulates the data layer from direct access. This is extremely important for critical functions (i.e. dealing with the user profile), cause business rules i.e. for handling privacy issues have to be defined and validated once.

We realised these objects as COM objects which could be used by ASP scripts. The business objects fosters the programmer to the following process shown in Fig 6..

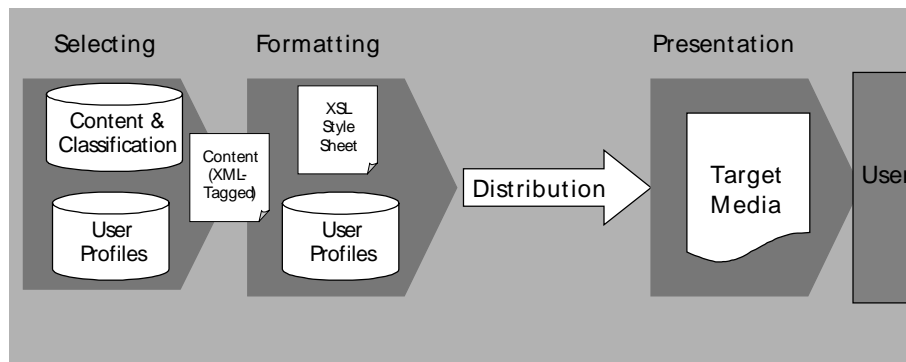


Fig. 6. The overall process of selecting and formatting contents could be shown as follows

First the information is selected by bringing user and content object together (this results in a set of contents appropriate for the recent user). Following to this the contents could be formatted within ASP pages (or any other COM capable scripting or programming language). Applying this principle, it became possible to easy generate different forms of personalized information (f.i. HTML and VoiceXML):

WWW

```
`SELECT
<%
Set CO =
CreateObject("ContentObject.ContentClass")

SET rs_contents_caption =
CO.Select_My_Categories(UO.UserID,"<web>")

%>

`FORMAT in HTML

<%

While Not qry_return.EOF

    response.write "<a
href='sample1.asp?Category=" &
qry_return("ID")& "'>" & qry_return("Name")
& "</a><br>"

    qry_return.MoveNext

Wend

%>
```

VoiceXML

```
`SELECT
<%

Set CO =
CreateObject("ContentObject.ContentClass")

SET rs_contents_caption =
CO.Select_My_Categories(UO.UserID,"<Voice>")

%>

`FORMAT in VoiceXML

<menu id="nyc_menu" dtmf="true">

    <prompt>

    <audio src="jingle.wav" />
```

```

Hi, <%=UO.User%>

        My PEACH daily news have the
following categories for you

</prompt>

<enumerate>

    For <value expr="_prompt"/>, press
    <value expr="_dtmf"/>

</enumerate>

<%

do while not rs_contents_caption.EOF
response.write "<choice
next='http://lt-tir/peach-
system/test/teaser_vxml.asp?
category='& rs_contents_caption("ID")
&''>" & rs_contents_caption("Name")

response.write "<grammar>" &
rs_contents_caption("Name") &
"</grammar>" & "</choice>"

rs_contents_caption.movenext

loop

%>

```

Table 2. Scripting of personalized navigation for HTML and VoiceXML browsers

Conclusion

After an introduction into personalized information services, the text presented a formal modelling method to describe personalized information services from an abstract and modular view. To prove its viability the model was applied for personalization based on categorized text's and user's. The model presented was mapped on a 3-tier architecture and a short introduction into the resulting object-oriented technology could be given. Finally the resulting business objects were shown "in action", to produce an personalized HTML and VoiceXML navigation based on a single content repository. In addition to the sample application referred to in this text, the model proved its viability within the PEACH project funded by the European Commission's IST programme (IST-1999-11345).



Dipl.-Inform. Thomas Ritz, born 1971 studied computer science and economics at the University of Bonn and obtained his degree in computer science in 1997. Since 1997 he is scientific employee at the University of Stuttgart IAT and the Fraunhofer IAO. He has a long standing experience in publishing industry and focuses his research activities on online communication and information. In several research projects he designed and implemented solutions for electronic information exchange, online communication in virtual communities, branch specific information portals and electronic commerce.

Mr. Ritz acts as project manager in the DISMED project (funded by the European Commission in FP 4 Innovation program RSE 067) and in the PEACH project (funded by the European Commission in FP 5 IST programme IST-1999-11345). Besides he joined several public and industry funded research, consultancy and development projects.

References:

- [1] J.W. Palmer and L.B. Eriksen "Digital Newspapers Explore Marketing on the Internet," *Communications of the ACM* vol. 42, no. 9, pp. 33-40, 1999.
- [2] H. Fisbeck "Customized Marketing im Internet," *Akademische Abhandlungen zu den Wirtschaftswissenschaften* 1999.
- [3] T. Ritz "Personalized Information Services - An Electronic Information Commodity," *e-Business Strategy Management* vol. 2, no. 2, pp. 105-114, 2000.
- [4] H. Berghel "Value-Added Publishing," *Communications of the ACM* vol. 42, no. 1, pp. 19-23, 1999.
- [5] C. Allen "Personalization vs. Customization," <http://clickz.com/cgi-bin/gt/en/pm/pm.html>, 2000.
- [6] J. Nielsen "Personalization is Over-Rated," <http://www.useit.com/alertbox/981004.html>, 2000.
- [7] IBM Developer Works "Web site personalization," <http://www-4.ibm.com/software/developer/library/personalization/index.html>, 2000.
- [8] T. Ritz "Personalized Information Services - an electronic information commodity and its production," *ICCC/IFIF conference on Electronic Publishing (ELPUB 2001)* pp. 48-58, 2001.
- [9] B. Smyth and P. Cotter "A Personalized Television Listing Service," *Communications of the ACM* vol. 43, no. 8, pp. 107-111, 2000.
- [10] M. Freter "Crossmedia-Publishing - Vorbereitung auf den Wettbewerb von Morgen," *Jahrbuch 2000 der Fachinformationen* 2000.
- [11] C. Shapiro and H.R. Varian "Information Rules," 1998.
- [12] E. Fritz "Cross-Media: Ein Inhalt aber mehrere Darstellungsformen," *Deutscher Drucker* no. 7, pp. w31-w32, 2000.
- [13] C. Fouchard "Eine ODBMS-basierte Lösung für "Cross-Media-Publishing"," *Objekt Spektrum* no. 3, pp. 91-95, 2000.
- [14] B. Oakly, D. Kueter and K. O'Heas "The Future of Content," *Discussions on the future of European Electronic Publishing* 1997.

- [15] H.-J. Bullinger, E. Schuster and S. Wilhelm "Content Management Systeme," 2000.
- [16] M. Schumann and T. Hess "Grundlagen der Medienwirtschaft," 2000.
- [17] J. Pine "Mass customizing products and services," *Planung Review* vol. 21, no. 4, pp. 6-13, 1993.
- [18] J. Göpfert and M. Steinbracher "Modulare Produktentwicklung leistet mehr," *Harvard Business Manager* no. 3, pp. 20-29, 2000.
- [19] N. Fuhr "Information Retrieval - Skriptum zur Vorlesung," 1998.
- [20] R. Baeza-Yates and B. Ribeiro-Neto "Modern Information Retrieval," 1999.
- [21] C.v. Rijsbergen "Information Retrieval," 1979.
- [22] T. Ritz "Modelling Mass Customization of digital news services," *e-Business Strategy Management* vol. 3, no. 3, 2002.
- [23] W. Dörfler and W. Peschek "Einführung in die Mathematik für Informatiker," 1988.
- [24] R. Peters "Business-Objekte, Workflow und die UML," *OBJEKTSpektrum* vol. 3, pp. 69-73, 1999.
- [25] J. Sturm "VB 6 UML Design and Development," 1999.

Adaptive Special Reports for On-line Newspapers

Sébastien Iksal and Serge Garlatti

Department of Artificial Intelligence and Cognitive Sciences
ENST Bretagne - Technopôle de Brest Iroise
B.P. 832, 29285 Brest Cedex, France
{sebastien.iksal, serge.garlatti}@enst-bretagne.fr

Abstract. Internet may bring new opportunities for electronic documents and Press agencies. Numbers of daily newspapers in the world propose their electronic version. The articles may be published in very different contexts which requires to be able to mix different sources, to provide different presentations, etc. Then, it is necessary to ensure reusability, sharing and exchange on the internet/intranet. Semantic web initiative can be an opportunity for on-line newspapers, news repositories or portals. Personalization/adaptation is an important issue in the semantic web. Indeed, adaptive web services have the ability to deal with different users' needs for enhancing usability and comprehension and for dealing with large repositories. Nowadays, it is not sufficient for a newspaper to provide raw information (news wires of Press Agencies). One of the most important task of journalists is to select, synthesize and analyze information and events for his readers. Special reports seem to be the most representative journalists' task. A special report offers news as well as analysis, debate, synthesis and/or development. A digital special report can be considered as an adaptive virtual document. The main goal of ICCARS is to assist the journalist in creating these adaptive special reports.

Keywords. Adaptive navigation, Adaptive presentation, Press, Semantic Composition, Virtual Document

1 Introduction

Press institutions on television, radio as well as the printed Press have web services, news repositories and/or portals. Some daily newspapers propose their "printed" edition and the digital one at the same time (Le Monde [1], Le Télégramme [2] ...). Others like monthly magazines differ their editions (Linux Magazine [3]). The high availability of Internet modifies the organization of newspaper's offices, as well as the Press behavior. Now journalists work with electronic mail, chat, search engines ..., and use Internet as a way for accessing information and getting contact with Press agencies. Numbers of daily newspapers in the world propose an electronic version. A lot of Web users (individuals, corporates, sme's, administrations, ...) are interested in

on-line newspapers and news repositories. So it's easy to understand that Internet may bring new opportunities for electronic documents and Press agencies. Most of Press agencies have to or would like to retrieve and/or sell or buy articles. These articles may be published in very different contexts which requires to be able to mix different sources, to provide different layouts, etc. Then, it is necessary to ensure reusability, sharing and exchange on the internet/intranet, and these features require to have a precise search engine. Indeed, it is well known that keyword-based information access presents severe limitations concerning precision and recall. On the contrary, intelligent search engines, relying on semantic web initiative [4] and semantic metadata, overcome these limitations [5, 6]. Semantic web initiative can be an opportunity for on-line newspapers, news repositories or portals.

Nevertheless, information space is so huge that it is not sufficient to have a precise search engine. It is necessary to take into account user interests – at least – to be sure to focus on relevant pieces of information. Personalization/adaptation is an important issue in the semantic web, but also for electronic documents. Some web sites personalize the access to information and others the search engine. Internet increased the need to satisfy the reader, that is why numbers of sites provide personalized services. Adaptation/personalization is one of the main issues for web services. But, it is not limited to filtering processes. Indeed, adaptive web services have the ability to deal with different users' needs for enhancing usability and comprehension and for dealing with large repositories. Indeed, adaptive web applications - also often called Adaptive Hypermedia Systems - can provide different kinds of information, different layouts, different navigation tools according to users' needs [7]. Creating adaptive web services from news repositories or portals requires the following features: i) methods to facilitate web application creation and management and ii) reuse, sharing and exchange of data through the internet/intranet. The notion of flexible hypermedia and more particularly that of virtual documents can lead to methods facilitating web application design and maintenance. According to Watters, "A virtual document is a document for which no persistent state exists and for which some or all each instance is generated at run time" [8]. Virtual documents have grown out of a need for interactivity and individualization of documents, particularly on the web. Virtual document and adaptive hypermedia are closely related – they can be viewed as the two faces of the same coin.

Nowadays, it is not sufficient for a newspaper to provide raw information (news wires of Press Agencies). One of the most important task of journalists is to select, synthesize and analyze information and events for his readers. In such framework, special reports seem to be the most representative journalists' task. A special report offers news as well as analysis, debate, synthesis and/or development. It can be viewed as an organized collection of articles offering a viewpoint on events. We consider the digital special reports as adaptive virtual documents. We are interested in adaptive virtual documents for author-oriented and reader-oriented web services providing several reading strategies to readers. In this paper, we focus on organizations called narrative structure for author-oriented reading strategies. An author-oriented reading strategy have the following characteristics: authors have know-how which enables them to choose special report contents and to organize them in one or more consistent ways – author reading strategies.

First of all, adaptive special reports by means of the ICCARS Project are presented. Secondly, we will show why it is interesting to manage the different views of the special report separately. Thirdly, the adaptation will be analyzed via our adaptive semantic composition engine. Finally, we will conclude by some perspectives.

2 Adaptive Special Reports

ICCARS is the acronym for Integrated and Cooperative Computer Assisted Reporting System. It is a joined project between the IASC Laboratory, a SME called Atlantide and a regional daily newspaper called Le Télégramme. It is funded by Brittany Regional Council. The ICCARS prototype is a computer assisted reporting system. Its main goal is to assist the journalist in creating adaptive special reports. These documents are able to include audio and video material, links, and they are no longer limited in size.

Internet increased the need to satisfy the reader, that is why numbers of sites provide personalized services. Someone provide all the most interesting news according to your preferences through e-mail such as e-revue [9]. For a web site, an interesting solution is offered by Crayon [10] which assist the reader in organizing his own newspaper (it is possible to name it like “The MyNewsPaper Post” or “The MyNewsPaper Tribune”). The internet reader is able to modify his newspaper and to select who is allowed to read it. But it has been made by the reader, which is very limited. We need to have automatic or semi-automatic processes able to filter information space for readers. A lot of Web sites propose to personalize the access and the layout of the information written for the printed newspaper. Two projects work with personalized news which can be read through a Web site. Sistemi Telematici Adattativi [11] is a project which propose to filter and display news and ads according to user’s preferences and characteristics. KMI Planet [12] is a kind of private on-line newspaper where all readers and writers are in a same group. It collects news through e-mail, processes and sends the result to the most interested readers. The tool is able to sort articles in order to fill in gaps, and after to inform the reader when the news is ready. It offers also an advanced interface for searching documents.

During a long time, local Press agencies were the main local news providers, but with Internet, new actors like “city-guides” propose local news. At the beginning, city-guides aimed at supplying information about public services, classified, association, weather ... Today, they have their own team of journalists [13] and propose national and local news. Then, the local newspapers created their own city-guides such as <http://www.vivabrest.com> for Le Télégramme. National Press agencies are also concerned by the phenomenon because some sites such as Internet providers propose classified [14]. They receive all news wire from agencies like AFP (Agence France Presse), Reuters. The main challenge for Press agencies is not to only be information providers but also to offer new services on the web. The printed Press loses numbers of readers, so Internet is a new medium useful for increasing their readership and in fact their income with advertising, e-business ... The Internet user is able to use various search engines to collect information. Nevertheless, this set of data needs to be

analyzed and synthesized. Due to internet features, numerous web sites proposed electronic special reports which are composed of a set of articles. Most of the time, they don't provide a relevant organization of these articles, sometimes they don't provide an organization. The structure proposed actually is a classification by the date of publishing and sometimes articles are grouped inside various headings. Our main objective is to propose various organizations for a same special report in order to increase the comprehension. It's possible to offer personalized organizations for digital special reports by considering virtual documents. We consider the digital special reports as adaptive virtual documents, we define them as follows:

- An adaptive/personalized virtual document consists of a set of information fragments, ontologies and a semantic composition engine which is able to select the relevant information fragments, to assemble and to organize them according to an author's strategy or reader's goals by adapting various visible aspects of the document delivered to the reader.

We are interested in adaptive virtual documents for author-oriented and reader-oriented web services providing several reading strategies to readers. An author-oriented reading strategy have the following characteristics: authors have know-how which enables them to choose special reports contents and to organize them in one or more consistent ways – author reading strategies. A reader-oriented strategy is an overall document structure computed from reader's goals. For instance, it can be based on geographic, history or topic criteria – a domain model – or a task model organizing the access to articles. Nevertheless, journalists have to be aware of such structures because they associate metadata with articles and special reports for these services. In this paper, we focus on organizations called narrative structure for author-oriented reading strategies. The reader has the ability to recognize – sometimes unconsciously – these structures. For instance, scientific papers, courseware, report, special report in journalism, etc., have each of them a distinct narrative structure. At present, the narrative structure is implicit in printed document, but also in digital one. Such author's know-how and skills can be represented at knowledge level and then be shared and reused among authors, used for generating web documents dynamically and for enhancing reader comprehension. A narrative structure provides an overall document structure which is a declarative description of web documents which offers a particular view on a special report. In electronic documents, there are different views which can coexist.

3 Special Report Views

In a digital document, three different views coexist: semantic, logical and layout [15]. For each view we have a specific structure. The semantic structure of a document conveys the organization of the meaning of the content of a document. This view fits the semantic level of the semantic web architecture. Indeed, it can be represented by ontologies. Ontologies are used to model types of fragment as well as their relationships. The overall document structure modeled from an ontology is a narrative structure designed by a writer for presenting a particular angle on a set of articles. A narra-

tive structure is composed of nodes and semantic relationships. Nodes are spans of texts. Relationships belong to those analyzed by Rhetorical Structure Theory (RST) [16]. RST defines relations between spans of text, each span have a role inside the relation (nucleus and satellite). Each relation is defined by some constraints on the nucleus, the satellite, the combination of the nucleus and the satellite, and an effect to the reader. Among these relations, we can find are antithesis, restatement, summary, interpretation and so on. For instance, “The fragment A which is an interview is the volitional cause of the fragment B which is an analysis”, the underlying relationship cannot be represented by a syntactic structure [17]. Interview and analysis are types of fragment. The interview is the satellite and the analysis is the nucleus of this rhetorical relation [16]. This relation is oriented and encode a particular reading guide. In this case, the fragment B will be better understood if the fragment A is read before. It could be interesting to show the type of relation to the reader as explanations or for increasing the comprehension.

The logical structure reflects the syntactic organization of a document. A document (for example books and magazines) can be broken down into components (chapters and articles). These can also be broken down into components (titles, paragraphs, figures and so forth). It turns out that just about every document can be viewed this way. The logical view fits the syntactic level of the semantic web architecture. A logical structure can be encoded in XML [18]. The layout view describes how the documents appear on a device and a layout structure describes it, (e.g. the size and color of headings, texts, etc). The layout view may be processed by an XSLT processor [19] for transforming an XML document into an HTML document that can be viewed by any web browser. It can also be processed by a java engine able to compute an XML document for presenting by a web browser.

In a printed document, these three views are intertwined and are not separable. There is no straightforward mapping between the semantic and the logical structure, that is to say, for instance, a paragraph does not correspond to a particular content’s meaning. On the other hand, the logical and layout structure are closely related. Indeed, the layout structure encodes the logical structure. For instance, each section element has a particular presentation – font, size, color, etc. The semantic structure is implicit and so it can be analyzed and/or recognized by a reader. Moreover, it is a key issue for reader comprehension. In a digital document, these three views may be represented and managed.

A special report model can be computed on the fly by means of a semantic composition engine using: i) an overall document structure – a narrative structure - representing a reading strategy for which node contents are substituted at run time, according to reader’s needs for adaptation, ii) an intelligent search engine, iii) semantic metadata, and iv) a reader model. This semantic composition engine architecture relies on these three views. Each special report model is computed when it is necessary, we don’t store the delivered reports. An authoring tool is provided for creating narrative structures, specifying their content and associating metadata.

4 Adaptive semantic composition engine

The semantic composition engine allows adaptive presentation and navigation. A reader chooses a particular special report and a corresponding reading strategy. Then, the system computes on the fly an adapted special report for this reader, web pages and their layouts. First of all, we present the semantic composition engine architecture. Secondly, our assumptions and design criteria are detailed. Finally, the adaptation process is described.

4.1 Semantic Composition Engine Architecture

Our semantic composition engine relies on OntoBroker for ontology management and intelligent search engine. OntoBroker is a knowledge management engine which is useful for filtering and information retrieval in a large amount of data as well as in the model specification – ontologies [6, 20, 21]. OntoBroker contains four ontologies [22] and facts closely related to them. These ontologies are: a domain ontology for representing contents, a metadata ontology at the information level which describes the indexing structure of fragments, a user ontology which may define different stereotypes and individual features and a special report ontology which represents the author's competences and know-how for creating special report models [23]. The domain ontology defines a shared vocabulary used in the metadata schema for the content description of data. It will also be used by the semantic composition engine as an overall document structure, by the user as an information retrieval tool because the user often has difficulty in defining his/her interests, and it is easier for him/her to recognize required information in a domain model than to specify it.

According to the three views of a document, our semantic composition engine architecture is described below (cf. fig. 1). One of the main ideas behind the notion of semantic composition engine is to declare as much as possible all the reader's tasks and interactions. The semantic composition engine is composed of three different stages: a semantic composition which manages the semantic structure of a special report model for defining a reader adapted special report and selects its contents, a logical composition which computes an XML web page from the reader adapted special report and a layout composition which computes the current web page layout from the XML structure. This architecture is based on two different studies: ICCARS Project and CANDLE Project [24] (Collaborative and Network Distributed Learning Environment) which is an European project.

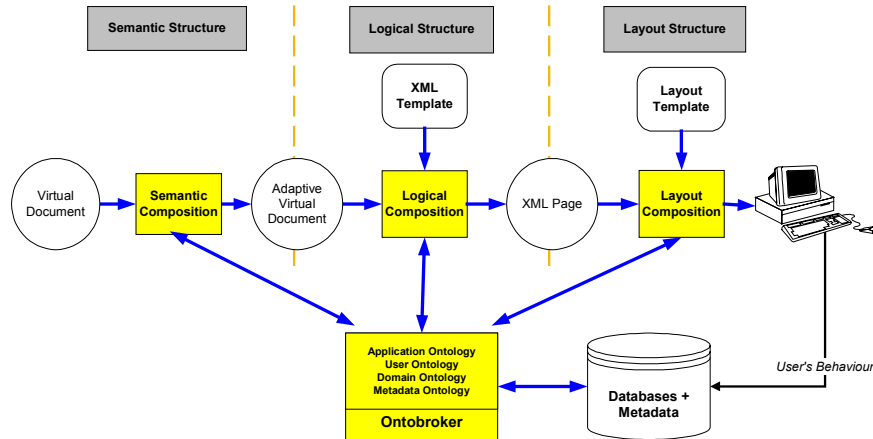


Fig. 1. The Semantic Composition Engine Architecture

4.2 Assumptions and Design Criteria

A special report model is composed of an information space - a set of fragments - and at least one narrative structure. We assume the structure is a directed acyclic graph. Each edge has a particular type which is a relation taken in the Rhetorical Structure Theory. Each node contains a specification which is used by an information retrieval process to find all relevant fragments. Fragments can be atomic or abstract information units. The latter are composed of atomic and abstract information units. Articles are atomic fragments and sub-reports are abstract fragments. A special report and corresponding reading strategies are modeled as follows in figure 2.

A sub-report is composed of a set of articles selected by the author – explicitly associated with it to define its relevant information space -, and one or more narrative structures – reading strategies. A sub-report can be organized according to one or more structures. A structure is a collection of components among which one is the root of the structure. A component is an abstract object, which exists only inside a particular structure. A component is linked to others through a semantic relation belonging to those of RST. This relationship gives the organization of the structure. That is to say, each component in the structure which is the source of a relationship, is a nucleus in RST and the corresponding destination (a component also) is a satellite. So, we use RST as a basis to build a narrative structure in which nodes are different categories of fragments. A component is a kind of information retrieval service which uses a description given by the author according to a subset of the metadata schema, to send a query to the intelligent information broker. The outcome of this service can be one or several atomic fragments – articles -, one or several sub-reports or both. The special report model is an input for the semantic composition engine which computes an adapted/personalized special report for a given reader.

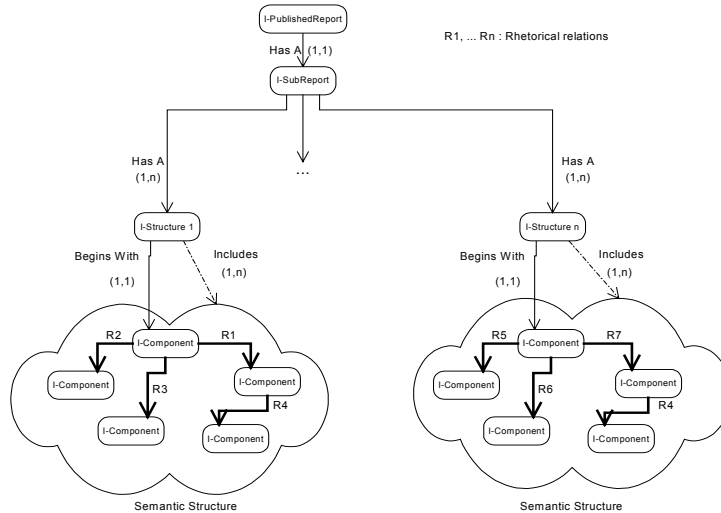


Fig. 2. A special report model

4.2.1 Adaptation Policies

Adaptation policies falls into two main categories: adaptive presentation and adaptive navigation [7]. The idea of various adaptive presentation techniques is to adapt the content of a page accessed by a particular user to current knowledge, goals and other characteristics of the user. And, the idea of adaptive navigation support techniques is to help users to find their paths in hyperspace by adapting the way of presenting links to goals, knowledge and other characteristics of an individual user.

