OmniPaper

Bringing Electronic News Publishing to a Next Level Using XML and Artificial Intelligence

Bert Paepen[1], Jan Engelen[1], Markus Schranz[2], and Manfred Tscheligi[3]

[1] Katholieke Universiteit Leuven
{bert.paepen, jan.engelen}@esat.kuleuven.ac.be
[2] Pressetext.austria, Technical University of Vienna,
schranz@infosys.tuwien.ac.at
[3] CURE (Center for Usability Research and Engineering)
tscheligi@cure.at

Abstract. In the last five years the Internet, intranets and search engines have brought unmanageable amounts of information to the average user's fingertips. Since this growth will only continue, it is vital that users are supported in converting this universe of information into improved productivity and opportunity instead of being swamped and paralyzed. The OmniPaper project is investigating ways for drastically enhancing access to many different types of distributed information resources. The key objective of OmniPaper is the creation of a *multilingual navigation and linking layer on top of distributed information resources in a self-learning environment*, thus providing a sophisticated approach to manage multinational news archives with strong semantic coupling, delivering to the user more than the sum of the individual service features.

1 Introduction

During the last decades, the amount of digital information has grown exponentially. The same holds for the number of computers and Internet connections. More and more information is becoming available in electronic form and its accessibility is, in terms of network presence, increasing rapidly. With this growth in availability, the need for information coupling has grown as well. Since it is physically becoming easier to compare information from geographically spread sources, the need for coupling information on a semantic level is on the rise.

Information and access to it are scattered. Electronic information is geographically spread throughout the modern world. It has numerous access methods, storage formats and information structures [2]. Many dozens of variations exist in operating systems and software necessary for handling it [8]. Countries in which information is physically stored all have their own legislation, bringing along different approaches how to handle information. And last but not least, the information can be stored in many different languages.

In spite of its unlimited possibilities in terms of access to information, the Internet is now becoming self-threatening. Because of the Internet's ever-growing diversity, the information overload is about to crush its users [5].

Information is the lifeline of decision making. The reasons, why users can not get the information are manifold. Often they are physical ones: in some cases, the information has not been digitally captured (e.g., a paper document that has not been scanned). In some situations, there is no network connection to bring the information to the consumer (e.g., at airports or in shopping malls or in traffic, or the software does not allow users to make the connection to the desired information). Both situations will become increasingly rare as time goes on and technology conquers the continents.

In most cases, however, information is available – indeed, available in abundance – and accessible. Users just can not locate the right information, they suffer "data overload" [5], a mismatch between the sheer amount of information and the time to select, read and make sense of it. "Infoglut", the overwhelming availability of information and data, has begun to measurably affect knowledge workers' productivity. Enterprises that understand how to navigate the information flood will have a distinct advantage over their competitors.

Nowadays, users have routine access to a huge number of heterogeneous and distributed digital libraries. To satisfy an information need, relevant libraries have to be selected, the information need has to be reformulated for every library with respect to its schema and query syntax, and the results have to be fused [6]. These are inefficient manual tasks for which accurate tools are desirable.

This paper explains the research work conducted in the OmniPaper project and is structured as follows: Section 2 defines the basic ideas and concepts of the OmniPaper project, Section 3 outlines the OmniPaper objectives, Section 4 explains the system architecture of the intelligent news archive prototype and modern approaches to satisfying information needs for the users. The current status of the project, scheduled activities and results, and planned achievements round up the paper in the conclusion.

2 The OmniPaper Project

Since the emerging boom of the Internet a lot of newspapers are being published electronically. In spite of this increasing availability of information, news items remain scattered throughout various archives, countries and languages. Furthermore, the distributed electronic information has different data structures, storage formats and access methods. Searching for news is still mostly done the "brute force"-way using full-text search robots and search result quality highly depends on the sophistication of the user's search input. The net result is that finding news from various international newspapers still is easier in an airport news stand than on the Internet.

The IST-funded European project OmniPaper (Smart Access to European Newspapers, IST-2001-32174) is investigating techniques to obtain a novel online news experience. These include XML- and Artificial Intelligence related technologies. The OmniPaper architecture starts from distributed news archives, all within different operating environments, database formats and indexing mechanisms. SOAP (Simple