The system manages adaptive presentation and five adaptive navigation methods [7, 25]:

- *Annotation* is a technique that presents differently each link according to the result of an evaluation (of the document pointed by the link). It is possible to use colors or pictures in order to differentiate the links.
- *Direct Guidance* means that the system determines which node is the next “best” node for the user to visit. What is best depends on the user’s knowledge and goal, both of which should be represented in the user model.
- *Sorting* is a technique that sorts all links according to their degree of relevance for the user. The system may use the user model and some user-valuable criteria, it will place the more relevant link at the top of the list.
- *Hiding* is a technique that displays only links which are the most relevant for the user.
- *Partial Hiding* is a technique that displays links which have a degree of relevance included in a particular interval.

An author can associate an adaptive navigation method to a special report model. Indeed, he can specify the methods available for a given reader stereotype. By default,

all methods are available for all readers. Otherwise, a reader stereotype is associated to each adaptive navigation method. For a given reader, methods for which the corresponding stereotypes are matched by his model are available. The reader stereotype is defined by a Boolean expression which is composed of reader's characteristics taken in the reader model. For instance, the author can say : "Annotation is possible for adults who works in the fishing industry". By means of the reader model, the reader can give his preferred adaptive navigation methods. But, author's constraints have priority over reader's preferences.

4.2.2 Adaptive Presentation

For adaptive presentation, we have to consider the special report model and the information retrieval service included in each component and then to take into account the adaptive navigation process.

Let A, B be components and R1 be a link from A to B. As soon as B has several fragments as an outcome of the information retrieval process, the link R1 is considered as several links of the same type, one per fragments from the source A to each fragment. A fragment has a single state and the various possible states are used by the adaptive navigation processes in order to manage links.

Some adaptive navigation methods like hiding, partial hiding or direct guidance, if they are the only methods available, allows the filtering and the removal of the irrelevant fragments and links. These methods have direct consequences on the special report content and structure and then on adaptive presentation.

4.2.3 Adaptive Navigation

The five adaptive navigation methods are based on the relevance states of fragments. It is possible to define up to five states (Very Good, Good, Interesting, Bad, Very Bad) which are ordered and mutually exclusive. A Boolean expression is associated to each state by an author. When the current fragment fits an expression, the related state is given to it. The Boolean expression uses some features of the fragment metadata and of the reader model (for instance his working area, his age or the knowledge model which is useful for counting the number of known concepts). Adaptive navigation methods are managed like this :

- *Annotation* : The system is able to associate at links a different picture or color according to the state of the fragment (which is the link destination).
- *Direct Guidance* : The system highlights the (or all the) most relevant link(s) according to the state of the fragment destination.
- *Sorting* : The system sorts all links according to the state of each fragments (all states are ordered).
- *Hiding* : The system keeps only the links corresponding to the fragments with the best state.
- *Partial Hiding* : The author chooses the list of states to keep, and the system will remove the links (and the fragments) through the fragments which are evaluated with the other states.

These features are useful for filtering or ordering all possible organizations of the special report.

4.3 Adaptation Process

The architecture of the system is closed to the three views of a document. Each composition level has its own adaptation features. The generation of the adapted special report begins with the management of the node content at the semantic level, next the logical level manages the adaptive navigation method fixed by the author or preferred by the reader and finally the layout level applies the adaptive navigation method on web pages.

4.3.1 At the semantic level

For an author's reading strategy, the main role of the semantic composition is to define the adapted special report content and to adapt the chosen structure to reader needs. Indeed, each node (component) has only a content specification. From this specification, one or more fragments may be selected from the relevant information space associated to the considered special report model. Indeed, only a subset of metadata entries are used for content specification by the authoring tool. The others are used for defining variants of fragments - according to adaptation policies. Our approach is very closed to the explanation variants of Brusilovsky [7]. With explanation variants, the page variants or fragment variants are clearly identified and the system choose one or another according to a stereotype. In our case the fragments contained in a node are selected from an information retrieval process. Then, in a same node we can only find fragment variants and the system can choose between them. Fragment variants in a node can be articles or sub-reports, it increases the richness of the content because the system can propose a sub-report to an expert of the domain and only an article for a reader who is in a hurry.

After the retrieval of nodes content, all the fragments obtained are evaluated with the authors rules. The system will associate a state to each fragment in order to filter the set of fragments at the other levels. When hiding, partial hiding and direct guidance are the only available methods, the system will manage adaptive presentation, that is to say adapting the content and the structure. Indeed, these adaptive navigation methods don't take into account some fragments (which have an irrelevant state) and the corresponding links. So, the system can remove these fragments and links. The structure is also modified if a node contains only a sub-report, because the structure of the sub-report is added to those of the special report model. Finally, the result of this composition process is a reader adapted special report.

4.3.2 At the logical level

The logical composition aims at computing the current web page structure - XML - with a content and navigation tools for accessing the rest of the adapted special report. A web page, represented as an XML structure, is generated from a particular template according to the reader task - in our framework reading a special report. A template describes the logical structure of a web page but without any content or navigation tools. It contains queries for computing navigation tools and for loading the content via OntoBroker. The content is given by the selected fragment in the current node of the adapted special report. The navigation tools depend on the selected adaptive navi-

gation method. For the adaptive navigation method, author's constraints have priority over reader's preferences. Adaptive navigation methods which are available are filtered according to the author's constraints and/or the reader model. Next, the reader model is used to select the preferred adaptive navigation method. For defining the hyperlinks in navigation tools, the logical composition engine has to browse the adapted special report. It has also to associate properties to hyperlinks for managing annotation, hiding, sorting and direct guidance. These properties come from the relevance states of fragments contained in nodes.

4.3.3 At the layout level

Finally, the layout composition has to map some presentation rules on the web page. The final process of this architecture is concerned by the design of the web pages of an adapted special report. The final layout of each page may be tailored to the reader's preferences: print size, color, and so on and/or use standardized styles from corporate, SMEs or institution style sheets. The layout composition has also to manage the adaptive navigation. From author specification or reader stereotypes or reader preferences, he has to hide, to annotate, etc; the different types of hyperlinks in a web page. There is a style sheet for each template. At this stage, the system will consider pictures or colors for showing the different states. It will interpret the XML page given by the logical composition process. At the end of this process, the reader has his/her HTML Page on his browser.

5 Conclusion and Perspectives

In this paper, we have presented our framework which consists in delivering adaptive special report according to an author-oriented viewpoint. Authors have know-how which enables them to choose special reports contents and to organize them in one or more consistent ways by means of narrative structures. Authors can share and reuse these narrative structures. A cognitive approach, a knowledge elicitation method based on verbal protocols, was used to acquire journalist's skills and know-how [26].

An adaptive semantic composition engine has also been presented. According to the three different views of a special report, different types of adaptation may be applied. At the semantic level, adaptive presentation and navigation can be applied. Indeed, there are closely related because adaptive presentation is mainly based on fragment variants which are very closed to the explanation variants of Brusilovsky [7]. Fragments are selected by an information retrieval process using a subset of metadata features. Some others metadata are dedicated to the specification of fragment variants. According to the current adaptive navigation methods – hiding for instance, adaptive link removal can be applied because the related fragments are not relevant. The adaptive semantic composition engine is able to manage up to five adaptive navigation methods and five states for annotation, to define adaptation policies. At the logical level, different XML page templates can be defined in order to provide different services to readers. The logical level is also able to manage the different adaptive navigation methods according to author constraints and reader preferences. Such an ap-

proach could be reused in very different areas. For instance, we are applying this approach in the CANDLE European project about distance learning. Indeed, it seems to be very convenient to specify different adaptation policies for different categories of learners.

We plan to offer a kind of free browsing mode which will use a narrative structure as a guide. In other words, the intelligent search engine will not be limited to the information space dedicated to the special report model. Indeed, the content specification of each component will be applied to the entire database. A reader will be able to access all articles fitting the different content specifications and then to get articles closely related to the current component. The notion of special report has to be refined and extended in some way. Indeed, corporates or institutions are interested in different categories of articles. For instance, the Télégramme is selling articles by email to different institutions. Then, we can apply our approach to provide special reports according to different corporates or institutions needs. These special reports could be updated automatically or semi-automatically. And they could be based on readers' strategies according to reader's goals for instance and then could be processed and updated automatically.

References

1. Le Monde, <http://www.lemonde.fr>.
2. Le Télégramme, <http://www.letelegramme.com>.
3. Linux magazine, <http://www.linuxmag-france.org>.
4. Berners-lee, T., *Weaving the Web*. 1999, San Francisco: Harper.
5. Decker, S., et al., *Knowledge Representation on the Web*. 2000, On to Knowledge Project: <http://www.ontoknowledge.org/oil/papers.shtml>.
6. Decker, S., et al. *Ontobroker: Ontology Based Access to Distributed and Semi-Structured Information*. in *Conference on Database Semantics*. 1999. Rotorua, New Zealand: Kluwer Academic Publishers.
7. Brusilovsky, P., *Methods and techniques of adaptive hypermedia*. *User Modeling and User-Adapted Interaction*, 1996. **6**(2-3): p. 87-129.
8. Watters, C. and M. Shepherd, *Research Issues for Virtual Documents*, in *Proceedings of the Workshop on Virtual Documents, Hypertext Functionality and the Web*. 1999, Eighth International World Wide Web Conference: Toronto, Canada.
9. E-revue, <http://www.e-revue.com>.
10. Crayon, <http://www.crayon.net>.
11. Ardissono, L., L. Console, and I. Torre. *Exploiting user models for personalizing news presentations*. in *2nd Workshop on Adaptive Systems and User Modeling on the WWW, AH'99 and UM'99*. 1999: Eindhoven University of Technology, Computer Science Reports.
12. Domingue, J. and E. Motta. *A Knowledge-Based News Server Supporting Ontology-Driven Story Enrichment and Knowledge Retrieval*. in *11th European Workshop on Knowledge Acquisition, Modelling, and Management (EKAW '99)*. 1999.
13. WebCity, <http://www.webcity.com>.
14. Tiscali, <http://actu.tiscali.fr>.
15. Christophides, V., *Electronic Document Management Systems*. 1998, UCH/FORTH: <http://www.ics.forth.gr/~christop/>.

16. Mann, W.C. and S.A. Thompson, *Rhetorical Structure Theory: Toward a functional theory of text organization*. Text, 1988. **8**(3): p. 243-281.
17. Decker, S., et al., *The Semantic Web - on the respective Roles of XML and RDF*. 1999. <http://www.ontoknowledge.org/oil>.
18. Bray, T., et al., *Extensible Markup language (XML) 1.0, (Second Edition)*. 1998, W3C: <http://www.w3.org/TR/2000/REC-xml-20001006>.
19. Adler, S., et al., *Extensible Stylesheet Language (XSL) Version 1.0*. 2000, W3C: <http://www.w3.org/TR/xsl/>.
20. Fensel, D., et al. *On2broker: Semantic-Based Access to Information Sources at the WWW*. in *World Conference on the WWW and Internet, WebNet 99*. 1999. Honolulu, Hawaii, USA.
21. Fensel, D., et al. *On2broker in a Nutshell*. in *the 8th World Wide Web Conference*. 1999. Toronto.
22. Iksal, S. and S. Garlatti. *Documents Virtuels et Composition Sémantique : Une Architecture Fondée sur des Ontologies*. in *NîmesTIC*. 2001. Nîmes, France.
23. Iksal, S. and S. Garlatti. *Revisiting and Versioning in Virtual Special Reports*. in *Workshop Adaptive Hypermedia in Hypertext 2001*. 2001. Aarhus, Denmark.
24. Candle, <http://candle.eu.org>.
25. De Bra, P. and L. Calvi, *AHA! An open Adaptive Hypermedia Architecture*, in *The New Review of Hypermedia and Multimedia*. 1998, Taylor Graham. p. 115-139.
26. Iksal, S. and S. Garlatti. *Semantic Composition of Special Reports on the Web : A Cognitive Approach*. in *H2PTM'01 - Hypertextes et Hypermédia 2001*. 2001. Valenciennes, France.

Cross-References in Web-Based Adaptive Hypermedia

Hongjing Wu, Erik de Kort

Department of Mathematics and Computing Science
Eindhoven University of Technology
PO Box 513, 5600 MB Eindhoven
the Netherlands

email: h.wu@tue.nl, erik.de.kort@asml.nl

Abstract: Many websites offer their users a lot of freedom to navigate through a large hyperspace. Adaptive hypermedia systems (or AHS for short) aim at overcoming possible *navigation and comprehension problems* by providing adaptive navigation support and adaptive content. The adaptation is based on a *user model* that represents relevant aspects about the user [2]. In most systems navigation support is tied to the existing link structure. To provide users with better understandable navigation, we discuss adaptive cross-references based on concept relationships in AHAM. We define an authoring tool to generate a cross-reference page for each page based on already defined relationships. With these added cross-reference pages, we can easily provide adaptive cross-references according to user features by AHAM.

Keywords: adaptive hypermedia, user modeling, navigation support, hypermedia reference model, adaptation rules.

1. INTRODUCTION

The introduction of World Wide Web has made hypermedia the preferred paradigm for user-driven access to information. Websites typically offer users a lot of freedom to navigate through a large hyperspace. Unfortunately, this rich link structure of hypermedia applications may cause some usability problems:

- A typical hypermedia system presents the same links on a page, regardless of the path a user has followed to reach this page. Even when providing navigational help, e.g. through a map (or some fish-eye view) the system does not know which part of the link structure is most important for the user. The map can thus not be simplified by filtering (or graying out) links that are less relevant for the user. Having personalized links or maps would eliminate some *navigation problems* that users have with hypermedia applications.
- Navigation in ways the author did not anticipate also causes *comprehension problems* for the user: for every page the author must take into account what foreknowledge the user has when accessing that page. In order to do so the author must at least consider all possible paths that lead to the current page. This is clearly an impossible authoring task because there are more ways to reach a page than any (human) author can foresee. In a traditional hypermedia application a page is always presented in the same way. This may result in users visiting pages

containing redundant information and pages that they cannot fully understand because they lack some expected foreknowledge. The website should really select the pieces of information that are shown on a page, based on a history of which pages the user has seen before, or provide easy access to extra information on the concepts presented in the page.

Adaptive hypermedia systems (AHS) in general, and adaptive websites in particular aim at overcoming the navigation and comprehension problems by providing *adaptive navigation support* and *adaptive content*. The adaptation (or personalization) is based on a user model that represents relevant aspects of the user such as preferences, knowledge and interests. We focus on simple Web-based systems that can only gather information about the user by observing the *browsing* behavior of that user. Each time the user “clicks” on a link the system uses the selected link to update the user model and to adapt the presentation accordingly.

In previous research we have shown that our AHAM model [2] can be used to describe content adaptation and link adaptation of AHS. Link adaptation as used in many AHS guides users towards interesting information by changing the presentation of link anchors. In general however, adaptation of the existing link structure alone is not enough to solve all users’ navigation and orientation problems:

- It may not be possible to select a guided tour consisting of existing links, and of only pages that are interested for a given user, because the interesting pages may not form a connected sub graph in the whole link structure.
- It may take too many steps (possibly going through uninteresting pages) to guide the user to the page(s) that deal with the topic the user is interested in.

To improve adaptive navigation support based on basic link structure in AHS, we propose *cross-reference navigation support* to supplement the above cases. Cross-reference navigation support (or CRNS for short) aims to help the user when s/he navigates through hyperspace following cross-references, i.e. conceptually related information. A cross-reference page provides meta-information for a “normal” page in the form of links, and can thus be adapted according to user features.

While cross-references could be investigated at a more general level, we concentrate on its use in AHS that can be described in the AHAM model. This paper is structured as follows: in Section 2 we briefly review the AHAM reference model, thereby concentrating on the parts that are needed to describe adaptation functionality at an abstract level. Section 3 discusses how to generate cross-reference pages and how to provide cross-reference navigation support based on these pages. Section 4 draws conclusions and describes our future work.

2. AHAM, A DEXTER-BASED REFERENCE MODEL

Many adaptive hypermedia systems share parts of their architecture. Just like the Dexter model [4][5], tried to capture the facilities offered by hypermedia systems of its time (and of potential future systems), AHAM [2], (for Adaptive Hypermedia Application Model) describes the common architecture of adaptive hypermedia systems. Part of this common architecture is typical for Web applications: their event-driven nature, where each page access results in a user-model update and an adaptive

nature, where each page access results in a user-model update and an adaptive presentation. AHAM's overall structure is an extension of the Dexter model. According to AHAM each adaptive hypermedia application is based on three main parts:

- The application must be based on a *domain model*, describing how the information content of the application or "hyper-document" is structured (using concepts and pages).
- The system must construct and maintain a fine-grained *user model* that represents a user's preferences, knowledge, goals, navigation history and other relevant aspects.
- The system must be able to adapt the presentation (of both content and link structure) to the reading and navigation style the user prefers and to the user's knowledge level. In order to do so the author must provide an *adaptation model* consisting of *adaptation rules*. An AHS itself may offer built-in rules for common adaptation aspects. This reduces the author's task of providing such rules. In fact, many AHS do not offer an adaptation rule language; the way in which the user model is updated and the presentation adapted is then completely predefined.

The division into a *domain model* (DM), *user model* (UM) and *adaptation model* (AM) provides a clear separation of concerns when developing an adaptive hypermedia application. The main shortcoming in many current AHS is that these three factors or components are not clearly separated. Modeling an existing AHS in AHAM may not be straightforward because AHAM requires these parts to be made explicit, and the adaptive behavior to be described using adaptation rules. However, using AHAM enables us to clearly describe how an AHS works, how different AHS compare, and also how to design new and more powerful AHS.

The AHS consists not only of the three sub-models (mentioned above) but also of an *adaptation engine* (AE). The AE describes implementation dependent aspects of the AHS. In previous work [7] we described design issues for a general-purpose AE, and defined AHAM CA-rules to illustrate how sets of rules work together. We will use these rules to describe the generation of adaptive cross-reference in Section 3.

Figure 1 shows the overall structure of an adaptive hypermedia application in the AHAM model. The figure has been made to resemble the architecture of a hypermedia application as expressed in the Dexter Model [4][5].

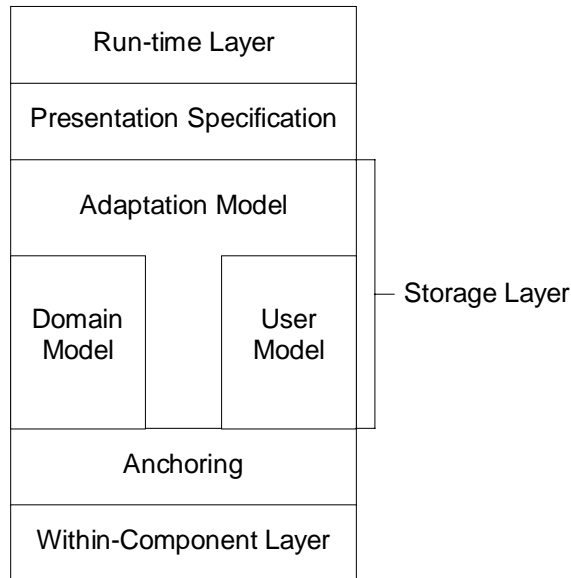


Figure 1 : Structure of adaptive hypermedia applications

In this section we only present the elements of AHAM that we will need to discuss how cross-reference navigation support (CRNS) can be expressed in AHAM.

2.1 The Domain Model

The domain model of an adaptive hypermedia application consists of *concepts* and *concept relationships*. Concepts are objects with a unique object identifier, and a structure that includes attribute-value pairs and a sequence of anchors.

A concept represents an abstract information item from the application domain. It can be either an atomic concept or a composite concept.

- An *atomic concept* corresponds to a fragment of information. It is primitive in the model (and can thus not be adapted). Its attribute and anchor values belong to the “Within-Component Layer” and are thus implementation dependent and not described in the model.
- A *composite concept* has a sequence of children (sub-concepts) and a constructor function that describes how the children belong together. The children of a composite concept are either all atomic concepts or all composite concepts. A composite concept with (only) atomic children is called a *page*. The other (higher-level) concepts are called *abstract concepts*.

The composite concept hierarchy must form a DAG (directed acyclic graph). Also, every atomic concept must be included in one or more composite concepts.

Figure 2 illustrates a part of a concept hierarchy.

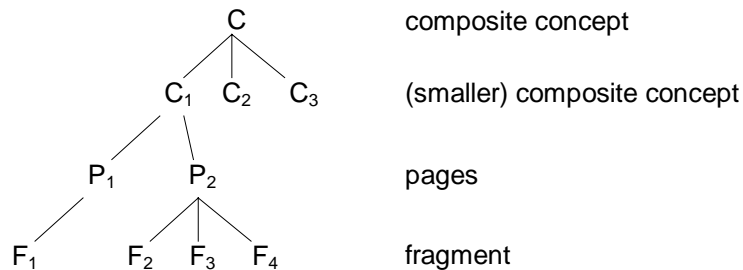


Figure 2 : Part of a concept hierarchy.

A *concept relationship* is an object (with a unique identifier and attribute-value pairs) that relates a sequence of two or more concepts. Each concept relationship has a *type*. The most common type is the hypertext **link**. In AHAM we consider other types of relationships (abstract relationships) as well, which play a role in the adaptation, e.g. the type **prerequisite**. When a concept C_1 is a prerequisite for C_2 it means that the user “should” read C_1 before C_2 . This does not imply that there must be a link from C_1 to C_2 . It only means that the system somehow takes into account that reading about C_2 is not desired before some (enough) knowledge about C_1 has been acquired. Through link adaptation the “desirability” of a link will be made clear to the user. **Figure 3** shows a small set of (only binary) concepts associated to one another by three types of concept relationships: prerequisite, inhibit, and link.

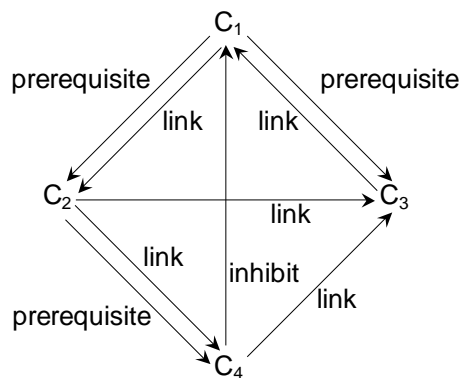


Figure 3 : Example concept relationship structure.

Apart from the “implicit” relationship type and set of relationships that form the concept hierarchy an AHS need not contain or support any other relationships. AHAM can thus also represent applications without traditional hypertext links (like e.g. in *spatial hypertext* [6]). A relationship graph defines a certain connection among a set of concepts. In Section 3 we will use relationship graphs to generate cross-reference pages; these can be used to supplement the basic adaptive navigation support in AHS.

The atomic concepts, composite concepts and concept relationships together form the *domain model* (DM) of an adaptive hypermedia application.

2.2 The User Model

A user model consists of named entities for which we store a number of attribute-value pairs. For each user the AHS maintains a *table-like structure*, in which for each concept in the DM the attribute values for that concept are stored. Because there can be many relationships between *abstract concepts* and *concrete* content elements like fragments and pages, a user model may contain many attributes per concept to indicate how the user relates to the concept. Typical attributes would be *knowledge level* (e.g. in educational applications) and *interest* (e.g. in encyclopedia, museum sites, or on-line mail-order catalogs). The user model may also store information about what a user has read about a concept, and for how long or how long ago, etc. Concepts can furthermore be used (some might say abused) to represent “global” user aspects such as preferences, goals, background, hyperspace experience, or a (stereotypical) classification like student, employee, visitor, etc. For the AHS or the AHAM model the true meaning of concepts is irrelevant.

In the sequel we will always consider UM as being the user model for a single user; we do not discuss adaptation to group behavior.

2.3 The Adaptation Model

The AHAM model targets adaptive hypermedia applications that follow the *request-response* paradigm that is typical for the Web. The interaction with the system is described through *events* generated by the user (or by an external source). Each event triggers user model updates and results in an adaptive presentation. In [7] we introduced a database-like language to express the effects of user actions as *condition-action rules* (AHAM CA-rules). This implies that we do not explicitly model *events* as events, but as updates to attributes that trigger rules. Accessing a web-page for instance will result in a Boolean “access” attribute of the page (in the user model) to become *true*. The small example below illustrates the structure of these rules. (A syntax description can be found in [7].)

```
C: select P.access  
A: update F.pres := “show”  
   where Fragment(P, F) and F.relevance = “recommended”
```

In this example we first see that the *condition* for this rule is that P.access has become true for some page P. When this happens (and because there is no additional **where** clause in the condition) the *action* is executed. In the action we look at fragments F of page P. If a fragment is marked as “recommended” then that fragment will be shown. This is indicated as a *presentation specification* and represented as a “pres” attribute of the fragment.

Note that the AHAM CA-rule language is just a vehicle for describing how an AHS should perform user model updates and adaptation. It does not imply that we require AHS to use such a language. Even when an AHS has only a built-in behavior, we can still describe this using the AHAM CA-rule language. Also, we partition adaptation rules into *phases* to indicate that certain rules must always be executed before certain other rules. The phases include IU, the initialization of the user model, UU-pre, the user model updates that are performed before generating the presentation, GA, the

generation of the adaptation, and UU-post, the user model updates that come after the presentation. The phases are a convenience for ensuring that the execution of the rules has desirable properties such as *termination* and *confluence*, as discussed in [7].

2.4 The Adaptation Engine

An AHS does not only have a domain model, user model and adaptation model, but also an adaptation engine, which is a software environment that performs the following functions:

- It offers generic page selectors and constructors. For each composite concept the corresponding selector is used to determine which page(s) to display when the user follows a link to that composite concept. For each page a constructor is used for building the adaptive presentation of that page (out of its fragments). Page constructors allow for dynamic content like a ranked list of links.
- It optionally offers a (very simple programming) language for describing new page selectors and constructors. For instance, in AHA [1] a page constructor consists of simple commands for the conditional inclusion of fragments.
- It performs adaptation by executing the page selectors and constructors. This means selecting a page, selecting fragments, organizing and presenting them in a specific way, etc. Adaptation also involves manipulating link anchors depending on the state of the link (like enabled, disabled and hidden).

It updates the user model (instance) each time the user visits a page. The engine will change some attribute values for each atomic concept of displayed fragments in a page, of the page as a whole and possibly of some other (composite) concepts as well (all depending on the adaptation rules).

The adaptation engine thus provides the implementation-dependent aspects, while DM, UM, and AM describe the information and adaptation at the conceptual, implementation independent level. Note that DM, UM and AM together thus do not describe the complete behavior of an AHS. The same set of *adaptation rules* may result in a different presentation depending on the *execution model* of the adaptation engine.

3. CROSS-REFERENCES IN AHS

Sometimes users do not want to follow the suggested navigation order, but rather skip across to related information. It would therefore be helpful to have access to meta-information for the concepts the user is reading about, especially through cross-references that make related information readily reachable. These cross-references are a collection of links to all concepts relevant to a (page)concept. Relevance is determined by examining the concept relationship graphs: there must be an (in)direct connection between the concepts. Section 3.1 explains how to generate cross-reference pages for each page in the DM. Section 3.2 explains how to update the user model for generating the CRNS. Section 3.3 explains how to personalize the cross-references.

3.1 Generate Cross-Reference Pages

An authoring tool can generate a set of cross-references for each page to connected or related concepts. It depends on the author and system what pages or abstract concepts are included in this set. Also depending on the system, cross-references can be show

as a part of the original page or in a separate page. For brevity and ease of reference, we choose to organize a cross-reference set as an independent page called a *cross-reference page* (CRP). A (page)concept may be connected with several abstract concepts through one relationship graph or with several abstract concepts through different relationship graphs. A CRP should contain a list of links that can be grouped according to the type of relationship so that the user can easily find related concepts and go directly to the related concepts. The CRP should be adaptable to a user's knowledge state for every referenced concept. Before generating CRNS, we first need to add some requirements on the general definition of relationships in the DM of AHAM. We then define a function to generate CRPs from the relationship graphs defined in the DM.

AHAM provides a platform to define all kinds of relationships. For a clear understanding we distinguish between the hypertext link relationship, and other more abstract relationships such as *prerequisite* or *consists_of*. To allow a user to go directly from one concept to a related concept, we have to create a hypertext link. If an abstract concept has no direct representation (e.g. C_1 in Figure 2) a choice has to be made as to which hypertext links need to be created. For the sake of argument we will assume that all related pages in the concept hierarchy are included, but this is really an implementation decision and could possibly be a configurable option in the authoring tool.

Once all required hypertext links are created for a (page)concept, the corresponding CRP can be generated. We can define a function to generate CRP for each page by using relationship graphs defined in the DM. There are different ways to generate cross-references according to relationship type and connection distance (the length of the shortest path in the relationship graph). For example, a CRP may contain links to relevant concepts for just one specific relationship type R , or for all types of relationships. Connection distance can be used to exclude certain links, and may be used as tool to determine relevance.

Let us assume that G is a relationship graph with type R , P is a page, and $P.cross-reference$ is a set consisting of relevant concept links to P . Relationship $R(C, P)$ represents that C is directly related to P through relationship R , while $R^+(C, P)$ represents that C is directly or indirectly related to P through relationship R . Function *Generate* generates the CRPs from the $P.cross-reference$ set. Function *Generate* is part of the authoring tool, so details are not relevant here. We illustrate four possible ways to the function to generate CRPs:

Algorithm 1 (only directly related concepts of relationship type R):

1. Select a relationship graph G with type R ;
2. **for** all P in G **do**
 $P.cross-reference := \emptyset$;
 for all C in G **do**

- if** $R(C, P)$ **then** $P.\text{cross-reference} := P.\text{cross-reference} \cup \{C\}$;
3. *Generate the CRP based on $P.\text{cross-reference}$.*

Algorithm 2 (directly related concepts of all relationship types):

1. **for** all G in Relationship_Graphs **do**
 for all P in G **do**
 $P.\text{cross-reference} := \emptyset$;
2. **for** all G in Relationship_Graphs **do**
 for all P in G **do**
 for all C in G **do**
 if $R(C, P)$ **then** $P.\text{cross-reference} := P.\text{cross-reference} \cup \{C\}$;
3. *Generate the CRP based on $P.\text{cross-reference}$.*

Algorithm 3 (all related concepts of relationship type R):

1. Select a relationship graph G with type R ;
2. **for** all P in G **do**
 $P.\text{cross-reference} := \emptyset$;
 for all C in G **do**
 if $R^+(C, P)$ **then** $P.\text{cross-reference} := P.\text{cross-reference} \cup \{C\}$;
3. *Generate the CRP based on $P.\text{cross-reference}$.*

Algorithm 4 (all related concepts of all relationship types):

1. **for** all G in Relationship_Graphs **do**
 for all P in G **do**
 $P.\text{cross-reference} := \emptyset$;
2. **for** all G in Relationship_Graphs **do**
 for all P in G **do**
 for all C in G **do**
 if $R^+(C, P)$ **then** $P.\text{cross-reference} := P.\text{cross-reference} \cup \{C\}$;
3. *Generate the CRP based on $P.\text{cross-reference}$.*

3.2 Updating the User Model

We briefly illustrate how the AHAM CA-rules are used to perform user model updates, in which the concept relationships play a role. Adaptation is normally based on *relevance* of or *recommendations* for concepts or pages. This is in turn deduced from aspects like *interest* or *knowledge* (which must exceed some *threshold* to be considered sufficient to be taken into account). Below we give a possible rule that decides whether a concept should be recommended based on whether the user has enough knowledge about all prerequisite concepts.

C: **select** $C_2.\text{knowledge}$
 where $C_2.\text{knowledge} \geq C_2.\text{threshold}$
 A: **update** $C_1.\text{relevance} := \text{"recommended"}$
 where Prerequisite(C_2, C_1)
 and not exists (**select** C_3)

where Prerequisite(C_3, C_1)
and $C_3.knowledge < C_3.threshold$)

The rule is triggered when the knowledge of concept C_2 is changed and when that knowledge then matches or exceeds the required knowledge threshold for C_2 . The action of the rule sets the relevance value for C_1 to “recommended” if there are no unsatisfied prerequisites left for C_1 . (For efficiency reasons only concepts C_1 are considered for which C_2 is a prerequisite. Other concepts cannot be influenced by the knowledge change for C_2).

In order to describe the user model updates, changes to relevance of pages, and also the resulting adaptation one needs many rules. We don’t describe them here, not just because of the size restrictions for this paper, but also because every AHS has different behavior and thus also a different description using AHAM CA-rules.

3.3 Personalize the Cross-References

In this section we discuss how to present a CRP. The relevant links presentation in the reference page should be adapted to the user’s features. For example, we use color metaphors as used for adaptive annotation to show the knowledge state for all links in the cross-reference page.

The following rule describes that when a CRP is displayed, the system shows the relevant concept links of this page by using the color metaphor for adaptation annotation.

C: **select** CR.access

A: **update** F.pres := “green”

where Fragment(CR, F) **and** F.relevance = “recommended”

Here F can be a page or an abstract concept. *F.relevance* has been calculated in UU-phase. The cross-reference page consists of a list of related page links or abstract concept links depending on constructors of the cross-reference page. From the cross-reference page users can go directly to related concepts. In this way we add additional navigation paths that we call link-independent navigation support.

4. CONCLUSIONS AND FUTURE WORK

With AHAM we can define any relationship within AHS. We can then automatically generate cross-references based upon these relationships. Using the basic functionality of AHAM we are able to apply adaptation and personalization to these cross-references to provide users with a better understandable navigation environment. This method could be very useful for educational web applications; teachers should provide the relationships among concepts for the course. It is less practical in large information systems without a clear notion of an author. But the idea to provide relationship based cross-references to assist the user while browsing through hyperspace is applicable in web-based information systems in general. All that is needed are meaningful relationships among pages, e.g. provided by semantic web techniques.

5. REFERENCES

- [1] De Bra, P., Calvi, L., “AHA! An open Adaptive Hypermedia Architecture”. The New Review of Hypermedia and Multimedia, pp. 115-139, 1998.

- [2] De Bra, P., Houben, G.J., Wu, H., "AHAM: A Dexter-based Reference Model for Adaptive Hypermedia". Proceedings of ACM Hypertext'99, Darmstadt, pp. 147-156, 1999.
- [3] EI-Beltagy S. R., Hall, W., De Roure, D. and Carr, L. "Linking in Context". Proceedings of The 12th ACM Conference on Hypertext and Hypermedia, pp. 151-160, 2001.
- [4] Halasz, F., Schwartz, M., "The Dexter Reference Model". Proceedings of the NIST Hypertext Standardization Workshop, pp. 95-133, 1990.
- [5] Halasz, F., Schwartz, M., "The Dexter Hypertext Reference Model". Communications of the ACM, Vol. 37, nr. 2, pp. 30-39, 1994.
- [6] Marshall, C.C., Shipman, F.M., "Spatial Hypertext: Designing for Change". Communications of the ACM, Vol. 38, nr. 8, pp. 186-191, 1995.
- [7] Wu, H., De Kort, E., De Bra, P., "Design Issues for General Purpose Adaptive Hypermedia Systems". Proceedings of the 12th ACM Conference on Hypertext and Hypermedia, pp.141-150, 2001.

Towards the tailoring of a ubiquitous interactive model applied to the natural and cultural heritage of the Montsec area

Montserrat Sendín¹, Jesús Lorés¹, Jordi Solà¹

¹ Computer Science and Industrial Engineering Department
University of Lleida, 69, Jaume II St., Lleida, SPAIN
Tel: +34 973 70 2 737 Fax: +34 973 702 702
e-mail: {msendin, jesus}@eup.udl.es
jordi@griho.net
URL: <http://www.udl.es/dep/diei>

Abstract. Starting from the multimedia information about the natural and cultural heritage from the Montsec area - that is already available on the interactive CD-ROM “La Memoria del Montsec” -, we propose an interactive model that offers such information with a double functionality. Firstly the diffusion via web, through as many internet devices as soon as they appear. The result is a multi-device dynamic web application which we have called “The Montsec Web”. Secondly, in order to be used in situ, to be a location-aware system. Thus, another level of interactivity is added, obtaining a mobile and context-aware multimedia environment which we call “The Interactive Montsec”. Both functionalities should be integrated on a common underlying technological base.

Additionally, the system will have to anticipate to the visitors’ requirements, adapting the selected information to his changing profile and history.

Keywords. ubiquitous computing, location-aware systems, augmented world model, adaptive hypermedia, user modelling.

1 Introduction

The mountains range of the Montsec are situated in the western sector of the Catalan low Pyrenees (in the province of Lleida). It is an area of outstanding geological, historical, paleontological and scenic wealth, boasting a valuable natural and cultural heritage, which is why it is so highly regarded as a truly natural laboratory. The application of the model presented here in the Montsec area will help to conserve and make people more aware of the park’s natural resources and cultural heritage all over the world.

The nowadays proliferation of Internet devices means that two-layer separation techniques – code and data – to create web pages will not be able to solve the problems that internet applications programmers are soon going to have. This is one of the reasons why it is being necessary to search for other technologies that allow to add another layer more in the applications development. It is the presentation layer,

that is to say, the visual appearance that the final hypermedia document has to take. The result is a three-layer functioning structure: data, code and presentation. This is the structure we have adopted to our prototype.

With the technological base we propose here we aim to spread the wide information about the Montsec zone to two well-differentiated environments. The first is the own Montsec zone, focusing on the generic visitor's requirements arising throughout his itinerary. We are referring to "Interactive Montsec", presented before. The other environment is the internet one, satisfying in this case to a generic net user who can use whatever existing internet device at present – referred before as "The Montsec Web" -. This last has to be possible without having to redesign the application. In both cases, the generic term used here makes reference to a user corresponding to a certain typology, which is gathered in the corresponding user modelling.

Our system solves the automatic generation of hypermedia documents in which the visual aspects related to documents presentation are based on previously designed templates – presentation models -. This avoids having to restrict documents presentation to an only particular knowledge representation. These templates are designed in order to adapt a same information to different devices. But this objective does not wholly cover all the expectations defined for the system.

The definition and processing and of the visual aspect separately will also solve the presentation of dynamic contents according to several changing contextual aspects. In the case of The Montsec Web these aspects are the user's profile and the kind of device used by him. In the case of The Interactive Montsec, the history of the visits and activities made by the user and, of course, the geographical location get also into consideration.

At the moment a prototype of The Montsec Web is already available as a multi-device dynamic application, in which the adaptation to different kinds of users and contextual situations should be improved. This means increasing the flexibility in several aspects. On the one hand, to set correctly a user modelling taking into consideration a bigger number of parameters. On the other hand, to offer the possibility of recording the evolution of each user along the time. Moreover, the user modelling constitutes a very important and determinant aspect to be considered when we are working the visual aspect of documents.

In relation to the second pointed functionality, the location-aware systems, under the name of location based systems, have recently received great attention as application field for mobile systems, and they have a very promising future.

In order to offer location-awareness – for instance, to show the position where the user is on a map or to make spatial queries, such as the nearest points of interest in relation to the location we are -, it is required a detailed model from the real world as a base. This model has to be developed and integrated in the architecture [1]. To make this non trivial task easier our proposal consists of delegating this responsibility to a spatial data managing module. This module will be in charge of building that model and of making possible its access. This is the objective of the Nexus project: providing a detailed model of the real world suitable to the functionalities and features of each location-aware application in particular. This model is called Augmented

World model (AW-model), and it combines objects from the real world with virtual objects. The Nexus platform is being developed at present at the University of Stuttgart [2].

These spatial questions will consist of identifying and locating the nearest points of interest, whether historical, geological, geographical, biological or even tourist nature – these are the thematic sections in which the information is structured -. These points of interest will be showed selecting or prioritising those that better adjust to a particular user's profile.

Under this expectation we can imagine the Montsec zone as a virtual museum where we can whenever obtain information from whatever is surrounding us in a implicit and tailored way. In an implicit way because the system solves automatically the location-awareness and, based on this it selects the relevant information anticipating so the visitor's requirements [3]. In a tailored way because the system takes into consideration contextual information related as much to the user's profile as to the visits and activities made by the user history, both of them brought up to date.

In relation to this second functionality we want to point out that a prototype based on GPS (Global Position System) has also been developed. According to the coordinates corresponding to the geographical point where we are, which are provided by a GPS device, the application shows the required information without having specifically to ask for it [4]. This prototype has been recorded on a videotape in November 2000 by the Catalan TV channel "Canal 33" [5].

The rest of the paper is made up of five sections more. On section 2 we present the technological structure that we have already developed up to date. In section 3 we expose the spatial and contextual data management that we propose to complete our Montsec web prototype. This way we intend our system to be anticipative and adaptive [6]. On section 4 we outline a sketch for the architecture which should integrate our starting system together the rest of the exposed functionalities. Next, on section 5, we present the tests we have carried out with the prototype. And finally, the conclusions.

2 Technological Structure of The Montsec Web Prototype

2.1 Layered Structure

As has been pointed previously, the application is structured in three layers which interact each other as we can see on the figure 1.

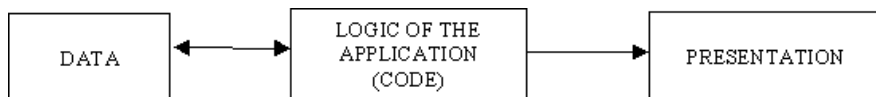


Fig. 1. Layered Structure

Let us see the aim of each layer.

Data layer. It is made up of a relational data base that at first gathers all the tuples of information selected for the prototype. By means of a JDBC driver we make possible to connect Java to the data base. In short, this layer provides required data to the code layer.

Code layer. This layer starts the processing of the request by sending a SQL query to the data base. When it receives asked data in jsp format, it translate them to XML format, sending the resulting XML file to the presentation layer which will have been validated and it will be well-formed.

Presentation layer. The aim of this layer is to translate data received from the code layer to different presentation languages, suitable to the current device (HTML, DHTML, WML, and FO among others). The code translation is carried out using some previously designed templates in XSL¹ code (eXtensible Style Language). It exists one template for each internet device we have planned. These templates combine the device specific code with XSL language labels. This labels are replaced by data extracted from the XML file at run time, in order to generate the corresponding presentation.

2.2 Underlying Technology

Now we want to present the operations to carry out on the code and presentation layers in more detail, showing how these are carried out and how they interlace each other. In addition, we want to detail the chosen technology for each part.

Data conversion to XML format (Code layer). The SQL query to the data base provides data in jsp format. The first step to do consists of converting this data to XML code, as we have already said.

Such a query is made by means of a jsp file. This file includes XML and JSP labels, as well as predefined labels of Cocoon² too. These labels basically show which template has to be used in the presentation layer to adapt the final web page to the kind of navigator that made the request.

The treatment of jsp code is carried out by the JSP servlet. It replaces the pertinent lines by data received from the data base. The result is the expected XML file. The JSP servlet referred resides in the Orion server we have used.

Validation and parser of the XML code (Code layer). The XML code obtained in the previous step has to be validated by means of a text file called DTD. The result is

¹ It is a language of styles to give format to XML data documents.

² It is a frame to publish web pages based 100% on Java and in the last W3C (World Wide Web Consortium) specifications [7] as DOM, SML and XSL.

a valid and well-formed XML file. Namely we have used the XERCES version 2.0 parser, from the Apache company.

Selection of the XSL page (Presentation layer). The selection of the XSL template to apply is worked out by the Cocoon (version 1.0.) module, which selects the one appropriate to the user's device. This module has been configured as a filter servlet. In that way, we achieve that the Cocoon servlet acts later than the JSP servlet. We want to underline that a filter is basically the main component which makes possible to combine the different technologies we have used. In particular, this was one of the most important requirements that the web server had to fulfil [4].

Data conversion to the final format (Presentation layer). The last step to realise after having the XML code validated and the XSL presentation template selected is to translate definitely data to a code directly treatable by the client navigator. Finally this resulting page will be returned to the client.

The XSLT³ (conversions language) converter used has been XALAN version 2.0. On figure 2 we can see how everything gets together.

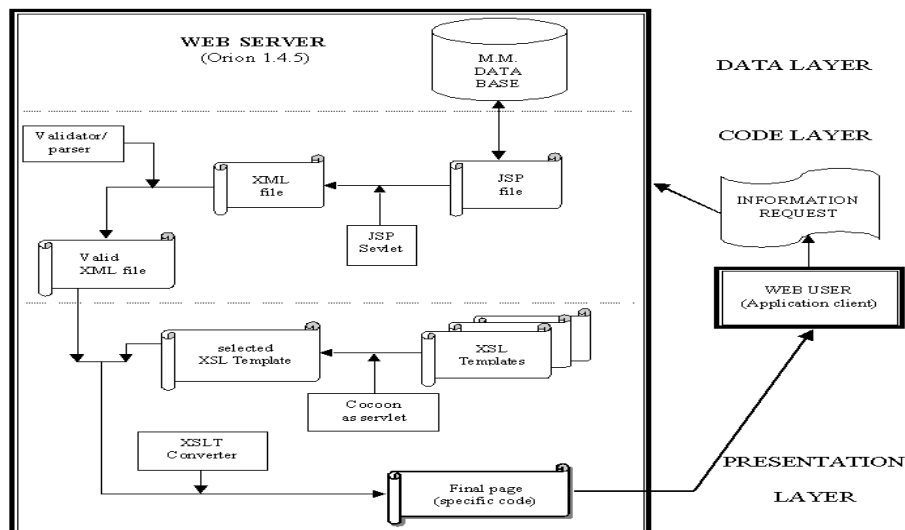


Fig. 2. Prototype technological structure.

³ It is a much more practical extension of XSL that allow us conversions of XML data to different formats.

3 Spatial Data Management and Tailoring

3.1 Augmented World Model Accessibility and Characteristics

The detailed model of the real world constitutes the central information model that provides an integrated and homogeneous view on the data available at each moment, in order to offer location-awareness [8]. Developing a detailed model of the real world turns out a costly task, specially if this one has to be updated, as it happens in our case.

Up to now, all the location-aware applications use their own particular world model. The goal of the Nexus project is to provide a detailed model of the real world suitable to location-aware applications, according to their concrete functionality – both for indoor and outdoor usage [9] -.

The result consists of representatives for real world objects and of virtual objects that provide, among other things, links to external information spaces like the web. This model is called the Augmented World Model (AW-model).

This model is kept up to date by integrating the update of sensor systems, for example to obtain the current position of mobile objects. In our case this updating is limited to know the position more or less approached where the visitor is at any moment. This one will be provided by the available GPS device.

Therefore, our application will formulate a query related to the visitor's zone. The Nexus platform will process this query accessing to the AW-model, and returning the requested information again to the application.

Location-aware applications may query the current state of the model by using the Augmented World Querying Language (AWQL) and receive as answer information about the model described by the Augmented World Modelling Language (AWML).

As much AWQL as AWML are languages based on XML. As we are working in XML format too, the compatibility is guaranteed. All that make us to think that the integration of the Nexus platform to our architecture will be feasible.

The Nexus platform basic interactions are graphically represented on figure 3. This interface offers to typical location-aware applications the required functionality hiding the details of the underlying data management. We should envision it as a middleware that brings together different providers and consumers of location-based information [2].

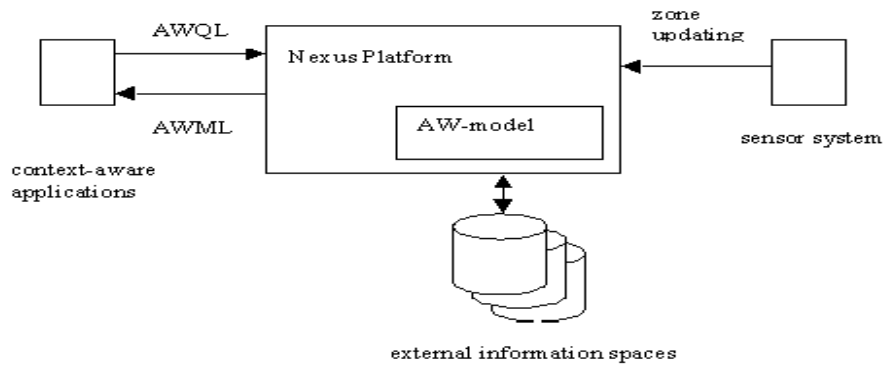


Fig. 3. Basic interactions of the Nexus platform

As the Nexus platform architecture – at the moment being in construction by the research group devoted to the Nexus project (Stuttgart University) –, does not constitute the central subject of our paper, it will not be showed in more detail.

3.2 Knowledge Representation

The essential components of the AW-model are the static, geographical objects data (spaces model: castles, towers, monasteries, rural areas, churches, etc), the location information of mobile objects (model of mobile objects: the zone visitors), and the virtual objects which provide additional information or services to Nexus users. The late have no counterpart within the real world, but they can be visualised by an appropriate application.

In order to define the AW-model, the Nexus platform uses an object-oriented approach, which allows the AW-model to be easily adapted to each concrete case. There are two categories of classes: members of the Standard Class Schema and the Extended Class Schema. The Standard Class Schema contains classes that we consider fundamental and that have a well known structure, so they can be used by every application. The Extended Class Schema will contain additional classes for special purposes, suitable to a concrete location-aware application.

It is aimed to identify the set of classes that better fits a certain geographical application area. This idea corresponds to the ontology concept.

The domain classes will be represented in AWML. So, the answer to an AW-model query will consist of a list of matching objects, in a particular zone. For example, a list of the nearest restaurants with their respective attributes – address and menu, for instance – and also a list of the castles feasible to visit in the zone, according to attributes as historical period and visits timetable. This would be suitable if our user has a certain preference for History. Depending on the interest degree – being a lover or an amateur – different information levels could get in action.

The restaurant object and the castle object are two examples of suitable entities of the Montsec area. So they have to be part of the Extended Class Schema. Another option would be objects that specialise some object from the Standard Class Schema with the suitable additional attributes, if it is possible.

The information level required by the user, or if, for example, he knows already a term related to the information such as the features of the romanic style of a church - these are aspects that we refer as contextual information -, will also have to be considered and modelled during the formulation of the query. We will so insert simple expressions on the query specification, for instance in Java. A possibility is to use an HTML extension based on JSP.

To define and formulate the query of this kind of adaptive elements in the AW-model we are going to introduce conditions on whatever class from the classes hierarchy. These conditions will have to match with the Java expressions included in the query. In that way it will be possible to construct dynamic structures which will take their definitive form at run time, according to the user model [10].

The contextual information will be used at run time to adapt to the user both the contents and the links selected – previous to the query time – and his presentation – at the generation of the hypermedia document time -.

3.3 Presentation Model

In The Montsec Web prototype developed we have only considered the device ubiquity. That is why designing a template suitable to each internet device was enough.

To also integrate the contextual information on the presentation – adaptation to the user's features and current situation – we have to diverse and extend the possibilities to generate a final presentation. We have to adapt the presentation to several circumstances, as much as possible [11].

In order to carry it out we are going to adapt two measures. On the one hand, we will enlarge the number of previously designed templates, taking into consideration not only the type of device, but also, for example, the preferred thematic section, according to the user's preferences. We consider this factor determinant in order to design the presentation.

On the second hand, we want also to update the templates contents. To generate personalised information we are going to introduce adaptive presentation elements, which are called presentation rules. They control aspects as the links generation, the spatial layout of lists, etc, and mainly take into consideration all kind of contextual information.

The combination of the templates language with the presentation rules will let us specify a wide set of non trivial presentations by means of a very simple syntax [10].

4 Integrating Architecture Sketch

We intend to integrate in our three-layer structured prototype the spatial data management proposed here to obtain a detailed model of the real world according to the purpose of our application, without having to carry out its management. This solves one of the functionalities that the starting prototype did not cover: being a location-aware application.

On the other hand, our final architecture has also to integrate all the aspects related to the contents adaptation according to the user's profile and history – as we call contextual information -. We will add two relational tables: the histories table and the profiles table, both of them indexed by the user's identification field.

As the contextual aspects will work out the query to realise, they will be solved a priori in order to outline the AWQL file used to make the query to the AW-model. Here is where the Nexus platform takes part. Additionally, the architecture has to be prepared to update as much the user's history as the user's profile according to the queries, visits or activities realised by the user, the concepts the user is assimilating and the preferences he is refining.

The identification module will solve all this management related to the contextual information. Later, on the presentation layer, it will operate again to select the template according to the thematic section preferred by the user. As we have pointed before, the offer of templates is now bigger, and in addition, their contents are adaptive.

Our aim is to obtain the adaptation degree that is proposed here, and to cover all the exposed expectatives.

In relation to the client of the application (the user), now he has a GPS device, and so he provides the information of the geographical coordinates, besides of his identification.

All that appears reflected on figure 4.

5 Tests Realised to The Montsec Web Prototype

We want to point out here that due to the developed application is only a prototype, the data base built is composed of a reduced number of tuples selected from each thematic section. Each tuple has been prepared to be consulted under three planned levels of information: basic, intermediate and expert. Besides, all that is available in two languages: Castilian and Catalan.

5.1 General Performance Tests

We have carried out part of the tests that usually are carried out for web applications. Our aim has been to validate on the one hand the robustness of all the code modules

that take part and also the server. On the other hand, we have also validated the runtime efficiency. Finally, we have checked the simplicity of maintenance. Let us see them separately.

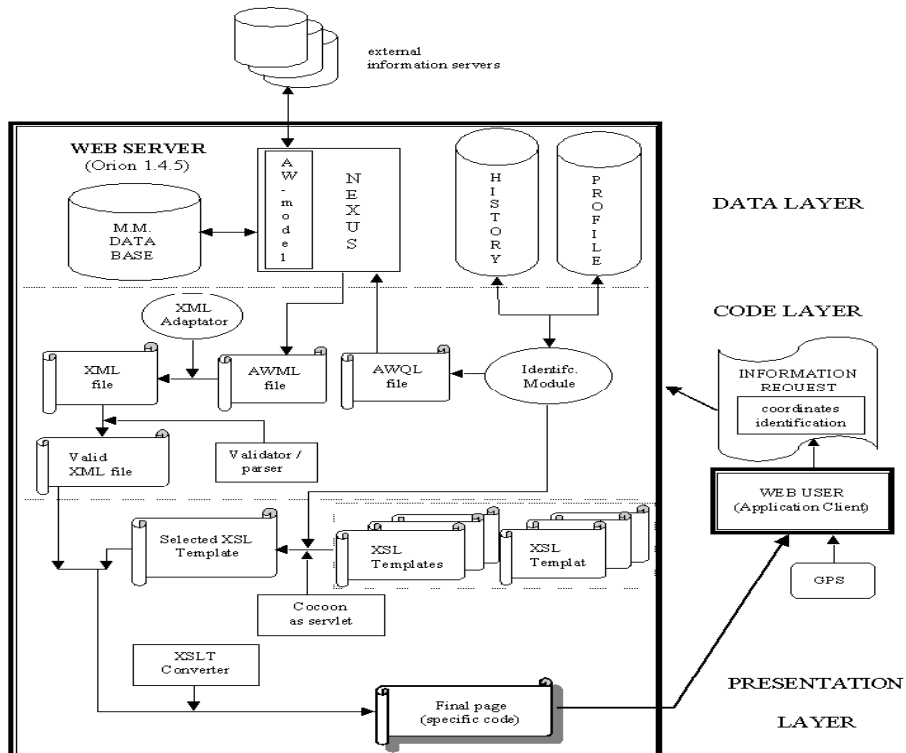


Fig. 4. Integrating architecture.

Server reliability. We have analysed the server efficiency in serving web pages and also the performance and mutual interaction among the servlets and the rest of the modules. We have made different tests with HTML, WML, JSP, CHTML and DHTML pages, and we have not detected any problem serving the different kind of pages.

In relation to the used parsers, validators and XML converters – XERCES and XALAN -, they have proved a high compatibility degree with the other modules of the web application, increasing with this the server reliability.

Download speed. We have evaluated this parameter in the different available devices. In general, the download speed has no been the expected one. Nevertheless, we have to take into account that the pages have to cross a sequence of stages (parse, validation and conversion from XML code) which implies a considerable and insuperable delay. With reference to that, we are expecting the version 2.0 of Cocoon in a short time. This will warrant an upper download speed.

Maintenance tests. The purpose of this kind of tests has been to prove the independence of the data layer from the code and presentation ones. In effect, thanks to the use of XML, no change has affected the application performance in any of the different versions of the presentations. So the tests have been fully satisfactory, as it was expected. Therefore we can conclude that our application behaves as a dynamic web.

5.2 Tests on Diverse Devices

It is obvious that the visual appearance and the format of presentation varies for each device, due to the graphic design and the navigation system have to be adapted to concrete possibilities of different devices performing the application. Here we detail the different devices we have used to prove the application.

WAP devices. This test has been realised on a WAP mobile emulator. As the possibilities of contents visualisation allowed by these kind of devices are very limited, we have offered the information only on the basic information level, as it is the most reduced.

IMODE mobiles. IMODE is another internet access protocol for mobile devices, developed by the Japanese company DoCOMo. The required language is CHTML. We have also used an IMODE mobiles emulator: the Internet Mobile Explorer, from the Microsoft company. This emulator also simulates aspects of download speed and screen visualisation capacity, which have also turned out useful. The tuples visualisation in this tests has been right. Nevertheless, it has been necessary to use scroll bars to see a whole tuple, due to the reduced size of the screen. The test realised with IMODE mobiles let us get a quite exact approach about the capacity of third generation mobiles.

Desktop Personal Computers. In this case, we have made two tests. The first one on one of the most used navigators: the Microsoft Internet Explorer version 5.5. The required language to prepare this kind of templates is DHTML. The second test has consisted of converting different information tuples on PDF binary format, to be visualised with Adobe Acrobat Reader. In this case the required language is FO.

6 Conclusions

In this paper we present the ubiquitous interactive model that we are developing. This model will solve diverse interactivity levels and it will also fulfil several challenger features; among them to be adaptive and anticipative. Applying this model to the Montsec environment will help to preserve and to promote all over the world the real state and natural and cultural resources of the park. On the one hand, it will offer adapted information to the zone visitors, obtaining this way a true “interactive space”, just as Weiser defined the ubiquitous computing in [12]. On

the other hand, it will also offer Montsec information to any net user from anywhere in the planet, without the device used offering any barrier.

The technological base obtained and presented in this paper turns out very flexible, adaptable and present, since it uses future technologies. The three-layer structure makes easier the reusability, modularity and consistency of the presentation, reducing in consequence the developing cost.

At present it exists a multi-device software prototype of The Montsec Web, which works with different templates suitable to each device, just as it was proposed in [13]. Nevertheless, all the aspects related to the user, which are going to be gathered on the user model, will be also incorporated to take part on the presentation dynamic generation.

The construction and use of an abstract presentation model taking into consideration all these aspects, will allow us to configure the adaptive presentation of contents independently of their developing [14].

Our future plans consist of constructing an architecture that incorporates, articulates and coordinates the device ubiquity as we have already solved, the location-awareness as we have presented here and finally the user modelling. Our final aim is to generate automatically hypermedia documents highly adaptive and tailored, using the technological base and the three-layer structure showed in this paper.

References

1. Cheverst, K., Davies, N., Mitchell, K., Friday, A.: Experiences of Developing and Deploying a Context-Aware Tourist Guide: The GUIDE Project. Proceedings of the Sixth International Conference on Mobile Computing and Networking (MobiCom 2000), ACM Press (2000) 20-31
2. Grossmann, M., Leonhardi, A., Mitschang, B., Rothermel, k.: A World Model for Location-Aware Systems. Published on behalf of CEPIS by Novática and Informatik/Informatique <http://www.upgrade-cepis.org>, Vol.II, Issue no. 5. Ubiquitous Computing (2001) 32-35
3. Brogni, A.: An Interactive System for the Presentation of a Virtual Egyptian Flute in a Real Museum. Virtual reality in archaeology. Archaeopress (2000).
4. Solà, J.: El Montsec Interactiu. Career Final Project, Politechnics Universitarian School, University of Lleida. (2001)
5. Sendín, M., Lorés, J., Balaguer, A., Aguiló, C.: Un Modelo Interactivo Ubícuo aplicado al Patrimonio Natural y Cultural del área del Montsec. Proceedings of the Second International Congress on Human Computer Interaction – Interaction'2001. Abascal, García and Gil (Eds.), Salamanca (Spain) (2001) 51
6. Pressman, R. S.: Ingeniería del Software. Un enfoque práctico. McGraw Hill, 4th edition (1997)

7. World Wide Web Consortium. W3C Note 08 May 2000, <http://www.w3.org> (2000)
8. Nicklas, D., Groámann, M., Schwarz, T., Volz, S., Mitschang, B.: A Model-Based, Open Architecture for Mobile, Spatially Aware Applications. Proceedings of the 7th International Symposium on Spatial and Temporal Databases (SSTD'2001) (2001)
9. Hohl, F., Kubach, U., Leonhardi, A., Rothermel, K., Schwehm, M.: Next Century Challenges: Nexus – An Open Global Infrastructure for Spatial-Aware Applications. Proceedings of the Fifth Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom'99), Seattle, Washington, ACM Press (1999) 249-255
10. Castells, P., Macías, J.A.: Un Sistema de Presentación Dinámica Hipermedia para Representaciones Personalizadas del Conocimiento. Proceedings of the Second International Congress on Human Computer Interaction – Interaction'2001. Abascal, García and Gil (Eds.), Salamanca (Spain) (2001) 45
11. Brusilovsky, P.: Methods and Techniques of Adaptive Hypermedia.. Adaptive Hypertext and Hypermedia, Kluwer Academic Publishers. Brusilovsky, Kobsa y Vassileva (Eds.) (1998) 1-43
12. Weiser, M.: Some Computer Science issues in Ubiquitous Computing. Communications of the ACM, Vol. 36, No. 7 (1993)
13. Sendín, M., Lorés, J., Aguiló, C., Balaguer, A.: A Ubiquitous Interactive Computing Model applied to the Natural and Cultural Heritage of the Montsec Area. Published on behalf of CEPIS by Novática and Informatik/Informatique <http://www.upgrade-cepis.org>, Vol.II, Issue no. 5. Ubiquitous Computing (2001) 23-26
14. Macías, J.A., Castells P.: A Generic Presentation Modeling System for Adaptive Web-Based Instructional Applications. ACM Conference on Human Factors in Computing Systems (CHI'2001). Extended Abstracts. Seattle, Washington (2001)

TORII

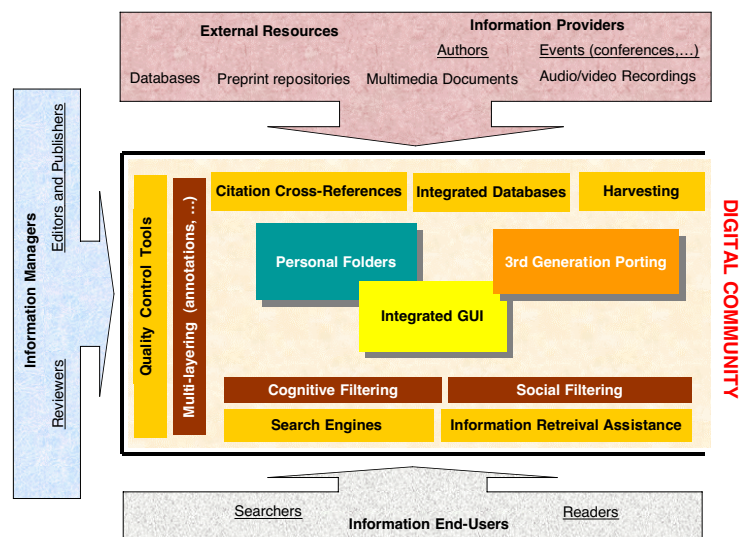
Access the Digital Research Community

Marco Fabbrichesi

INFN/SISSA
via Beirut 4
34014 Trieste
Italy

The communication of the results of scientific research and in many ways research itself have changed in recent years as digital means of information production, distribution and access have become widespread.

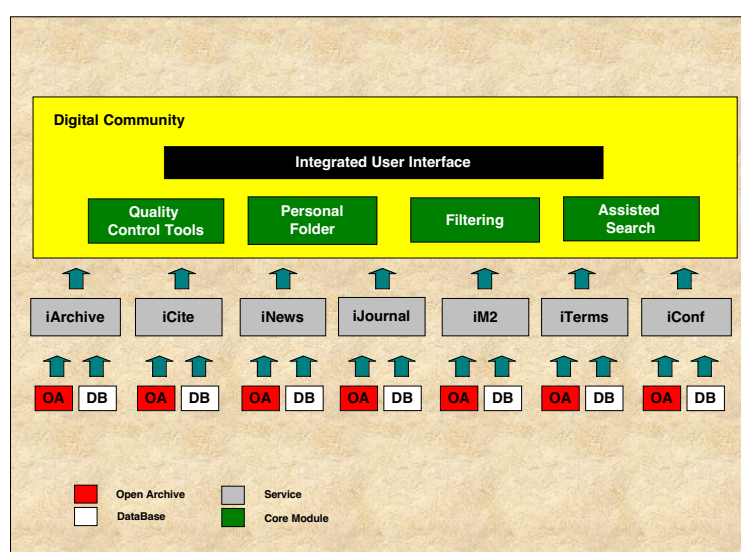
Think of your desk, or your office. The way you work has already changed. Paper preprints have been replaced by electronic archives,



mail and phone calls by e-mail, typewriters and hand drawing by text and graphics software programs, cabinet files by saved directories on hard disks. These new tools, together with multimedia presentations and conference websites, constitute the growing digital network of information that is taking over many aspects of the working place of research. It is a system in which the information flow is regulated, integrated and made available by the software and the network.

The digital network of research is currently organized in three layers:

- I Repositories of information: open archives and databases. This first level is the analogous of library and publishers stacks.
- II Services over and for information: e.g. review journal, cross-citation. They are the analogous of, for instance, library desks and paper journals.
- III Digital communities: synergic union of services and information. Ideally, they replace your desktop environment by giving access to the tools you use in your everyday work.



Only the first two have been fully digitalized: Torii is the first attempt to complete the structure.

Torii at torii.info gives you direct access to the digital research community. It works the way you work. All tools and documents you need are collected under an unified access point, organized according to your needs and ready for you everywhere you are and at any time you may need them. An intuitive user interface helps you to navigate. All the tools you need are at your finger tips. Choice of archives and subjects are easily costumized to fit your interests.

And the platform grows as the digital community grows. New features will be added as they become available in the future.

The personal folder is the hub of the system. You can store your documents here for future reference or to be printed or sent to others. The personal folder is easy to use by means of its drag-and-drop interface. It ideally replaces the cabinet filer where paper documents used to be stored. Stored documents can be ranked according to your profiles, impact factors or evaluation tools.

And there is more: you will find in your personal folder new documents suggested by the social filtering engine and you can attach to any documents comments for yourself or to be shared by the community.

The multi-layered document is a stack of documents that we want to manipulate. It could be an entry in a database, and as a new layer is added so is the entry column modified in the database, or it could be a collection of documents managed by a web server that keeps track of their relationships and modifications. The access to a multi-layered document is dynamical. According to who you are at a given moment—reader, author, referee, editor—you have access to different layers. Dynamical access requires an appropriate interface between the multi-layered documents and the users. It also requires intelligent agents to sift through the increasingly large amount of information to shape it into some hierarchy and thus making it usable.

Key features for the integration of dynamic access to the information into the portal are the XML language and the Open Archive Initiative protocol. The XML language is used to encapsulate in a common structure the exchanged information, originally stored in a variety of formats. This XML metadata structure will represent the semantic aspects related to the data.

The Open Archives Initiative Protocol for Metadata Harvesting defines a HTTP-based mechanism for harvesting XML files containing metadata from repositories.

This is the basic communications protocol between Torii and the underlying services, operating in a location-transparent way. The Open Archives Initiative, indeed, develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. The goal of the Open Archives Initiative is to supply and promote an application-independent interoperability framework that can be used by a variety of communities engaged in publishing content on the Web. Through the use of the Open Archive Initiative protocol, Torii will be easily extensible to any archive implementing the protocol.

In a user-friendly information society, the information overload is limited and the information delivery is personalized: the broad-casting of information is replaced by a more effective narrow-casting and mass-media are replaced by personal media tailored to each user's needs. These aims can be reached by more effective systems for information access. Torii provides a filtering component to skim too large a set of retrieved information and thus providing the user only with the information nearest to his interests.

The user defines his research interest profiles by filling in a form; from this a user profile is derived, based on a semantic network. The profile is automatically updated every time the user provides explicit relevance feedback on some new documents. Documents to be evaluated by the cognitive filtering module are processed through information extraction techniques aimed at capturing the meaning of the document content. These techniques exploit linguistic processing and statistical analysis. Every day the filtering module filters the submissions to the archives accordingly to the user profiles and the graphical user interface provides tools to rank displayed documents accordingly to the user's profiles.

Out of the 30-50 daily submissions, the user is able to see the 3-4 most relevant at the top of the list.

Social filtering circulates interesting documents among users who share interests. It automatically feeds in the personal folder those documents that are potentially relevant for the user. The relevance of the documents is evaluated for similarity with the selections done by other users with similar interests. The process is the digital analogous of sharing paper among colleagues. It fosters the growth of the digital community.

Quality control tools memorize and exploit human evaluation of documents. They provide users with the possibility to express their evaluation of a document by filling in a predefined form and writing free textual comments. The form results are used to statistically evaluate numerical scores about the scientific quality of the document, the comments are general, each user can choose whether his comment will be public or for himself. Users can read all public comments on a document. These tools embody a first instance of open peer review in which the community as a whole participate in the review process.

A search engine, Okapi, is accessed directly from Torii. It offers a sophisticated search environment where you can look and search among the more than 150,000 documents currently stored in the archives. Okapi offers advanced retrieval mechanisms based on the probabilistic model of retrieval and relevance feedback. It runs on both the document metadata and their full text. It is fast and accurate.

You are not left alone in searching. An assistant monitors your search and helps you with helpful hints and terminological and contextual suggestions. It alerts you for dead-end searches leading to hundreds of documents or no document at all. You are made aware of strategic aspects of searching that allow you to fully exploit all information resources and services. The assistant comes fully integrated into the Okapi search engine of Torii.

Every day iCite extracts all citations from all the documents submitted to the archives. These are used to rank documents in Torii so that you can order them according to their impact factors. It is a completely automatic system that creates a net of cross-references inside the archives. It is an instance of service, the second level of the three-layer structure, that can also be accessed independently at icite.sissa.it to search for citations patterns and ranking.

The role of libraries in the near future will be that of maintaining archives and databases and providing access for its users to the digital communities by subscribing to them. Aside from collections of old paper resources, new documents accessed through archive and services providers will be available on screen and, on request, printed locally by the new generation of digital printing/binding machines. the budget nowadays spent on journal subscriptions will progressively be transferred to community subscriptions, archive maintenance and hardware for users.

Torii is ready to move on into the future of digital networking. As the next generation of wireless systems comes into production, Torii will be accessible from your mobile phone. You can connect already and use it via WAP at

`torii.sissa.it/wml/ia.wml` but the full potentiality of the system must wait for the 3G broad bandwidth to come into being. At that point, you will be able to browse documents use your personal folder and any other of the features of Torii as you travel.

Torii is the web platform of the TIPS consortium (`tips.sissa.it`), a project sponsored by the Information Society Technology program of the European Commission. The ideas and concepts here summarized come from the work of many people. In particular, I would like to acknowledge the contributions of the people working at SISSA: Fabio Asnicar, Sara Bertocco, Lorian Bonora, Marina Candusso, Marco Mizzaro, Fabrizio Nesti, Feliz Sima and Cristian Zoicas

Okapi in TIPS: The Changing Context of Information Retrieval

Murat Karamuftuoglu, Fabio Venuti

Centre for Interactive Systems Research
Department of Information Science
City University
{hmk, fabio}@soi.city.ac.uk

Abstract. In this paper the changing context of information retrieval is discussed by examining the role of the Okapi text retrieval system in the “Tools for Innovative Publishing in Science” (TIPS) project. TIPS project poses a number of new challenges for end-user probabilistic retrieval, including field-based searching, Web access, and integration with other software. These and some other problems involved in designing a sophisticated Web-based best match retrieval system are highlighted. The architecture of the implemented Okapi Web system, its integration with the other systems that comprise the TIPS portal and design and evaluation of the user interface are discussed.

1 Introduction

In this paper we discuss the changing “context” of information retrieval by examining the role of Okapi in the EC funded “Tools for Innovative Publishing in Science¹” (TIPS) project [1].

Okapi is the name given to a family of experimental retrieval systems that have been developed over the last two decades². It is based on the Robertson-Sparck Jones probabilistic model of searching [4]. Okapi started life as an online library catalogue system and since has been used as the basis for real services to groups of users in various contexts. The TIPS project uses Okapi as the main retrieval tool to index and retrieve full text papers in High Energy Physics (HEP) area and to serve the information needs of that scientific community. The overall aim of the TIPS project is to provide a unified, desktop-like access to various services and tools to be used by the HEP community [1]. While Okapi has been the subject of a considerable amount of research, the TIPS framework presents significant new challenges.

The next section describes the overall structure of a typical Okapi system and reviews briefly the past Okapi research in relation to the TIPS project. The following section discusses in some detail issues that surround the implementation of the Web-based Okapi retrieval service in TIPS. In this section the architecture of the Okapi Web system, its integration with the other systems that comprise the TIPS portal and

¹ Information Society Technologies Program: IST-1999-10419

² For an overview of past Okapi research see [2,3]

design and evaluation of the user interface are discussed. The final section summarises the main points examined in the paper.

2 Okapi and the Changing Context of Retrieval

A typical Okapi system involves (Fig. 1):

- A search engine, the BSS (Basic Search System), which provides efficient low level functionality for weighting and ranking searches, and also for full Boolean and pseudo-Boolean (proximity) operations. Term weighting and document ranking is based on the Robertson-Sparck Jones probabilistic model of searching [4, 5].
- Indexing routines for indexing and inputting text documents. These accept raw text files and create a database in a form suitable for searching.
- Various interface systems that provide different representations of the underlying BSS functionality.
- Query Model layer, which manages the interaction between the user interfaces and the BSS and supporting additional functions such as incremental query expansion and passage retrieval.

BSS and indexing routines run on Solaris & Linux for Sun and Intel architectures.

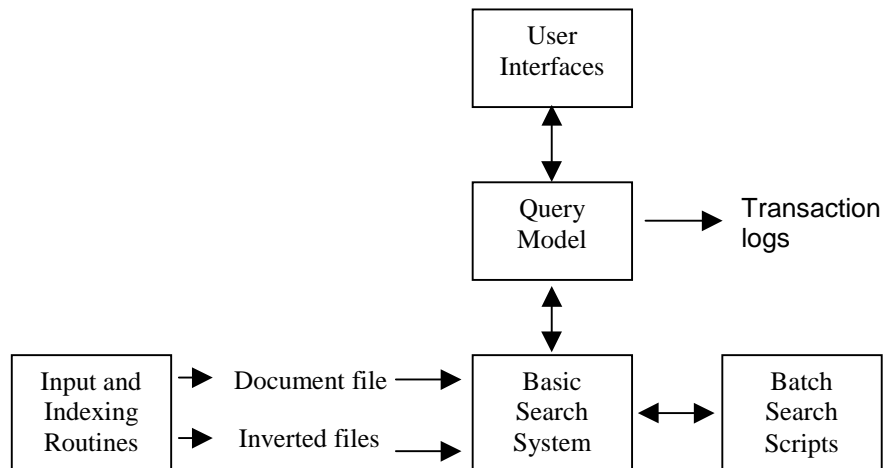


Fig. 1. Overview of Okapi architecture.

A typical Okapi search session involves the following:

User:

- Enters search terms.

System:

- Pre-processes, parses and stems the search terms to remove capitals, hyphens, punctuation and similar linguistic devices to convert the user input to the standard form used in indexing the documents in the databases. After input terms are pre-processed they are parsed to remove stop words and identify common phrases. The remaining terms are stemmed and weighted and document score are calculated. A ranked list of matching documents is presented to the user.

User:

- Provides relevance judgements feedback on the documents (relevance feedback).

System:

- Selects terms from the relevant documents, expands the query and performs a new search based on the expanded query. Term selection is based on the model described in [5].

There are four main sources of data used in query term weighting in Okapi [5]:

- Collection frequency - Terms which occur in only a few documents are likely to be more useful than ones occurring in many
- Term frequency - The more frequently a term appears in a document, the more important it is likely to be for that document
- Document length - A term that occurs the same number of times in a short document as in a long one is likely to be more important to the short document than it is to the long one
- Relevance information - The more frequently a term appears in relevant documents, the more important it is likely to be for that query

All these elements are combined to give a weight for each term-document combination, and then all these term weights are combined to give a total score for how well the document matches the query.

Okapi was set up specifically to provide an environment in which ideas could be tested in live settings that involve real users and information needs. Although some of the Okapi projects involve laboratory type experiments, as exemplified by the involvement in the TREC program, the main focus of Okapi-based projects has been to explore the inherent interactive nature of the retrieval process.

User-oriented Okapi research [2,3] in the past decade concerned with information seeking behaviour of users given different representations of the underlying probabilistic retrieval model at the user interface level. The past Okapi experiments involved a database of short bibliographic records (INSPEC) and/or library catalogues. The systems allowed users to enter "free-text" queries, which were matched against an index derived from the titles and abstracts of the records. Similarly the query expansion was based on the same index. These systems, hence, are basically free-text search tools that did not implement searching on specific fields

(e.g. author) or combination of fields (e.g. title and author). Users in these experiments were students and other members of the City University. Although they were all studying/researching in the general area of IS&T, they constituted rather a heterogeneous group with different levels of knowledge, skills and interest in the subject. Laboratory experiments in the context of TREC (both purely automatic runs and interactive experiments with human subjects as users) used full-text databases of various kinds, including newspaper and newswire stories. Again these were treated in a purely “free-text” fashion, with no field searching.

Involvement in the TIPS project poses a number of new challenges for end-user probabilistic retrieval: The document collections used in the project contain the full-texts of the articles as well as bibliographic information (or metadata) associated with the records, such as title, author names, date of publication, assigned keywords (controlled and/or free-text), and abstract. The source documents to be indexed are in TeX format that contain mathematical formulae and special symbols as well as plain (ascii) text. We have a research community to serve, who are experts in a particular area of Physics, namely HEP. The system will need to have client-server architecture and will be accessible on the World Wide Web using standard Web browsers. There will be *many* concurrent users of the system at any time. Finally, Okapi needs to work with and be integrated into other software developed by our partners in the project. These points are discussed further below.

Table 1. Changing context of the retrieval process.

Past	Present
Bibliographic (plain text)	Full-text (TeX format)
Free-text searching	Field searching
User community mainly novice	User community mainly experts
LAN-based	Web-based (HTTP)
Client and sever tightly coupled	Client and server separated
Few concurrent accesses	Many concurrent accesses
Stand-alone	Integrated with other software

3 Okapi in Tips

3.1 System Architecture

One of the main challenges in implementing a Web version of Okapi is that, HTTP is a stateless protocol. Between the requests sent to the Web server the client-server connection is broken, so information about the search history is not automatically maintained. However, one of the main defining characteristics of a probabilistic

retrieval system is that relies on the history of interactions between the user and the system within a given search session to optimise its performance. In particular it needs to store the current query state and documents selected as relevant by the user.

So the most important technical problem to be addressed in the Web environment is how to reconcile the Web's essentially "stateless" mode of operation with the tight browser-server integration and session continuity considered essential for an effective Okapi implementation. There are several different technologies available to maintain session continuity. Rapid evolution of the Web means that several new tools are added to the available set all the time, and selection of any one of the methods at a given moment may turn out to be non-optimal.

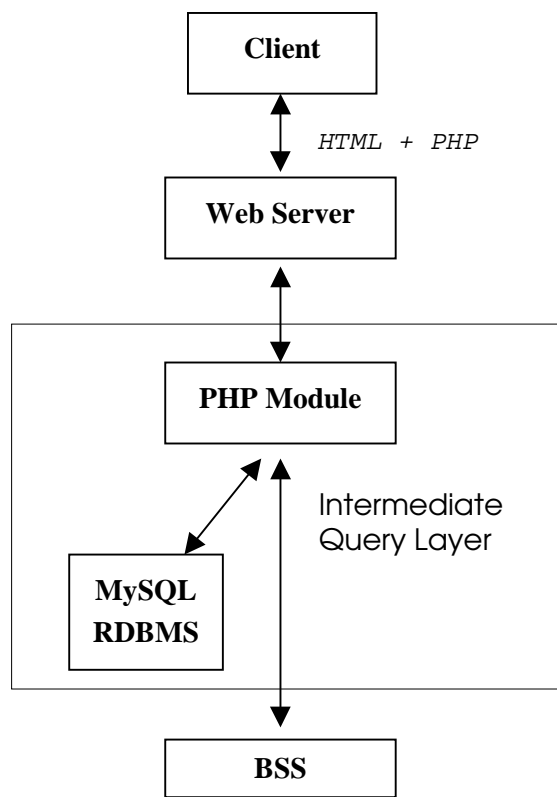


Fig. 2. Prototype Okapi Web System.

Fig. 2 illustrates the architecture of the Web-based system implemented. In this system, the client communicates with the Web server by means of PHP code embedded in a standard HTML document. Communication between the Web server and BSS is achieved via an Intermediate Query Layer (IQL), and session continuity is

maintained using a relational database holding data about the current retrieval state. IQL interacts with the BSS to perform various low-level BSS operations used in searching and retrieving documents and provide additional functions for query manipulation and expansion, plus transaction logging. The Web browser acts as a *thin* client, most of the processing are done at the server-end. The system can service simultaneous requests from multiple clients, maintaining the current query state for each of them.

The combination of PHP and relational database management system (RDBMS) to hold session information has proved to be a practical solution in the time-scale of the current project. PHP scripts are relatively simple to write, they are portable, all the standard browser functionality provided via the normal HTTP interface is retained and the resulting client is lightweight. PHP supports XML, which is used by the portal into which Okapi is integrated (see 3.2 below).

3.2 Integration with the Portal

The above described system is integrated into the TIPS portal, which is based on the Apache Software Foundation's JetSpeed groupware [6]. All portal content is stored in XML format. XML documents are dynamically rendered and converted to client-dependent formats such as HTML, PDF and WML (for mobile devices), at the server end (see a separate paper on the portal in the proceedings of the workshop).

The Retrieval Assistant and Advanced Search. One of the sub-systems of the portal is the Information Retrieval Assistant which is a rule-based expert system developed by our project partners. The retrieval assistant employs a relatively complex rule base to "reason" about the search process; it monitors user actions and gives contextual suggestions to improve the search outcome. The suggestions aim to help the user exploit the various resources available to them, and to resolve specific problematic situations (see a separate paper on the retrieval assistant in the proceedings of the workshop).

A distributed architecture is developed to establish the communication between Okapi and the retrieval assistant. The communication is one-way and consists of messages sent by Okapi to the retrieval assistant, which is called by means of three java servlets: one for the creation of a new session, one for the analysis of user's actions within the session and one for the deletion of the current session. Whenever a user interacts with the search interface, the PHP module of the Okapi system processes the requests and saves temporary data into the relational database. At the same time Okapi calls the proper servlet and passes the message regarding the user's actions (e.g.: search queries, document judgements, etc.) and the results obtained (e.g.: number of postings for each search term, list of documents retrieved, etc.) as a list of parameters to the servlet itself. Hence the retrieval assistant can monitor the user's activity and send suggestions directly to the portal to be displayed to the user. The two systems operate as independent 'black-boxes', neither need to know about the internal structure of the other. All information needed by both the systems to conduct an interactive retrieval session is stored in the relational database (cf. Fig. 2).

The same relational database serves for system evaluation and user behaviour research.

Communicating with the Portal. Contrary to communication with the retrieval assistant, communication with the portal is two-ways: Okapi receives data from the portal (e.g.: a query), processes it and sends back the results to the portal (e.g.: the list of documents retrieved). For this task the XML-RPC protocol [7] has been chosen. Okapi is seen by the portal as a set of functions remotely called through the web. The remote call from the portal is received in the form of an HTTP post by the web server where Okapi is installed. The web server then routes the request to the XML-RPC server that is part of the PHP module of the Okapi system. The XML-RPC server parses the request written in the XML-RPC syntax and calls the proper PHP function (for instance, the function `okapi.SearchDb` that performs a search given a query and a username). The function processes the data passed as parameters and sends back the results to the XML-RPC server, which prepares and sends back the appropriate response to the portal.

3.3 User Study

A small-scale user study was carried out to evaluate an earlier version of the system. The study involved a small number of handpicked expert users of HEP document collections. Its purpose was to improve the usability of the system based on the feedback elicited from this focus group. The subjects performed their searches on an indexed sample of about 4000 full-text arXiv documents. Five expert users in High Energy Physics were asked to participate in the experiment. All five users perform daily online searches that fall into two main categories:

- Filtering or monitoring: the user searches for new relevant documents in an established area of interest. The area is precisely defined (e.g.: “*localised particles on branes in extra-dimensional models*”). This kind of search is performed daily and often more than once per day. Users know exactly the keywords commonly used in the area and the most interesting authors, and demonstrate a high degree of expertise and confidence when making relevance judgements.
- Starting a new search to explore a new area of interest, often following leads suggested by other colleagues. The documents in most demand are review and “top cited” articles.

The five users were first given a short introduction to the system by the experimenter. They were then asked to perform a search on a topic concerning their research interests and give their comments. Their actions were logged. The experimenter took note of the comments made by the subjects and intervened whenever necessary to clarify a user action or comment.

The comments gathered from the focus group are used to improve the design (functionality as well as the user interface) of the system. Specific results drawn from the interviews and observations made by the experimenter are summarised below (more detailed discussion could be found in [8]). Users expressed their wish to:

- Search in document ID, full-text, abstract, title, author name, date and keywords fields.
- Search by phrases as well as single keywords.
- Restrict the search to some specific arXiv sub-area (category) and possibly combine them (e.g.: search in astrophysics and hep-experiment).
- See detailed information about the documents in the hit list, including, the names of the authors.
- See the number of citations the document received by a document. (The need to see in the hit list whether or not the document has been published in a referred journal was also expressed).
- See in the document record the reference list and a list of other papers that cites the document.
- Recover a past search session, with queries, results and relevant documents.
- Find similar documents to those chosen as relevant.

3.4 The User Interface

Based on the feedback elicited from the experimental study a new interface is designed which is comprised of the following main elements:

Query Frame. Fig. 3 illustrates the user interface of the Okapi search system. The system allows users to enter single words or phrases delimited by double quotes to describe their information need. The query language includes the “+” and “-” operators to include or exclude documents that contain the marked term(s) from the search results. The system assigns a high positive weight when a term is marked by “+” sign and inversely a negative weight when marked by the “-” sign.

There are currently two collections available in Okapi for retrieval purposes. First one, arXiv³, is a collection of pre-prints in the HEP and related areas. The other, Journal of High Energy Physics (JHEP)⁴, is an electronic journal in the same subject domain. The fields searchable in these two collections are: document ID, full-text, title, author, abstract, date (of submission in arXiv, publication in JHEP) and keywords. Keywords are assigned by authors, so not all documents contain them.

For field-based searching the weight of a term in a given field is calculated on the basis of its frequency in that field. Weights of all matching terms for a given

³ arXiv is divided into four areas: Physics, Mathematics, Nonlinear Sciences, Computer Science. The physics area is divided into twelve sub-areas, of which six are of interest to the HEP community. As of April 2002, arXiv contained about 150,000 full-text papers in these six HEP related sub-areas. Number of new papers submitted to arXiv was about 2500 per month. arXiv is available from <http://xxx.lanl.gov/> and a number of mirror sites. The Web site (excluding the mirrors) receives over 100,000 daily (in weekdays) connections. Detailed usage statistics about the arXiv collection can be found in [9] and at http://arXiv.org/todays_stats

⁴ JHEP is a small database (a few hundred) of full text papers. The JHEP site and its mirrors receive roughly 650 paper download requests per week. More detail about JHEP can be found in [9].

document are combined to calculate the document score as described in section 2. When more than one field is used in the search, the sets of documents that match the search terms in a given field are retrieved for each field separately and the final hit list is created by merging the sets using the Boolean AND operator. Users can search simultaneously in arXiv and JHEP or separately in one of the collections. Users can also limit their searches to one of the sub-areas (categories) of arXiv.

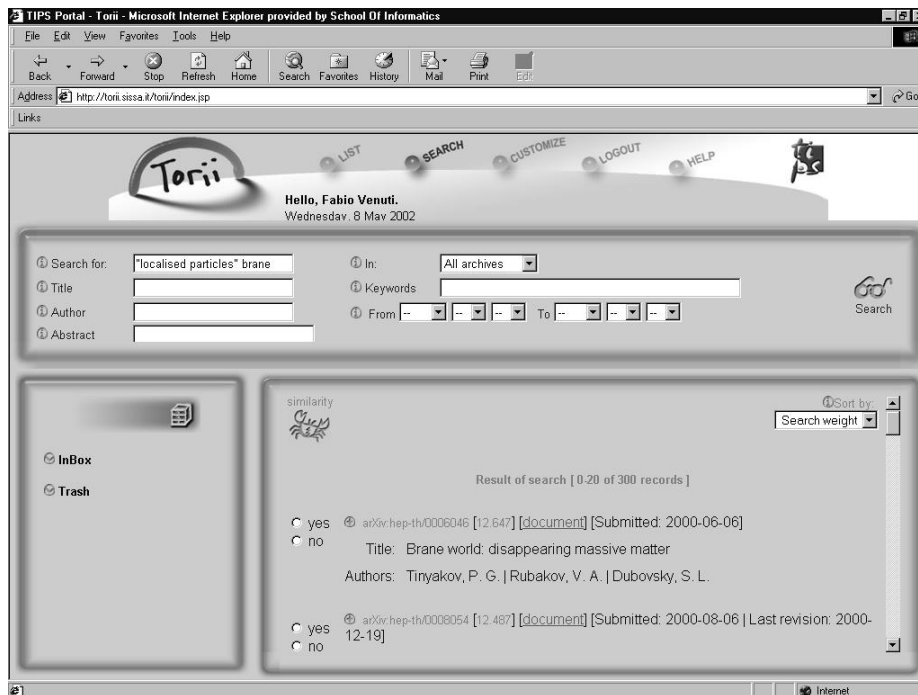


Fig. 3. The Search Window.

Hit List. Document ID, title, author name, and date fields are displayed in the hit list (Fig. 3). By default the documents in the hit list are sorted in descending order of their Okapi scores. The user could also choose to order the list by document ID, impact factor, or quality score. The latter is assigned by other users of the system (see a separate paper on the quality control tools in the proceedings of the workshop). Clicking a title in the hit list brings up the document record (see below). In the hit list next to each document entry there is a binary relevance feedback option in the form of yes/no radio buttons. Documents that are marked with yes are used to expand the user's current query in the manner described in section 2. The user could then repeat the search with the new expanded query using the similarity search option (see further below).

Document Records. The information shown in the document record window includes the bibliographic details and abstract of the document. It is possible to view the reference list given in the document. One can also view other documents that cite the document in this window. It is also possible to evaluate the quality of the document filling a form. The user can save the document to a personal folder and/or use it to modify her/his profile(s). The user profile(s) is used by the document filtering system (see a separate paper on the portal in the proceedings of the workshop). There is also a link in this window to the source document in various formats.

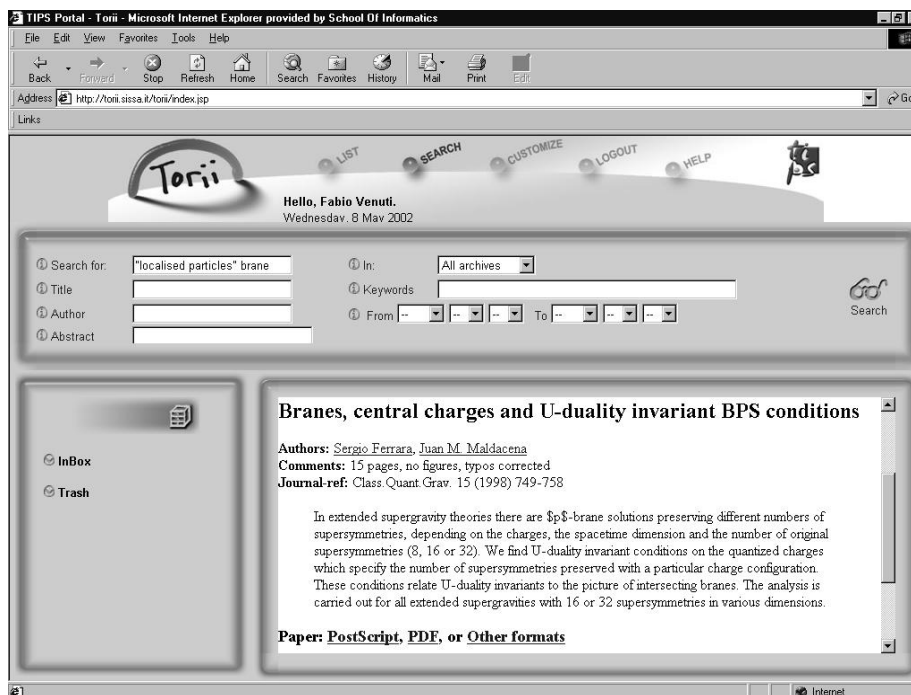


Fig. 4. A Document Record.

Similarity Search. After two positive relevance judgements⁵ the user can activate the “Similarity Search” option⁶. This triggers another search using the expanded query,

⁵ The more documents judged relevant, the better the term selection performance of the probabilistic formula becomes. A single positive relevance judgement is not enough for effective term selection.

⁶ The “Similarity search” label is used in the user interface instead of the more accurate, but perhaps less intuitive label of “Relevance Search”.

comprising terms extracted from the full text of relevant documents. Use of field-based criteria in the original query is disregarded at this stage, as are multi-word phrases and the “+” and “-” operators: the search is based solely on single word stems selected from the full text of relevant documents. The problems involved maintaining all user criteria during the query expansion process are discussed in detail elsewhere [8].

History and Judged Documents. A search “History” function has been implemented (but not yet integrated into the portal at the time of writing this paper) to backtrack to a previous state of the search process within a given search session. To keep it simple and avoid a too cumbersome interface we save only the original user query and the result of the last similarity (relevance) search – if the user made use of this option – for each of the queries within a single search session. Results of intermediate searches are not saved. The user queries are represented in the History view by the terms submitted by the searchers. A user can submit any number of queries by filling in the search form and pressing the submit button in a given search session. Each search request made in this way is considered as an independent query. A search session starts when the user submits a query for the first time and terminates when s/he logs out from the portal (or timed out). It will also be possible to save the complete search session (or any part of it) before logging out from the portal.

4 Conclusion: The Future

Some of the problems involved in designing a sophisticated Web-based retrieval system and solutions implemented are discussed. Our experience from our involvement in the TIPS project showed that creating a user-friendly interface while maintaining the power of the underlying system remains to be one of the main challenges of information retrieval research. The HEP community with geographically distributed user population who depend on document retrieval to support their day to day activities proved to be a highly fertile ground to study the complexity of the information retrieval problem.

As well as providing challenges, such an environment provides new opportunities to tackle the information retrieval problem from new directions. We are barely starting with the TIPS project to appreciate the collaborative and social nature of the retrieval process. A substantial amount of theoretical and empirical research is needed in this direction to understand the social contexts of the retrieval process and how to harness the potential offered by information networks to enable people to discover, create and share information.

References

1. TIPS project home page. Available at <http://tips.sissa.it>
2. Special issue of *Journal of Documentation* 53 (1), 1997.
3. Okapi projects home page. Available at

- <http://web.soi.city.ac.uk/research/cisr/okapi/okapi.html>
4. Robertson, S.E. and Sparck Jones, K. Relevance weighting of search terms. *Journal of the American Society for Information Science* 27, 1976, 129-146.
 5. K. Sparck Jones, S., Walker and S.E. Robertson, A probabilistic model of information retrieval: development and status. University of Cambridge Computer Laboratory Technical Report no. 446, 1998. Available at <http://www.ftp.cl.cam.ac.uk/ftp/papers/reports/#TR446>
 6. JetSpeed project home page is available at <http://jakarta.apache.org/jetspeed/site/index.html>
 7. XML-RPC home page. Available at <http://www.xmlrpc.com/>
 8. Karamuftuoglu, M., Jones, S., Robertson, S., Venuti, F., Wang, X. Challenges posed by web-based retrieval of scientific papers: okapi participation in TIPS. *Journal of Information Science* 28, 2002, 3-17.
 9. TIPS Information Resources. Available at <http://tips.sissa.it/docs/UR-R1/UR-R1-IR.ps>

Personalization techniques in the TIPS Project: The Cognitive Filtering Module and the Information Retrieval Assistant

Stefano Mizzaro and Carlo Tasso

Artificial Intelligence Laboratory
Department of Mathematics and Computer Science
University of Udine
{mizzaro, tasso}@dimi.uniud.it

Abstract. Persistent and ephemeral personalization techniques can be exploited to implement more adaptive and effective information access systems in electronic publishing. Within the TIPS project, we effectively applied both these techniques in information filtering and retrieval systems used, via the specialized Torii portal, by physicists in their daily job.

1. Introduction

Internet has changed, and is changing, the standard communication mechanism adopted in science. Nowadays, a peer reviewed journal can be distributed by electronic means, and the peer reviewing can take place completely electronically, drastically reducing time and money for publishing (see, e.g., JHEP at jhep.sissa.it or Earth Interactions at EarthInteractions.org). Many publishers now allow their subscribers to electronically access the full text of the papers published on standard journals. Beyond modifying the standard scholarly journals and proceedings, the Web has also introduced a new way of disseminating scholarly knowledge: *e-prints*, i.e., open online repositories of scholarly papers (see, e.g., arXiv.org, mainly about physics, or cogprints.soton.ac.uk, about disciplines concerning cognition).

As a result, the scholar is nowadays overloaded by a large amount of highly structured hypermedia information, in the form of scholarly publications, online repositories, commentaries, and so on. In this scenario, it is important to allow the scholar: (i) to stay up-to-date, being notified when new information on some topics of interest is published, and (ii) to quickly and easily find, on demand, information on specific topics. Both goals can be approached by advanced personalization techniques. At the University of Udine, we have been investigating the issue of personalization in information access for several years [1, 7, 8, 9, 17, 19]. Within the 5th FP IST project TIPS (Tools for Innovative Publishing in Science, contract number IST-1999-10419), see tips.sissa.it, we applied adaptive and personalized information access techniques to the electronic publishing field, and more specifically in scholarly publishing.

Personalization is needed and useful in information access, and especially in scholarly publishing. Users (i.e., researchers) are interested in it for two important reasons: (i) detecting newly published information relevant to their interests and

preferences, and (ii) accessing stored information satisfying specific information needs. However, this twofold situation requires a novel approach, in which two distinct and complementary personalization techniques (i.e., ephemeral and persistent personalization) can be applied together to meet user's requirements. Personalization plays indeed a fundamental role not only for the highly subjective nature of the information seeking process, but also because the job of a researcher is highly innovative, it does not conform to any standard behavior, and it is therefore quite different for each researcher.

Personalization techniques are very numerous and are ranging from simple user-controlled customization of Web content, to autonomous system-controlled adaptation [14, Reader's Guide, p.6]. We distinguish two types of personalization [20]: *persistent* (or *long term*), i.e., based on a user profile which lasts over time and is stored in a persistent information structure; and *ephemeral* (or *short term*), which is not based on a persistent user profile. The main differences are the temporal features of the process aimed at building and managing the user profile: in persistent personalization, the user profile is incrementally developed over time and at the end of each session it is stored in order to be used later on in subsequent sessions. These two personalization techniques (ephemeral and persistent) nicely match with the two classical kinds of information access [6, 10]: information retrieval (IR) [2] and information filtering (IF) [12].

On the one side, personalization in IF means capturing the long term information interests and preferences of the user, in order to tailor the selection process to the specific personal characteristics. On the other side, persistent personalization is not feasible in IR, since in that context information needs have a short term nature and are different, for the same user, in the different sessions. However, ephemeral personalization can be effectively exploited, with the goal of modeling the search session, rather than the information need, for immediately providing personalized support during the searching session. The resulting approach to the personalization in IR systems is innovative for two reasons: (i) a short term modeling is performed through ephemeral personalization, which restricts the scope of observation to the current session only, and (ii) we do not build a model of the information need (difficult, if not impossible, during just one session), but rather a session model.

In the remaining part of the paper, we briefly describe the two Torii modules in which we have applied this approach. A longer description is available in [18].

2. The Cognitive Filtering Module

In previous work, we have developed and evaluated several content-based filters [15] for persistent personalization. Among them [1, 17, 19], the most effective has been the information agent *ifT* (*information filtering Tool*) [17], which is based on the user modeling shell *UMT* (*User Modeling Tool*) [9].

ifT exploits lightweight natural language processing and co-occurrence-based semantic networks for building long term user profiles and for evaluating the relevance of text documents with respect to a profile. The main mechanism for building user profiles exploits explicit relevance feedback provided by the user on both positive and negative examples. The learning capabilities of this mechanism have been evaluated by means of several laboratory experiments [1].

Given the performance reached by ifT, we decided to adopt it as the filtering engine of the Torii portal. More specifically, the problem approached with persistent personalization has been the high (and currently increasing) rate of incoming documents: about 100-200 new e-prints are submitted every day and included in arXiv, which is accessible through Torii. Normal users (researchers in high energy physics) were used to start the working day by browsing the long list of new e-prints. By adding a personalized filtering engine to Torii, each user can now define one or more profiles related to his interests, and all the new incoming information is automatically filtered. In this way, Torii displays (in the first positions) only the documents which best match user's interests. Information overload is then reduced, as well as the cognitive load of analyzing many documents every day.

Torii has undergone a validation phase through field testing in July 2001. Twenty users were using the system for 29 days. All their sessions have been monitored and tracking logs of all actions acquired. Final interviews were also delivered. Cognitive filtering was working well and judged well by the users, who proposed to extend the system with the possibility to rank any set of documents (possibly coming as the result of a search in one of the available collections) by means of ifT.

3. The Information Retrieval Assistant

The interactive nature of IR is advocated since years [13] and is now widely accepted: between the user and the IR system a dialogue takes place [5], during which the user should receive adequate support [3]. The help should be provided proactively by the system and suggestions should be given "on the background", with the user retaining the control of the interaction [4]. A basic kind of support is *terminological* help, which identifies and suggests to the user terms that improve the query [11, 16]. Another kind of support is *strategic* help, which provides to the user useful hints on how to improve the strategy adopted for organizing the searching process (see a survey of this issue in [8]).

We use ephemeral personalization techniques to provide both strategic and terminological support to IR users. We have been doing research on this issue for several years. We implemented the FIRE and SAM prototypes [7, 8] that, by means of thesauri and of a detailed conceptual model of the session, are capable of suggesting to the users of a boolean IR system alternative terms and strategies to better (re)formulate their information needs. After some laboratory experiments involving several participants, we had evidence that terminological and strategic help are useful and nicely complement each other [7]. Following this positive evaluation, we applied ephemeral personalization to the IR system deployed in the Torii real setting: we implemented the Information Retrieval Assistant (IRA), a system providing various kinds of suggestions to users that are searching the paper and e-print database available in the Torii portal.

The most important innovative features in IRA concern the new models on which ephemeral personalization, i.e., both terminological and strategic suggestions, is based. Terminological help is obtained by a new spreading activation algorithm capable of browsing an heterogeneous, dynamically generated, and integrated thesaurus, starting either from the last inserted search term, or from the set of all the

search terms used by the user so far. Terminological help is then presented to the user by means of a ranked list of suggested terms.

The basic reasoning process exploited for ephemeral personalization is described in the following. Each user *action* (i.e., any operation performed by the user) on the user interface is notified to IRA. IRA monitors these time-stamped actions and builds a model of the session history, that is made up by a sequence of interleaved actions and states. A *state* is a set of parameters describing the current state of the system, like number of terms in the query, number of retrieved, read, and judged (as relevant or not relevant) documents, etc. At each state, i.e., after each action, a new set of situations is inferred. A *situation* is a history pattern, or an abstract description of the session history. Situations can be very simple, like ‘insertion of a zero posting count term in the query’ (a term that is not contained in any document), or they may concern a longer time interval, like ‘two consecutive searches with no changes to the query’. Moreover, they can be more abstract and difficult to infer certainly, like ‘user not reading the content of the retrieved documents’. The derivation of new situations is triggered by the last user action, but takes into account the whole session history.

From each situation, a set of *suggestions* is derived. One of the most important suggestions is terminological help, but IRA suggestions also include simple *hints*, that merely make aware the user of alternative actions (like reminding the user to have a look at the full text of the documents, or to judge, by clicking on the appropriate button, the relevance of the read documents), and more complex *advices*, i.e., a set of operations which are carried out collaboratively by the user and IRA (like author search, that suggests to look for documents written by the same author as the documents already judged relevant by the user). IRA suggestions are always contextual and are provided in two kinds of situations: *critical* (i.e., the user is experiencing some problem, as repeatedly retrieving no documents, or not making progress) and *enhanceable* (i.e., when the user could follow other – possibly more – appropriate alternative routes). Finally, IRA suggestions are ranked and proposed to the user.

We performed a laboratory evaluation that highlighted some positive qualitative results: the sample users that used IRA were satisfied with the adequacy, timeliness, comprehensibility, and usefulness of the suggestions. Moreover, as foreseen, terminological help has been especially appreciated.

4. Conclusions and Future Work

In this paper we have shown how persistent and ephemeral personalization techniques can be exploited to implement more adaptive and effective information access systems. More specifically, the research presented here approaches two problems of the user of a scholarly publishing system: the need to be timely and accurately updated about new relevant information and the request for adequate, effective and easy-to-use support during search of archive information. Several experimental results show that persistent personalization is useful for information filtering systems, and ephemeral personalization leads to more effective and usable information retrieval systems.

References

1. F.A. Asnicar, M. Di Fant, C. Tasso, User Model-Based Information Filtering, in M. Lenzerini ed. *AI*IA 97: Advances in Artificial Intelligence – Proc. of the 5th Congress of AI*IA*, LNAI 1321, Springer, Berlin, D, 1997, 242-253.
2. R. Baeza-Yates, B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley, New York, NY, USA, 1999.
3. N.J. Belkin, Helping People Find What They Don't Know, *Comm. of the ACM* 43(8), 2000, 59-61.
4. N. Belkin, C. Cool, D. Kelly, S.-J. Lin, S.Y. Park, J. Perez-Carballo, C. Sikora, Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval, *Information Processing and Management* 37(3), 2001, 403-434.
5. N. Belkin, C. Cool, A. Stein, U. Thiel, Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems, *Expert Systems with Applications* 9(3), 1995, 379-395.
6. D. Billsus, M.J. Pazzani, User Modeling for Adaptive News Access, *User Modeling and User-Adapted Interaction Journal* 10(2-3), 2000, 147-180.
7. G. Brajnik, S. Mizzaro, C. Tasso, Evaluating User Interfaces to Information Retrieval Systems: a Case Study on User Support, *Proc. of the 19th Annual International ACM SIGIR Conference*, Zurich, CH, 1996, 128-136.
8. G. Brajnik, S. Mizzaro, C. Tasso, F. Venuti. Strategic help in user interfaces for information retrieval, *J. of the Am. Soc. for Information Science and Technology*, 53(5):343-358, 2002.
9. G. Brajnik, C. Tasso, A shell for developing non-monotonic user modeling systems, *International Journal Human-Computer Studies* 40, 1994, 31-62.
10. W.B. Croft, S. Cronen-Townsend, V. Lavrenko, Relevance Feedback and Personalization: A Language Modeling Perspective, *DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries*, 2001, www.ercim.org/publication/ws-proceedings/DelNoe02/.
11. E.N. Efthimiadis, Query expansion, *Annual Review of Information Science and Technology (ARIST)*, M. E. Williams ed., vol. 31, 1996, 121-187.
12. U. Hanani, B. Shapira, P. Shoval, Information Filtering: Overview of Issues, Research and Systems, *User Modeling and User-Adapted Interaction* 11(3), 2001, 203-259.
13. P. Ingwersen, *Information Retrieval Interaction*, Taylor Graham, London, UK, 1992.
14. A. Jameson, C. Paris, C. Tasso eds., *User Modeling – Proc. of the 6th Intl. Conference UM97*, Springer-Verlag, Wien New York, 1997.
15. T. Malone, K. Grant, F. Turbak, S. Brobst, M. Cohen, Intelligent information sharing systems, *Comm. of the ACM* 43(8), 1987, 390-402.
16. R. Mandala, T. Tokunaga, H. Tanaka, Query expansion using heterogeneous thesauri, *Information Processing & Management* 36, 2000, 361-378.
17. M. Minio, C. Tasso, User Modeling for Information Filtering on Internet Services: Exploiting an Extended Version of the UMT Shell, *UM96 Workshop on User Modeling for Information Filtering on the World Wide WEB*, Kailua-Kona, Hawaii, USA, January 1996.
18. S. Mizzaro and C. Tasso. Ephemeral and Persistent Personalization in Adaptive Information Access to Scholarly Publications on the Web. In Paul De Bra and Peter Brusilovsky editors, Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, In press. Malaga, Spain, 29-31 May 2002.
19. C. Tasso, M. Armellini, Exploiting User Modeling Techniques in Integrated Information Services: The TECHFINDER System, in E. Lamma and P. Mello eds., *Proc. of the 6th Congress of the Italian Association for Artificial Intelligence*, Pitagora Editrice, Bologna, I, 1999, 519-522.
20. C. Tasso, P. Omero, *La personalizzazione dei contenuti Web: e-commerce, i-access, e-government*, Franco Angeli, Milano, I, 2002.

Building Thesaurus from Manual Sources and Automatic Scanned Texts

Jean-Pierre Chevallet

Laboratoire CLIPS-IMAG
385 avenue de la Bibliotheque
B.P. 53 38041 Grenoble Cedex 9, France
`Jean-Pierre.Chevallet@imag.fr`

Abstract. This paper describes the work done in the TIPS project about the construction of a thesaurus base. This construction is a merge from a thesaurus manually built and one automatically extracted from large text corpora. Several manually built thesaurus have been semi-formatted to be merged in a consistent common base. The automatic extraction is based on both syntax and statistics. We present in this paper the way thesaurus are built and the results on Scientific corpus in the context of the TIPS project.

1 Introduction

The TIPS project aims to offer an integrated tool in order to manage scientific documents. This system provides a searching tool that retrieves documents from a query proposed by a user. The IRA module (Information Retrieval assistant), proposes to guide the user in improving the retrieval results. The Terminological Tool is a part of the IRA module. The first aim of this tool is to provide help for the user at search time using a terminological database. An other goal is to improve the building of the query. For that we propose to the user an interface that enable to browse among a structured set of terms where some are indexing terms. Finally, browsing among a set of terms extracted from the actual set of documents, enables a perception of its content. In fact, it can improve the perception of the system answer to the user; because, in this way, browsing the set of extracted terms look like browsing summarization of the all corpus content. In this paper we present the building stage that leads to the construction of the terminological database with some links : we then have a semi structured thesaurus.

2 Thesaurus and indexing

By definition, "thesaurus" is the study of term usage in given domains associated to a human activity. There are thesaurus for medical domain, mathematics, computer science, etc. A term is a sequence of words used in a given domain and which makes sense in this domain. Terms then refer to concepts of this domain.

The "Quebec Terminological Base" (Base terminologique du Quebec, also known as "Grand Dictionnaire"), or WorldNet are a good example of general thesaurus. Therefore, thesaurus is on the domain knowledge side and it is used for domain description. A thesaurus is often a human manual activity because it requires human domain expertise. In technical and scientific domain, terms are often composed probably because it is the simple way to build new terms. We can also notice that a multi-term is less ambiguous than a single term. A thesaurus is a sort of terminological base: it is a collection of terms, plus a set of relations among them. In some ways a thesaurus can be a bridge from a terminological base to document indexing. It can be used as a normalization of indexing terms. An index term is used for document description. It is therefore on the document side if it is automatically built, or on the user side if manually chosen by librarian. To sum up, terms of a thesaurus are used to describe a domain, whereas index terms are about the description of document content. The role of an index term is also to discriminate documents in order to retrieve them using a term-built query.

As we can see, a term that belongs to Thesaurus seems very different from a term that is used for an indexing process. Nevertheless, in this project, we propose to use terms from Terminological base and from Thesaurus, in order to help document access.

In the following, we present the way we use thesaurus in the TIPS project. In the rest of this paper we use the word Thesaurus instead of Terminological base because we are not only interested by terms, but also by relations among them.

2.1 Using a Thesaurus in IR

A Thesaurus can be used in the indexing process. We have already tested this approach in TREC test collections (see [6]). The idea is to enhance the precision of indexing using precise multi-terms. There are major difficulties underlying this approach: the use of single and multi-term together raise a counting problem, because single terms are included in multi-terms. As index weighting is grounded on frequency measure, the discrepancy between the frequency of single and multi-term must be solved in a consistent way. Using multi-terms crosses the frontier between Statistical Text Analysis and Natural Language Processing. It is not possible and even not desirable to take into account all natural language phenomenon for IR purposes. On the other hand, it seems important to us to take into account syntactical variation (ex: "detector", "neutrino detector" "underground detector", "deep underground detector", "deep underground neutrino detector", etc), and to take into account also references and elliptical expressions (ex: "We uses a deep underground detector...", "this detector...", "it is used for neutrino") because it changes the way frequency has to be computed.

On the other side, thesaurus can be used at retrieval time. This thesaurus is presented to the user so he can choose terms among it. If these terms are extracted from the actual corpus, they can be used into the query. Structuring the thesaurus can help the user finding the right term in a given domain. The

drawback is of course information overload : user will have to browse among a huge set of terms. An other drawback is the discrepancy between available data through the thesaurus and the actual data stored in the index. In that situation, user could chose terms in the thesaurus that do not exist as index terms either because it is not a good index term (from the system point of view) or because the thesaurus does not cover the same domain as the one covered by the documents of the corpus. An other important reason is the inevitable increase of the term set : specially in scientific domain, every rise of new concept, every breakthrough in the technology is the occasion of new terms creation. At the same time, some terms tend to disappear as technology changes. A thesaurus has to follow this natural evolution. The better choice is to follow it from the sources which are scientific publications.

In the TIPS approach, we have decided not to use thesaurus at indexing time: indexing aspects are not fundamental in this project, and classical single term indexing has been chosen.

The TIPS portal proposes a thesaurus allowing to select possible query terms, and also to perceive the domain covered by the indexed corpus by browsing its content. This is possible because a lot of terms are directly extracted from inline document content.

Before going into details of the thesaurus construction in TIPS, we just mention some general facts about thesaurus construction.

2.2 Thesaurus construction

Manual thesaurus building is a hard task but in this way, one can guarantee a good quality of the collected terms. So we can present these data to the end user for browsing. Maintaining such a thesaurus up to date is also costly. On the other hand, automatic Thesaurus building is quite human costless but the quality is not guaranteed. It relies on the content of document sources and also on the Natural Language treatment implemented. Our goal in this project is to combine both approaches. We will compile manually-built data, and extract terminological knowledge from documents, and finally merge these two sets into a final structure that will be proposed for browsing. Our building steps are then the following:

Extract a terminological base from documents by means of automatic full text analysis;

Compile existing accessible thesaurus and terminological sources;

Validate and filtering automatically obtained terms by confrontation with manual thesaurus and by limited manual inspection;

Merge both data sources. In this step, one can propagate some information from manual thesaurus to automatic thesaurus like the known domain of a term.

Structuring the term set using and propagating extracted links from existing thesaurus, and by the computation of syntax variations.

Integrate the final thesaurus into TIPS portal through the Information Retrieval Assistant.

In the next section we go through these steps in detail.

3 Automatic thesaurus construction in TIPS

Thesaurus is extracted from full text by means of syntax analysis. In this part we detail terms extraction and structuring steps that define the automatic thesaurus construction. In this thesaurus, we have a generic relation based on syntactic variation. We also obtain a non typed relation (a sort of "see also") based on conditional concurrence probability which are known as Knowledge Discovery in Text techniques (KDT) [2]. We will not develop this aspect in this article as we don't have yet the results.

3.1 Term extraction

We have used our IOTA system for all tasks except the first one: the full corpus tagging using a part of speech tagger. We have used the Brill tagger [1] because our IOTA system accepts only French texts as input. Thus we have had to develop a coder from Brill tagger to our IOTA format in order to use the rest of our system for all of the other text treatments.

The second step is term extraction. It is based on part of speech templates. These templates are used to extract noun phrases. In English as in French, most of these phrases are about 2 or 3 full words long. Full words are nouns or adjectives. Longer terms are less frequent and are less numerous. It is useful to extract longer terms if we take into account term variation. If not, we then have two different terms that are synonym in the sentence context. In fact long terms (noun phrases) usually appears once and rather at the beginning of texts. Shorter version then appears in texts as variations of longer terms.

Knowing this linguistic fact, it seems then important to compute co-references between terms and also between terms and pronouns. In TIPS, we do not compute these co-reference paths. Our goal is only to extract and structure terms from the all corpus. Moreover ambiguity between two term variations is very rare because size length is a sort of guaranty against term homonymy. Hence we promote the resolution of term variation and then the co-reference phenomenon, not at the sentence level, but at the end of extraction, so at the corpus level. This approach enables us to use frequency term information to choose the right term variation. This is the next treatment detailed in the next part. This choice explains why we extracted full size terms and so why we do not limit ourself to 2 or 3 terms length. Here are some examples of extracted phrases related to the word "algorithm":

```
randomized bidding algorithm ADJQ SUBC SUBC  
optimal randomized bidding algorithm ADJQ ADJQ SUBC SUBC
```


pseudopolynomial time algorithm ADJQ SUBC SUBC
forward search algorithm ADJQ SUBC SUBC
algorithm for matrix multiplication SUBC PREP SUBC SUBC
simple dynamic programming algorithm ADJQ ADJQ SUBC SUBC
cubic time algorithm ADJQ SUBC SUBC
iterative algorithm ADJQ SUBC
simple polynomial time algorithm ADJQ SUBC SUBC SUBC
algorithm for query evaluation SUBC PREP SUBC SUBC

Terms are followed by the corresponding part of speech. In the next section we present the structuring of this set of terms that leads to the thesaurus.

3.2 Term structuring by means of syntax

We used two sorts of term structuring. One is based on syntax and cover the term variation phenomenon, the other is based on global document term concurrence and expresses a more general sort of term relation. There are some attempts to automatically acquire from text a given type of relation, like hyponyms [5]. Some other approaches uses context defined by syntax [3, 4]. The Sextant system, uses syntax dependences between noun/noun, noun/verb, and noun/adjective. The underlying hypothesis used is that terms sharing contextual dependencies are semantically related. This approach is not able to qualify the extracted relation. Other systems like Xtract [7] are only based on co-occurrence statistics computed into a five word windows.

For this project we have chosen the combination of two methods : one based on syntax and term variation, combined with one based on term co-occurrence in document using dependence probability.

The syntax driven structuring deals with the all set of full length extracted terms from the all corpus. The system tries to link terms using variation rules. A variation rule is a couple of two part of speech patterns. The left pattern is the trigger of the rule. A rule is fired if the input term matches the part of speech tag sequence of the pattern. The right pattern is the production part. It produces a shorter term by reordering and reducing the set of tags of the right pattern. Applying a rule produces a short reordered term. The goal of such a rule is to link two term variations: a larger and a smaller variation of terms. Here are some examples of such rules. For each rule, one have an example of derivation and the rule itself.

```
deterministic algorithm -> algorithm  
ADJQ SUBC <VGEN> 2 .
```

This rule expresses the variation from a term without the adjective that qualify the substantive. The right part of the rule is a sequence of part of speech. The left part is the sequence of word that are kept for the associated term. In this rule, we only keep the second word of the term.

positive acceptance probability -> positive probability
ADJQ SUBC SUBC <VGEN> 1 3 .

This rule illustrates a term variation by insertion of substantive.

probability distributions for sequences of every finite length
-> probability distributions
SUBC SUBC PREP SUBC PREP PREP ADJQ SUBC <VGEN> 1 2 .

This last example, shows a term split at a preposition.

In order to avoid combinatory explosion and production of meaningless terms, the system only attempts to link actually existing corpus extracted terms. Hence, in this approach we have to first extract all possible terms from all documents before the application of these rules. All these rules have been proposed if we have at least one good example of term variation. A set of derivation rule is then language dependent. Here is an example of linked terms produced in this way.

optimal randomized bidding algorithm for the case of multiple bidders
-> optimal randomized bidding algorithm
-> randomized bidding algorithm
-> bidding algorithm
-> algorithm

known optimal algorithm
-> optimal algorithm
-> algorithms

In case of two rules that can be fired simultaneously, we have a preference for the one producing the most frequent term. If both possible terms have the same frequency in text, we produce them both.

4 Results

In this part, we present some information about thesaurus and documents that have been treated in this project. First the list of available online thesaurus that have been treated and merged and then, some data about documents that have been analysed.

4.1 List of treated thesaurus

We have treated a list of seven thesaurus. These thesaurus have been chosen because they are related to domains that are present in the ArXiv document base, and second, because there where available on the web. We have extracted from them four relation types:

Generic is hierarchic relation. A term a is a generic of a term b if the meaning of a includes the meaning of b . Hence b is a specific term of a .

Synonym is used when a term a can be used in place of a term b .

Context is a relation that express that a term can be used in the context of an other term.

see is a general relation without a precise meaning. It is often called "seealso".

The table 1 sum up the results. We have found very few synonyms : in only one thesaurus. The context relation is also not very frequent (two thesaurus). The more common relation is generic and after the "see also". The set of term after merging is 13 809. Only 5% of terms are common. Finally, we have obtain an average of 2 relations per terms, which not very important.

Here is the list of thesaurus treated:

aa0 This Astronomy thesaurus is very important. It is composed of 2 846 terms. (<http://darmstadt.gmd.de/lutes/thesalpha.html>)

arxiv ArXiv is the organisation of the document repository. It is not really a thesaurus but rather a classification scheme for clustering documents in the base. We used 440 terms. (<http://arxiv.org/archive/>)

jhep is a very short list of terms on High Energy Physics (<http://jhep.sissa.it/JOURNAL/keywords.html>)

msc MCS is a thesaurus dedicated to mathematics. It is structured in three sub levels. (<http://www.ams.org/msc/>)

pacs PACS thesaurus is about physic and astronomy. It contains 4324 terms related to condensed matter physics, material science and microelectronics. We only used these sections of the PACS thesaurus. (<http://www.aip.org/pubservs/pacs.html>)

schlagw The SCHLAGW thesaurus is more a list of recommended indexing terms than a real thesaurus. We have extracted 1552 terms from it. (<http://www-library.desy.de/schlagw.txt>)

spires SPIRES is an important thesaurus. We used only the physics part. (<http://www.slac.stanford.edu/spires>)

We sum up in table 1 some figures about the treatment of the manual sources.

4.2 About the analysed documents

We have analysed quite all content of the ArXiv content. This base has been indexed for the TIPS portal demonstration. We have analysed about 300,000 English documents in latex format. All theses documents are splited into 40 categories and sub categories (see the ArXiv thesaurus above). In the table 4.2 we present some figures obtained on some of them. This table shows the following information:

Doc nb is the number of treated documents in the sub category. We can notice that some categories have very few documents compared to others.

Voc size is the number of single terms found in the collection of documents. We can notice this number is not directly related to the number of documents.

Table 1. Treated thesaurus

Thesaurus	Theme	See	Generic	Context	Syn	relation	term
AAO	astronomy	8 111	2 432	429	0	11 972	2 846
ARXIV	high energy physic	0	440	0	0	440	115
JHEP	high energy physic	0	124	0	0	124	126
MSC	mathematiques	1 450	4 971	0	0	6 421	4 810
PACS	astronomy, physic	488	3 836	0	0	4 324	3 912
SCHLAGW	physic	1 228	964	186	64	2 142	1572
SPIRES	physic	1 198	343	0	0	1 541	1 191
Total		12 475	12 810	615	64	26 964	14 572
Total	After Merging					26 964	13 809

Term nb is the total number of full length terms found. We can see the impressive amount of different terms found. These figures show that word combination produces between 5 and 6 times composed terms more than single terms. In fact it is not such a quantity as we know lot of terms are more than 3 words long.

Hapax is the ratio of terms that appears only once in the corpus. This figure is important because we notice that for every corpuses this value is stable. About 80% of composed terms appear only once !

Max frequency is the maximum frequency of terms. It means the maximum number of documents in which a term can appear. This value is always 3 or 4 times less than the number of document. This value is interesting because we can suspect a term to be useless if it appears in too many documents.

Variation is the number of relations that has been computed. Generally speaking, we notice that we do not have found a lot of relations regarding the number of terms extracted. This is probably due to a reduce set of rules. Theses rules have been set up in incremental and manual ways. We do not know exactly how many rules are useful to cover the maximum of interesting term variations.

Relation is the number of terms that are found in the variation relation. Again we note an important loss of terms due probably to a lack of relation rules.

5 Conclusion

We have built for this project an important thesaurus related mainly to physics, astronomy and mathematics. We have produced a very huge amount of terms

Table 2. Analyzed documents

Theme	doc Nb	voc size	term nb	hapax	max freq	variation	relation
acc-phys	71	5 578	4 912	87.0 %	12	471	609
adapt-org	781	19 729	46 399	85.2 %	171	9 912	11 881
alg-geom	1 913	34 442	103 669	81.0 %	989	22 379	26 467
astro-ph	18 051	232 567	1 234 090	83.0 %	4 014	273 747	315 298
chao-dyn	3 762	58 528	257 239	83.7 %	1 150	59 364	68 624
cond-mat	62 973	388 581	3 426 576	83.0 %	21 568	618 998	712 604
computer	2 500	53 103	158 570	83.0 %	354	4 177	46 525
hep-ph	63 703	323 485	1 851 639	80.8 %	13 747	350 261	403 059
math	28 444	276 423	1 198 857	80.5 %	27 700	233 887	270 127

from the available scientific article of the ArXiv pre-print document base. Hence, we have proven that it is possible and useful to run some simple Natural Language techniques in order to automatically built a very important collection of terms in an automated way. The resulting user interface has not been tested with real users yet. The test done was only on a small set of terms. So we do not know at this moment, the pros and cons brought by the capacity of browsing through such a huge base of terms.

I thank Carole Bergamini for her help in extracting data from existing thesaurus and launching the processing of the latex file for the construction of the automatic built thesaurus. I thank also Christophe Hoang for the development of the part of the IOTA system that computes the syntactic variation and for the code that merge all data into one unique base of term and relation.

References

- [1] Eric Brill. English tagger. In <http://www.cs.jhu.edu/~brill/>.
- [2] R. Feldman and I. Dagan. Kdt - knowledge discovery in texts. In *Proceeding of the First International Conference on Knowledge Discovery KDD'95*, pages 112–117, August 1995.
- [3] Gregory Grefenstette. Use of syntactic context to produce term association list for text retrieval. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM press Copenhagen, Denmark*, pages 89–97, 1992.
- [4] Gregory Grefenstette. Automatic thesaurus generation from raw text using knowledge-poop techniques. In *Making sense of Words 9th annual Conference of the University of Waterloo Centre for the Oxford English Dictionary and Text Research, Cambridge*, pages –, September 1993.
- [5] Marti Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistic, Nantes, France, July 1992*.
- [6] Nie Jian-Yun and Chevallet Jean-Pierre. Using terms or words for french information retrieval ? In *Text REtrieval Conference 1997 (TREC-6), Gaithersburg, Maryland, USA*, pages 457–462, November 19–21 1997.

- [7] F. Smadja. Retrieving collocation from text : Xtract. In *Computational Linguistics*, pages 143– 177, 19(1) 1993.

QCT and SF services in Torii: Human Evaluations of Documents Benefit to the Community

Nathalie Denos

CLIPS-IMAG 385, rue de la Bibliothèque - B.P. 53
38041 Grenoble Cedex 9 – France
Nathalie.Denos@imag.fr

Abstract. This paper describes two services of the Torii portal dedicated to the High Energy Physics research community, and developed within the context of the TIPS European project. These services both relate to the reuse of evaluations performed by humans on scientific publications. The first one, called QCT (Quality Control Tools) aims at collecting human detailed evaluations of documents in order to enrich the traditional topical indexing of documents with quality-related information. The second one, called SF (Social Filtering) integrates a push functionality as a alternate and complementary tool to traditional pull services such as Information Retrieval; documents are pushed to users with respect to the evaluations they have made in the past, and as compared to other users' evaluations.

1 Quality Control Tools and Social Filtering

This paper describes two services developed for a community portal, in the context of TIPS european project. After setting the context, we describe each of these services.

1.1 TIPS Project

1.1.1 Purpose of the Project

TIPS (Tools for Innovative Publication in Science) is a european project funded by the European Community (IST-1999-10419) for the period February 2000 to July 2002. It involves the High Energy Physics research community (HEP), that is in need of tools to access the great quantity of available information, especially the scientific articles that are published everyday in this domain.

The aim of the TIPS project is to develop tools that help researchers in their daily activities: search, and read articles, write and disseminate papers, communicate with other researchers.

1.1.2 Services in Torii Portal

The TIPS project has developed a Web portal in order to offer services to the HEP research community. The portal, called Torii, has two aims:

- centralize and facilitate the access to various third part resources, such as open archives of scientific publications, and
- serve as a personal desktop where a given researcher can find the most relevant information for him.

The Grenoble partner in TIPS project (CLIPS-IMAG research laboratory) is in charge of two services: Quality Control Tools, and Social Filtering. Both of these services are related to the evaluation of scientific publications.

1.2 Evaluation of Scientific Publications

The two services are related to the general idea of the evaluation of scientific publications. In their everyday activity, researchers need to evaluate articles that they read. Such evaluations occur in two different contexts: peer review, and evaluation for personal use.

1.2.1 Peer Review

The peer review activity is a widespread process for evaluating the quality of research articles. For instance, when a journal calls for papers to be published in a new issue, the editorial board mandates researchers to review the submitted papers in order to decide whether an article reaches the quality standards required to be published in the journal. The reviewers are selected with respect to their personal research skills, that determine their ability to evaluate the article.

Reviewers are generally asked to fill an evaluation form that reflects the quality standards of the journal or conference. These quality standards can vary depending on the aim of the journal or conference.

1.2.2 Evaluation for Personal Use and Social Reuse

Another context where evaluation occurs is the evaluation that a researcher does when he reads an article that he retrieved from some archive. When a researcher searches for information, he retrieves a number of articles selected on the basis of the topic that the articles deal with. Among the retrieved documents, some will be more relevant than others, because of quality-related features of the document.

To select the most relevant ones, the user needs to read the documents in order to elaborate his own opinion on the document. This opinion generally encompasses several criteria, that are likely to be shared by other users with the same information needs.

This evaluation process is generally not formalized, as the user does it for his own use. In some cases, it can be made partly explicit by the user when he tries to organize the set of interesting documents, not only by topic, but also by the extent to which it can be useful in some situation: for instance as a reference to quote in an article, or as a reference to recommend to students, or to research colleagues.

1.3 Overview of QCT and SF Services

1.3.1 QCT

The Quality Control Tools (QCT) aim at collecting the evaluations that researchers have done, in order to enrich the traditional indexing information that is associated with documents for retrieval. Documents are generally indexed by information relating to topics. With human participation, quality-related information can be added, that is highly valuable to help subsequent users in their selection of the most relevant documents.

In Torii, a user can evaluate any document retrieved via the portal, by using an evaluation form. The portal keeps track of the evaluations made by any user, and combines them into statistical indicators that are displayed together with the document.

In order to easily produce ad-hoc evaluation forms, a tool to produce new forms has also been developed. This tool is dedicated to the editorial boards who would like to formalize the quality standards of their publications.

1.3.2 SF

Social Filtering (SF), also known as Collaborative Filtering, automatizes the process of people recommending documents to other people that have similar interests. A typical social filtering system pushes documents to a given user on the basis of the evaluations that he has made in the past, as compared to the evaluations other people have also made.

In Torii, a user can evaluate any document that he encounters in the portal with a single click, and the system pushes relevant documents into a dedicated subfolder of his personal folder.

In the remainder of this paper, we describe the QCT and SF services that are available in Torii.

2 QCT in Torii

The Quality Control Tools collect user evaluations of documents and combines them into statistical indicators, that are in turn made available to all users.

2.1 Evaluation Criteria

A study has been conducted in order to identify important quality criteria for the evaluation of scientific publications.

Quality criteria are criteria that allow selecting a subset of documents out of a set of topically relevant documents. They allow users to specify quality standards when they search documents. Quality features are the properties of documents that are associated to quality criteria. For a given document, a quality feature must be assessed to allow for later searching along the corresponding quality criterion.

We have identified the following 9 top-level quality features (for more details, see document UR-R1-QCT1 from the TIPS project Web site <http://tips.sissa.it>). The full set of candidate quality features is given in figure 1.

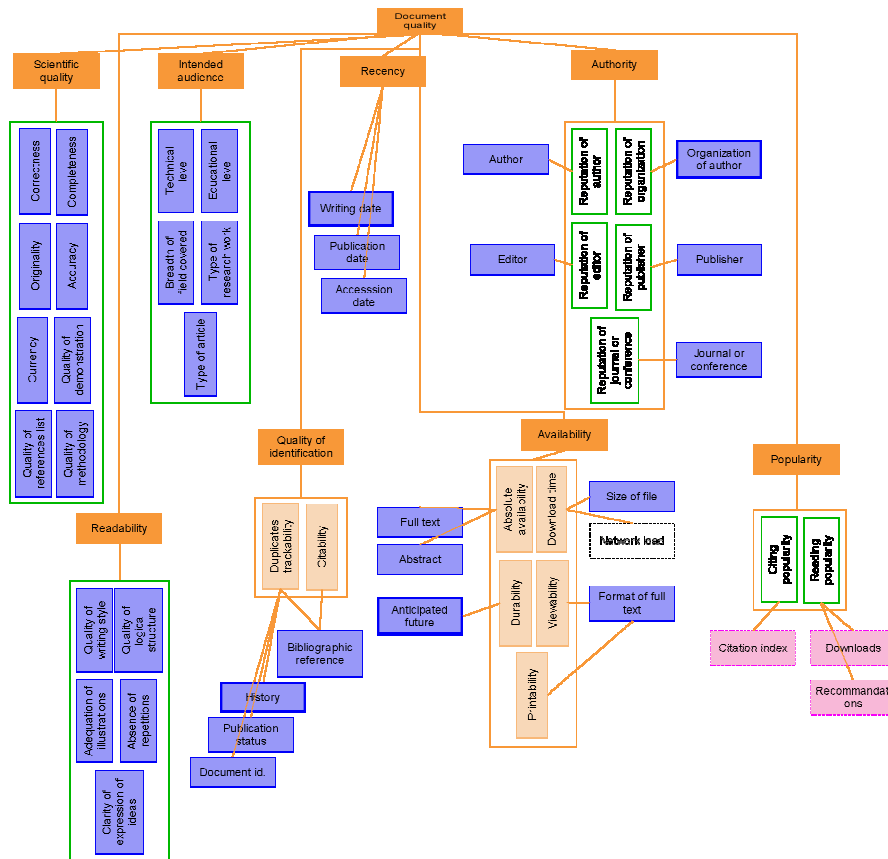


Fig. 1. Hierarchy of the full set of quality features. There is a wide variety of quality features. For instance, “correctness” is a quality feature that reflects a strict meaning of the word “quality”, as it vehicles a clear-cut judgment: it is always bad for a document not to be error-free. On the opposite, “recency” is reflects a less clear-cut meaning for the word “quality”, as a document may be a good one, although it is not a recent one.

2.2 Evaluation Forms

Evaluation forms reflect the quality standards that depend on the intended use of the evaluation. In Torii, the emphasis was made on the evaluations as made by a simple

reader, for his own use. A standard default evaluation form was defined with the help of experts in the domain, for it to be easily understood and filled up by users. An additional functionality was also provided to account for the fact that users are not always willing to formalize their evaluations. This is the “free comment” functionality, that allows a user to comment freely on a document.

Besides this, the peer review context was also accounted for, with a tool that allows to create ad-hoc forms for more specific intended uses.

We present here the general structure of an evaluation form via the presentation of the user interface to create forms, the standard default evaluation form, and the “free comment” functionality.

2.2.1 Creating Forms

In Torii, the definitions of evaluation forms are stored into a database, in order to automatically generate the corresponding forms.

We first describe the general structure of a form.

- A form is a set of “scaled criteria”.
- A scaled criterion is a criterion associated with a “scale” which can be a set of predefined values, or a free text field.
- Each criterion can have a set of subcriteria, which are also associated with a scale.
- The predefined values of a scale can be associated with numbers when it is relevant to order these values.

Figure 2 shows a sample of user interfaces for creating forms.

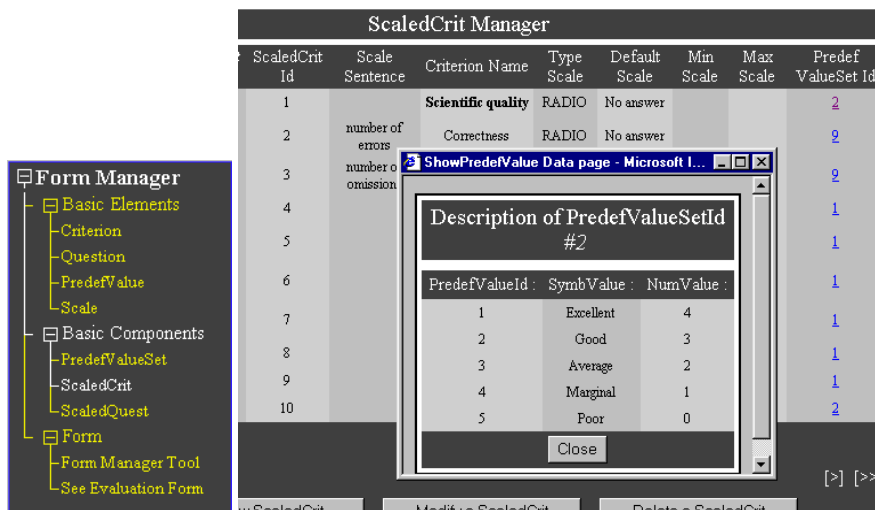


Fig. 2. Sample of interfaces for the creation of a form.

The administration tool allows to:

- Create new predefined values and associate numerical values to them,
- Group predefined values to build a scale,
- Create a new criterion, and optionally specify whether it is a subcriterion of another criterion,
- Associate a criterion to a scale (to produce a scaled criterion),
- Group criteria to build a form.

2.2.3 Standard Default Form

Figure 3 shows the default evaluation form that is prompted when a user wants to evaluate a document in details.

The screenshot shows a web browser window with the following content:

About

Document ID : oai:JHEP:042002025
 User ID : clips
 Date/Hour : 2002-05-06 17:49:09

Evaluation

Research quality Gives a global measure of the quality of the scientific contents presented in the article	✓	X
<input type="radio"/> Excellent <input checked="" type="radio"/> Good <input type="radio"/> Average <input type="radio"/> Marginal <input type="radio"/> Poor <input type="radio"/> No answer		
• Validity Scientifically sound and not misleading		🔑
• Importance New results		🔑
Presentation Well organized, clearly written, appropriate length,...	✓	🔑
Intended audience		X
<input type="radio"/> Broad interest <input type="radio"/> Narrow interest <input checked="" type="radio"/> No answer		

Fig. 3. Default evaluation form for the simple reader. The forms interface allow users to expand and collapse the scale for each criterion. The right-hand column recalls to the user whether he has already evaluated a criterion or not.

For the standard default form, to be used by simple readers as opposed to mandated reviewers, a selection of the most important quality features has been made, with the help of researchers in the HEP domain.

Three criteria remain: “Research quality”, “Presentation” and “Intended audience”. Research quality can be detailed along two subcriteria: “Validity” and “Importance”.

2.3 Free Comments

Users can attach free comments to the documents they access via the portal. A given user can edit a private comment, only visible by himself, and a public comment, accessible to any other user.

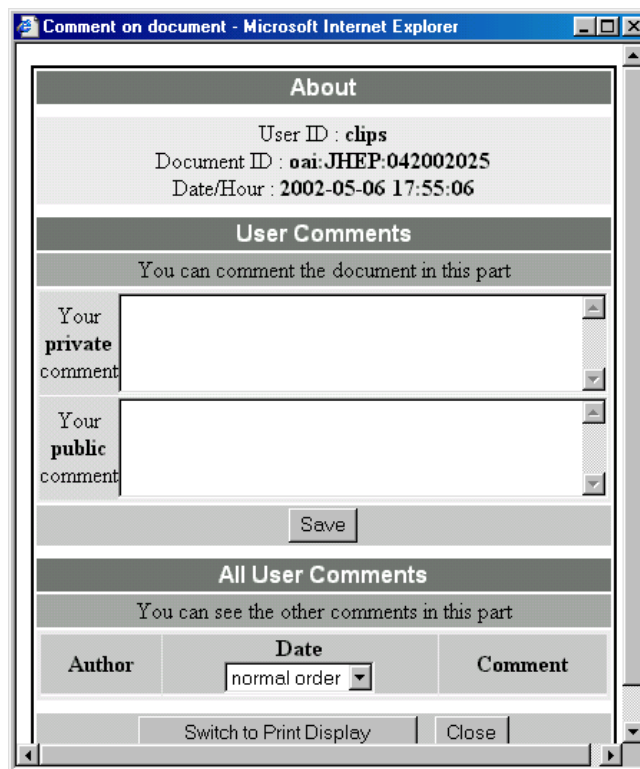


Fig. 4. Free comments functionality. A user can both edit his own comments and view other users comments in this window.

2.3 Statistics

Statistics can be drawn from users evaluations of documents (Number of Evaluations, Research Quality, Presentation, Intended Audience). These statistics are then displayed when a user accesses a document via the portal.

Towards the realistic fermion masses with a single fam

Author: M. V. Libanov and E. Ya. Noughev

Abstract: In a class of multidimensional models, topology of hierarchical masses and mixings in the effective for vector-like generation. We carry out numerical sim masses and mixings in one of these models.

Subject: Extra Large Dimensions

Source: hep-ph/0201162

SPIRES: [References](#), [Citations](#)

Evaluations

QCT

Of Evaluations :
 Research Quality :
 Presentation :
 Intended Audience :

Fig. 5. Display of the QCT statistics in the document view.

3 SF in Torii

The Social Filtering service in Torii pushes documents into the personal folder of an authenticated user.

3.1 General Principle of Social Filtering

The profile of a given user is the set of document identifiers associated with the evaluation that the user has given in the past for these documents.

For a given user U , the filtering engine computes a score for each document that is new to U , but has already been evaluated by one or several other users. This score is the prediction that the system makes of how much user U will like the document. It accounts for the evaluation of the other users, weighted by the similarity between the profile of these users and the profile of user U . The engine only pushes the documents for which the prediction is over a given threshold.

In Torii, the engine follows the Pearson-R correlation method for the computation of profile similarity, and the standard memory-based algorithm for the computation of the ranking (see for instance [1]).

3.2 Torii Interfaces for Social Filtering

The Social Filtering service allows users to indirectly define their profile, via a rating of documents on a single scale with the meaning “I like it”.

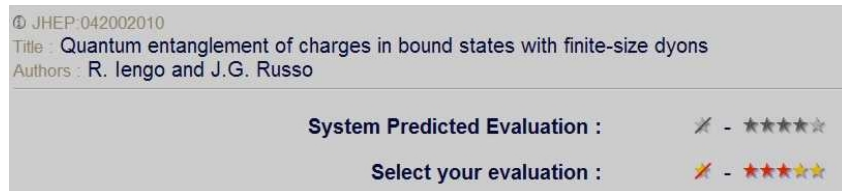


Fig. 6. User interface for the rating of a document to be used for social filtering. The number of red stars indicate how much user U likes the document, and the grey stars indicate the system prediction of this value, when available.

Users can check the documents that have been pushed by the system into a specific subfolder of their personal folder.

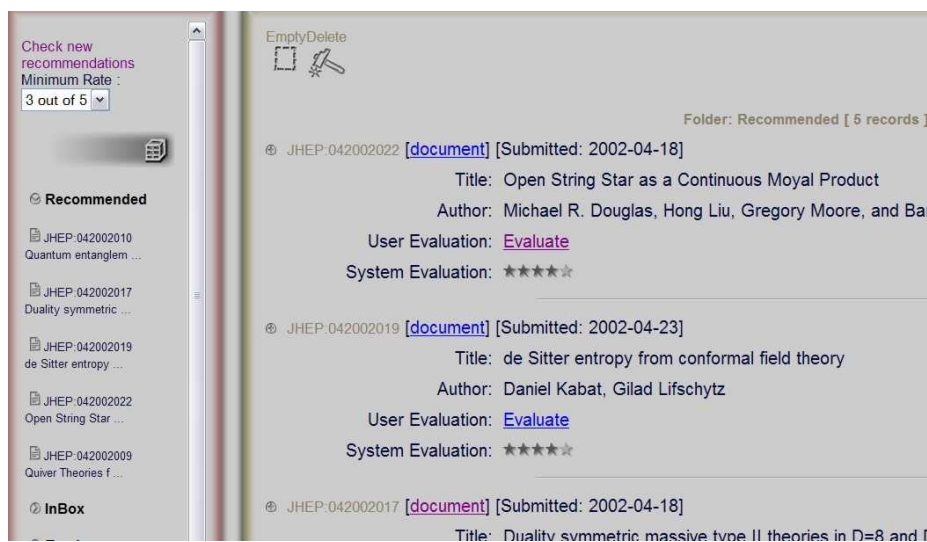


Fig. 7. The “Recommended” subfolder, that contains the documents pushed by the system.

4 Conclusion and perspectives

The two services presented in this paper increase the possibility of reusing individual evaluation efforts for the benefit of a community. They tackle the information

overload problem that researchers encounter, by encouraging and amplifying some of the naturally arising social phenomena.

For the time being, Torii integrates several services dedicated to the personalized and assisted access to information. QCT and SF are two of them, but cognitive filtering and assisted information retrieval are also available. Cognitive filtering, based on the topical content of documents, also requires evaluation feedback from the users, as well as information retrieval (the well-known “relevance feedback”). A stronger integration of these related services will be studied for the future, in order to allow users to better benefit from the feedback that they provide in these various contexts.

References

1. Resnick P.,Iacovou P., Suchak M., Bergstrom P., Riedl J.: GroupLens: An Open Architecture for Collaborative Filtering of Netnews, Proceedings of ACM Conference on Computer Supported Cooperative Work, pp. 175-186, 1994.

Toward conceptual indexing using automatic assignment of descriptors

Arturo Montejo Ráez

CERN - European Laboratory for Particle Physics, Geneva, Switzerland
Data Handling Group

Abstract. Indexing techniques have reached a well matured state. Digital libraries and other digital collections make an intense use of these algorithms to store and retrieve documents. In the other side, we have browsing techniques, which lets the user to gather the information. Current approaches are not yet advanced enough in order to satisfy the user. At CERN we are working in an indexer based on thesaurus descriptors. With a collection of documents related to thesaurus, user can manipulate them in a more conceptual way. Here we describe the core of this system, the automatic descriptor assigner.

1 Introduction

Indexing techniques has, mainly, focused the attention of *Information Retrieval* (IR from now) researchers, because it was clear that they represents one main problem to be solved and optimized. It is easy to understand such tendency, since indexing has a vast repercussion on the rest of components in an information retrieval system. From older works in IR by Rijsbergen [19] and Salton [16], we can summarize the accessing to the information in this case of use:

1. The user has a specific need of information.
2. The user transmits his/her need of information to the system.
3. The system access the collection and retrieves to the user a set of documents.
4. The user browses the collection and returns a feedback to the system.
5. The system gets feedback from the user so it can perform a better search.
6. The dialog user-system finishes once the user is satisfied with the results obtained.

The well-known *full-text search* is considered as a “philosophal stone” for the implementation of this dialog. We can find several query languages that can enhanced this type of search to enable the user specify more detailed queries. The main problem is the ambiguity of the query, that is, the not trivial task carried out by the user when determining his/her requirements. Some approaches try to solve the problem providing enhanced interface for browsing, and seems that a mixing of good ranking algorithms like *PageRank* [13] together with more intuitive and fast browsing tools, like clustering the result set to make easier the discrimination by the user [14], are in the path to the most suitable solution.

In research environments like CERN, the browsing of documents can be a complex task which involves the gathering on a very large collection. We are working in the developing of more semantic tools which will provide to researchers the ability of jumping from one document to another in a conceptual basis. We use the DESY thesaurus [6] to assign descriptors to High Energy Physics (HEP) related documents. In that way we are adding meta-data which tell us about the semantic content of the document. Since all the descriptors are belonging to a structured thesauri, documents are, therefore, interrelated. We could think in a network in the aim of the *Semantic Web* proposed by Tim Berners Lee et al. [4], but in a very well domain, our digital library.

2 Adding semantic meta-data to the document. Descriptors

Some articles do contain some subject information supplied by the authors (usually only when the journal makes it a condition of publication!). So some journals do have keywords, and quite a few have adopted the PACS classification supported by the American Physical Society [12]. However, these approaches are far from being complete over all documents, so they are not useful for global searching. Therefore, any added data have to be supplied by the creator of the database used for the searching.

This kind of adding of subject material is called subject indexation or keyword¹ enhancement. There are two very different ways of doing this: to choose terms from a fixed thesaurus or to use keywords which can be chosen by the indexer at will. The efficient allocation of keywords from a fixed thesaurus makes the most demands on the indexer, as the documents have to be well understood. The indexed terms may not even appear in the text at all, which can give this method a big advantage over any strategy which just uses the text of the document. Examples of fixed thesauri are those used by INIS [1] (International Nuclear Information System, Vienna) and INSPEC [2] (Physics, Computing and Electrical Engineering Abstracts, UK). as well as the DESY Thesaurus.

Many new documents arrive every day at the CERN Library, nearly all of them in electronic form. The task of indexing is mainly performed by indexers working at DESY. Due to the growth in the production of HEP-related papers a new approach to assignment has been developed. Since full-automatic indexers are still far from providing a realistic solution, a computer-based help tool for indexing might be able to be used in order to ease the work of human indexers.

The *HEPindexer* project intends to propose a preliminary solution, opening the door to research on automatic indexing tools in the area of HEP. This tool proposes descriptors for a given document. In the development of such a system a first step has been achieved: the generation of *main DESY keywords*. These descriptors are generated following a statistical approach [19].

¹ Please note that here, the use of the terms *keyword* and *descriptor* are interchangeable.

```

*coherent interaction
  coherent state (for quantum mechanical states)
  cohomology
*coil
-coincidence ('fast logic' or 'trigger' or 'associated production')
-Coleman-Glashow formula (baryon, mass difference)
-Coleman-Weinberg instability (symmetry breaking)
*collective (used only in connection with accelerators)
*collective phenomena ('field theory, collective phenomena' or
'nuclear physics, collective phenomena' or 'nuclear matter,
collective phenomena')
-collider ('storage ring' or 'linear collider')
  colliding beam detector (use only in instrumental papers)
*colliding beams (for accelerator use 'storage ring' or
'linear collider')
  color (for colored partons)
  colored particle
  communications

```

Fig. 1. Extract from DESY thesaurus

Figure 1 shows an extract from the DESY thesaurus. Descriptors labeled with “*” are descriptive (secondary) keywords; those with “-” are non-keywords, while those preceded by a blank are main keywords.

3 Previous work

The availability of large collections of documents in full text format has represented the beginning of a new era in information retrieval. Much research is being done around natural language processing. The early work of Salton [17] provides a good introduction. Many relevant algorithms have arisen for this approach, from classic conflation algorithms to reduce the representation of a document to its essential items (see [15]), to those which treat the document as a whole, identifying discourse trees [11] or conceptual phrases [5].

In the pure sense of descriptors assignment we identify two different tendencies: those ones where the goal is the use of descriptors by humans, and those ones where descriptors are intended to be used by other computed tasks. For the first ones we can cite some systems that have been developed during past years, such as BIOSIS, MeSH, the NASA MAI System [10]. For the second use, we have approaches like the probabilistic one of Reginald Ferber [9] and some multilingual approaches like the indexer used in the European Commission [18] for cross-lingual purposes and the MAGIC system of Kutschekmanesch et al. [8].

For us, the use of the descriptors is a mixture of both tendencies. They let us interrelated documents, and they let the user to gather the collection using them. Our system: the *HEPindexer*, is the core of all this, it will propose automatically descriptors for a given full text document.

4 HEPindexer

The algorithm used needs a set of data which must be [3] generated in a *training* process beforehand. Later, this system will be able to propose main descriptors with a reasonable degree of success, as proved through a *testing* process. These two processes require a set of documents as input. HEPindexer is supplied with the training collection of 3.700 documents. This collection was a sample of HEP-related documents and the DESY keywords were supplied for each document. That is, we have a list of documents already labeled by DESY from which our system can learn. After training, we are able to pass a new document to the system and receive as output a list of automatically-proposed descriptors.

4.1 Algorithm

The *training* consists on:

1. Each document is parsed, eliminating *stop words* (articles, prepositions and other words without meaning) and applying a *stemmer* (in order to get the “stem” of each word). Finally, the frequency of every remaining term in the document is computed.
2. For each descriptor, we compute a vector of terms using the following formula:

$$weight(k, t) = \lg \frac{M}{M_t} \sum_d TFIDF_{t,d} \bullet KF_k$$

where

$weight(k, t)$ is the weight of the term t for the descriptor k

M is the total number of descriptors

M_t is the number of descriptors related to term t (that is, the term t appears in M_t documents labeled with k)

d is a document

$TFIDF_{t,d}$ is the *document frequency* multiply by the *inverse document frequency* of the term t in document d

$KF_{k,d}$ is the infrequency of the descriptor k for the document d

The *assignment* of descriptors given a new document is performed ranking all descriptors in the thesaurus with a weight computed as follows:

1. The document is parsed, as in the training phase, to get a vector of terms by frequency.
2. The vector is multiply by the matrix of weights between descriptors and terms, given as result a vector of weighted descriptors.

4.2 Results

The system interacts with the user through a web-based interface. Using a web browser, the user can test the system with documents from the test collection or obtain proposed keywords by supplying a new full-text document, either in Postscript or PDF format, or plain text. Although the system can still only propose DESY main descriptors the results are close to 60% in both **precision** and **recall**. That means that an average of almost 60% of the keywords proposed are also contained in the list proposed by DESY, and that almost 60% of keywords proposed by DESY are the same as the returned ones by the system.

This system has now integrated within the CERN Document Server [7]. Improvements are still being made, as the project is only in its initial phase. Secondary keywords and more refined algorithms (using linguistic resources) are being studied in order to enhance the performance of the system.

5 Conclusions and future work

Some improvements on the system have to be performed yet. A new HEPindexer is now being programmed based on *Java* and *MySQL*. Other measures will be tested and it is planned to incorporate the capacity of dealing with *multi-words*. After it, the system will be ready to be integrated in a browsing tool for users, providing the feature of gather documents and citations with descriptors from the thesaurus as added values.

References

1. Inis thesaurus. <http://www.iaea.or.at/worldatom/publications/inis/inis.html>.
2. Inspec thesaurus. <http://www.iee.org.uk/publish/inspec/>.
3. R. Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Acm Press Series, 1999.
4. Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic Web. *Scientific American*, 284(5):34–43, May 2001.
5. Christopher Culy. An extension of phrase structure rules and its application to natural language. Master's thesis, Stanford University, 1983.
6. DESY. The high energy physics index keywords, 1996. <http://www-library.desy.de/schlagw2.html>.
7. CERN. DH Group, ETT division. The cern document server, 1996. <http://cds.cern.ch>.
8. Said Kutschemanesh et al. Automated multilingual indexing: A synthesis of rule-based and thesaurus-based methods. In Pub Deutschen Gesellschaft fur Dokumentation, editor, *Information und Markte*, pages 211–224, Donn, Germany, 1998.
9. Reginald Ferber. Automated indexing with thesaurus descriptors: a cooccurrence-based approach to multilingual retrieval. In Carol Peters and Costantino Thanos, editors, *Proceedings of ECDL-97, 1st European Conference on Research and Advanced Technology for Digital Libraries*, pages 233–251, Pisa, IT, 1997. Lecture Notes in Computer Science, number 1324, Springer Verlag, Heidelberg, DE.

10. Oak Ridge Gail Hodge. Cendi agency indexing system descriptions: A baseline report. Technical report, CENDI, 1998. <http://www.dtic.mil/cendi/publications/98-2index.html>.
11. Daniel Marcu. Discourse trees are good indicators of importance in text. Technical report, Information Science Institute, University of Southern California, 1997.
12. American Institute of Physics. Physics and astronomy classification scheme. <http://publish.aps.org/PACS/>.
13. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Computer Science Department, Stanford University, 1998.
14. Christopher R. Palmer, J. Pesenti, Raul E. Valdez, Michael G. Christel, Alexander G. Hauptmann, D. Ng, and Howard D. Wactlar. Demonstration of hierarchical document clustering of digital library retrieval results. In *ACM/IEEE Joint Conference on Digital Libraries*, page 451, 2001.
15. A. M. Robertson and P. Willett. Evaluation of techniques for the conflation of modern and seventeenth century english spelling. In Tony McEnery and Chris Paice, editors, *Proceedings of the BCS 14th Information Retrieval Colloquium, Workshops in Computing*, pages 155–168, London, April 13–14 1993. Springer Verlag.
16. G. Salton. A vector space model for automatic indexing, 1975.
17. Gerard Salton. Automatic text analysis. Technical Report TR69-36, Cornell University, Computer Science Department, June 1969.
18. Ralf Steinberger. Cross-lingual keyword assignment. In L. Alfonso Ure na López, editor, *Proceedings of the XVII Conference of the Spanish Society for Natural Language Processing (SEPLN'2001)*, pages 273–280, Jan (Spain), September 2001.
19. C. J. van Rijsbergen. *Information Retrieval*. London: Butterworths, 1975. <http://www.dcs.gla.ac.uk/Keith/Preface.html>.

Digital content sewed together within a library catalogue WebLib - The CERN Document Server

Jens Vigen

CERN, Geneva, Switzerland

Abstract. Aggregation, harvesting, personalization techniques, portals, service provision, etc. have all become buzzwords. Most of them simply describing what librarians have been doing for hundreds of years. Prior to the Web few people outside the libraries were concerned about these issues, a situation which today it is completely turned upside down. Hopefully the new actors on the arena of knowledge management will take full advantage of all the available "savoir faire". At CERN, the European Organization for Nuclear Research, librarians and informaticians have set up a complete system, WebLib, actually based on the traditional library catalogue. Digital content is, within this framework, being integrated to the highest possible level in order to meet the strong requirements of the particle physics community. The paper gives an overview of the steps CERN has made towards the digital library from the day the laboratory conceived the World Wide Web to present.

1 Setting the scene

CERN, the European Organization for Nuclear Research, is the world's largest particle physics centre, used by half of the world's particle physicists. These scientists, altogether 6500 users, represent 500 universities and over 80 nationalities. The CERN staff, who comprise just under 3000 people, have as their global aim to support these scientists in their research. CERN staff encompass a wide range of skills and trades - engineers, technicians, craftsmen, administrators, secretaries, workmen, ... and of course librarians who are there to meet all their information needs. The CERN staff design and build CERN's intricate machinery and ensure its smooth operation. Then help prepare, run, analyse and interpret the complex scientific experiments and carry out the variety of tasks required to make such a special organization successful. Constructing these highly advanced machines is an extremely costly operation [1]. The high costs imply that there is absolutely no financial room for research and development already carried out somewhere else in the world. This constrain lead t he particle physics community into a culture based on preprints to accelerate the communication process more than 40 years ago. Driven by the same requirements Tim Berners-Lee, a CERN computer scientist invented the World Wide Web, conceived and developed for the large high-energy physics collaborations which have a demand

for instantaneous information sharing between physicists working in different universities and institutes all over the world.

2 "In the beginning ..."

We are back in December 1990 and the Web was created. "... the earth was formless and empty, darkness was over the surface of the deep ..." and all the world's web servers were at CERN - maybe even in the same building. A prophet has however no honour in the prophet's own country, so instead of just moving one floor up, the later so famous web went off to Stanford Linear Accelerator Center where it shortly after was set up as a new gateway for accessing the bibliographic database SPIRES-HEP [2]. Looking back, information retrieval from SPIRES-HEP became the application that compelled the community to start using the Web. A few months later at Los Alamos National Laboratory, "in the middle of nowhere between two Indian pueblos", theoretical physicist Paul Ginsparg applied the new technology to set up arXiv, a system facilitating distribution of drafts to other theoretical physicists. 10 years later Ginsparg is often credited by his peers for having revolutionised scientific communication by setting up this system [3]. At present the archive contains 190 000 papers, augmented with some 100 new papers every day. ArXiv is estimated to distribute about 25,000 daily e-mail alerts and there are probably at least 35,000 distinct daily users via the web.

3 Back at CERN

In spite of not having been the first library in the world with web access to its catalogue, the various developments world wide were closely monitored by the CERN Scientific Information Service. Actually, an experimental initiative of scanning documents received on paper from laboratories and universities from all over the globe had started more or less at the same time, with the idea of diffusing the information to the whole of the community via the information networks. The bibliographic data was kept in the library catalogue while the fulltext was kept separately, so in order to retrieve a paper one first had to search for the key in the bibliographic database before one could navigate down to the paper itself on the preprintserver. [4]. January 1994 stands out as a paradigm change as from then onwards all new preprints were provided in an electronic format, a fully integrated service, or a "WWW GUI" as it so nicely was referred to at the time, would however not be launched until two years later.

4 Take-off for WebLib - the CERN Document Server

The times with no linking capabilities between metadata and fulltext was perceived as having lasted for ages, at least by the library staff who had spent quite

some energy in guiding lost readers through this initial period. There were however no major reason to complain, in 1996 most libraries still had not yet started thinking of having hyperlinks to external resources from their bibliographic catalogues - for most libraries there were basically no relevant materials to link out to, at least not materials considered to be relevant at the time. CERN librarians, having ALEPH as the in-house library system, realised quickly that the standard web interface of the system did not correspond to their ambitions of integrating digital content to the highest possible level. It was consequently decided to build a CERN specific interface, using extendable application program interfaces [APIs] enabling an expansion without expensive source-code modifications. The die was cast, WebLib, the CERN Document Server, was about to be implemented. To start with the "high ambitions" were made up of simply imagining a basic set of links between related records and links to the corresponding fulltext for preprints as the killer application ...

5 The publishers getting onboard

In parallel to the e-publishing activities in academia, all the major publishing houses were carrying out tests with the intention of launching electronic journals. The appearance of the web, which of course was a gift to anybody involved in electronic publishing, kind of wiped out all existing initiatives as being all of a sudden obsolete. "If there is no library system available or if it is decided to develop an entirely new system, the best option at this moment seems to be to use the very popular World Wide Web as user interface." was, not surprisingly, one of the conclusions from the TULIP project, initiated by Elsevier, in 1996 [5]. Institute of Physics (IOP) adapted quicker than the other publishers and was therefore the first publisher to offer its entire journal portfolio across the web in the spring of 1996. The launch was well received by an innovative library community and a group of enthusiastic scientists, in spite of a response time one would wish not to think about - even back in 1996. The new resources were warmly welcomed, although not yet integrated into the libraries traditional retrieval tools. The level of integration of the electronic journals was simply restricted to setting up links to the available titles from the various libraries' website. Nobody seemed to think about cataloguing an e-journal. Why bother with cataloguing as the counterparts on paper were already in the catalogue? One year later the American Physical Society (APS) entered the arena and now things finally started to happen in terms of integrating the electronic collections. For the architects behind the electronic version of Physical Review D it must have been clear from the very beginning that they had to facilitate direct access to any article in the journal by using a URL scheme similar to the scheme used for article identification in the traditional library. This was an enormous progress as it permitted users to access articles without having to navigate through a whole set of pages before getting to the desired information. For information collectors it was thus possible to automatically generate links from their metadata repositories to the corresponding articles, but so far only for articles published in

Physical Review D. The handling system developed at CERN to facilitate this feature was named "Go Direct" and caused quite some excitement, even though to start with it only could handle a few titles [6]. In addition to Physical Review D it could also handle some Springer titles using a set of cleverly thoughtout lookup tables which had to be manually updated whenever a new issue of the journal appeared. With "Go Direct" CERN had implemented SFX [7] - without knowing it - after all not so strange, SFX had still not yet been conceived ... Triggered by the enthusiasm shown by the library's avant-garde users, it was now time for the CERN librarians to start a systematic lobbying of all physics publishing houses. At every occasion letters were sent, interventions were made during seminars etc. with the aim to make the publishers introduce URLs for their journal articles based on the triplet journal, volume, page. Surprisingly enough the idea did not generate the same amount of excitement among the publishers as it did at CERN, however, a few months later it was silently implemented for all APS journals [8] and the APS Linkmanager has become a defacto standard for all publishers handling systems. The philosophy of the linkmanager is simple, but powerful: the triplet which has served the world as a unique, or at least close to unique, identifier for articles since the appearance of the first journals centuries ago, should always be associated to a persistent and robust URL.

6 Homework to be done

The publishers had started doing their homework, so then it was just necessary to keep on going with more innovations on the library side in order to maintain the pressure. Entering into the new era it was rewarding to see the importance of having fully streamlined metadata. With an ISBN associated to each record it became straightforward to create links to the corresponding Amazon records, records which more and more often contain samples of the books itself in addition to the table of contents, reviews etc. Without clean data no links would have been created - finally the fruits of years of librarians accuracy could be harvested. But how long was Adam in paradise? Library users were quickly getting acquainted and it was realised that in order to provide an effective service it would be indispensable to collect a maximum of metadata for each record and these data would have to be added to the database with the shortest possible delay.

In the case of CERN this meant starting to add publication references to all preprint records as soon as the papers became published. Up to that moment such references had only been added to papers originating from CERN, so more automation was clearly required in order to absorb the additional workload [9] [10]. To identify the preprints' published counterparts is not straightforward as no publishers are willing to give the correspondence between the articles they publish and the "original" preprint numbers. Publishers even claim that the correspondence is unknown to them, this in spite of the fact that several of the major actors (Elsevier, IOP, APS etc.) even seem to prefer, or at least appreciate, the possibility of picking up the fulltext of the submitted manuscripts from arXiv

... A comprehensive matching procedure, matching data from various sources of published material against the CERN collection of preprints, had therefore to be implemented. Matched records were updated with publication references, permitting "Go Direct" to automatically create links to the corresponding fulltext of the published articles.

7 Library system becomes CERN's main scientific information system

Not only the library world profitted from the developments of the digital media. Photographs, posters, newspapers etc. went all electronic, so they did at CERN. This technological step lead naturally so that what so far had been managed as isolated pieces of information, were all at a certain point brought into one single system, offering users the ability to search across all the various collections at once. Within this optic the CERN Document Server has moved from being a pure library catalogue, towards becoming the CERN "global" scientific information system. The collection covers for the time being preprints, books, periodicals, reports, photographs, press cuttings, posters, exhibition objects and much more - in the foreseeable future it will cover, or point to, all scientific information resources needed by any particle physicist or CERN staff member to perform her or his work efficiently.

8 Enhanced reader services

Having reached the "maximum" of links based on the metadata, it was time to investigate what could be done with the actual content of the documents. The policy change actually implied that the long way from mainly being a document depository to become a real subject portal had started - based on Avi Saha's definition : "A portal is a single integrated point of comprehensive, ubiquitous, and useful access to information (data), applications, and people." [11]

As researchers spend quite some time just retrieving referenced papers, it was obvious to the CERN developers that this would be a field with great potential for savings. A program for automatic extraction of the references, using the fulltext documents, was therefore implemented. The extracted data were later parsed through a "normalization filter" to be made compliant to standard notations. The result was loaded into a devoted part of the database and links were automatically created to the corresponding texts [12]. This operation generated in total about 2,1 million links to fulltext, a number which has since been augmented with about 1500 per day, as the extraction is now a part of the regular routines. For the astrophysics papers the results are particularly striking due to the fact that a significant part of the journal literature in the field is available in fulltext, also retrospectively [13]. Literally speaking nearly all references belonging to this collection have been converted into active links.

Search in fulltext has often been pushed forward as a replacement of costly indexing. At CERN both approaches were considered to be important for providing an efficient information retrieval service. The CERN library however never indexed its preprint collection in a systematic way due to the lack of resources, but entering the digital era opened new possibilities: Hosting a vast collection of electronic preprints made it natural to start experimenting with searching in fulltext. Ultraseek is used as the search engine and the fulltext search interface permits searching through more than 90,000 fulltext documents stored on the CERN Document Server. A fulltext search can retrieve the one and only document describing the most rare concept, but it might also retrieve lots of noise. So in parallel to providing the fulltext search, it was decided also to look into the area of automatic indexing. Automatic indexing can be considered to be a branch of automatic summarization, which aims at the generation of abstracts from fulltext documents. The developers at CERN went ahead and made the HEPindexer that proposes a preliminary solution which can open the way for further research into automatic indexing tools in the area of particle physics [14]. So far a first step has been achieved, namely the generation of main DESY keywords [15]. These keywords are generated following a statistical approach. Investing more resources in the development of the HEPindexer will certainly give results which will improve the precision and recall searching of the particle physics literature.

Having extracted keywords and references for each document, a range of possibilities for automatically connecting related documents are opening up. The system can then in accordance with predefined rules, propose other sets of documents to the user as a function of what the user him/herself and similar users, have consulted earlier. In the case of few or zero documents as a result of a research, the system should verify the possibilities for misspellings and propose alternatives when that would be considered as appropriate. The system should further propose to carry out the same query in other relevant databases by simply transmitting the search arguments, applying the different syntaxes, to the various search engines.

Given that the system has captured knowledge about the nature of a certain set of documents related to specific users, the system should further be used to add additional links to other relevant sources. I.e. new-comers to the subject field can be proposed links from all keyword to the corresponding concepts in Encyclopedia Britannica or in any other general source, while the experienced readers will be directed towards sources as the Review of Particle Physics. The system will in the near future also link out to non-bibliographic information as hompages of authors, publishers pages etc.

In principle links can be created to any related entity provided that one finds a scalable solution for introducing and maintaining the links.

9 From one central library to thousands of satellites

The pendulum is swinging back, having centralized services for years the CERN library is again establishing satellite libraries and even personalized libraries. These libraries are of course digital libraries, fully integrated on the readers desktop. So far the service is mostly restricted to information provision [alerts and searching] and private records management [personal e-shelves and loans]. Enhanced reading tools represent a path which is only partly explored so far, even if the links to the full text of the citations have a quite striking effect. What CERN Scientific Information Service have not yet started to explore is having authoring tools as a part of the library system. It is however clear, if librarians want to continue being advocates for integrating digital content, one has to start expanding in that direction, then the ring will be closed.

10 Conclusion

Integrating digital content is only partly a technical challenge, the main challenge is to get all involved parties "to speak the same language". However, if you just believe in it strongly enough, nothing will be impossible, even if the interest of a commercial publisher and a research library are quite different in spite of the fact that the end users are the same.

References

1. LHC Cost Review <http://info.web.cern.ch/info/LHCCost/2001-10-16/LHCCostReview.html>
2. Documentation of the Early Web at SLAC (1991-1994) <http://www.slac.stanford.edu/history/earlyweb/index.shtml>
3. An Online Archive With Mountain Roots New York Times, August 28, 2001, Tuesday <http://college3.nytimes.com/guests/articles/2001/08/28/864834.xml>
4. Electronic pre-publishing for world-wide access : the case of high energy physics Dallman, D P ; Draper, M ; Schwarz, S ; Interlending and Document Supply : 22 (1994) , pp.3-7
5. The TULIP Final report Borghuis, M et al. Elsevier Science, 1996 ISBN 0-444-82540-1 <http://www.elsevier.nl/homepage/about/resproj/trmenu.htm#ToC>
6. Link managers for grey literature Lodi, E ; Vesely, M ; Vigen, J ; CERN-AS-99-006 <http://weplib.cern.ch/abstract?CERN-AS-99-006> 4th International Conference on Grey Literature : New Frontiers in Grey Literature, Washington, DC, USA, 4 - 5 Oct 1999 Ed. by Farace, D J and Frantzen, J - GreyNet, Amsterdam, 2000. - pp.116-134
7. SFX - context sensitive reference linking <http://www.sfxit.com/>
8. Citing and Linking in Electronic Scholarly Publishing: A Pragmatic Approach Doyle, M 3rd ICCO/IFIP Conference on Electronic Publishing. 1999. Ronneby, Sweden. Smith, John W ed. ; Ardo, Anders ed. ; Linde, Peter ed. ; Washington DC : ICCO Press, 1999. - pp.51-59 ISBN 1-891365-04-5 <http://www5.hkrr.se/EIPub99.nsf/>

9. Automation of electronic resources in the Scientific Information Service at CERN Pignard, N ; Geretschlger, I ; Jerdelet, J ; High Energy Phys. Libr. Webzine : 3 (2001) , pp. 3 ;<http://library.cern.ch/HEPLW/3/papers/3/>;
10. Using Internet/Intranet Technologies in Library Automation Vesely, M Thesis : Univ. Economics Prague : 2000 ;<http://weblib.cern.ch/abstract?CERN-THESIS-2000-040>;
11. Application Framework for e-business: Portals Avi Saha IBM developerWorks, 1999 ;<http://www-106.ibm.com/developerworks/library/portals/>;
12. From fulltext documents to structured citations : CERN's automated solution Claivaz, J B ; Le Meur, J Y ; Robinson, N ; High Energy Phys. Libr. Webzine : 5 (2001) , pp. 2 ;<http://library.cern.ch/HEPLW/5/papers/2/>;
13. The NASA Astrophysics Data System ;<http://adswww.harvard.edu/>;
14. Experiences in automatic keywording of particle physics literature Montejo Rez, A ; Dallman, D High Energy Phys. Libr. Webzine : 5, (2001), pp. 3 ;<http://library.cern.ch/HEPLW/5/papers/3/>;
15. DESY. The high energy physics index keywords ;<http://www-library.desy.de/schlagw2.html>;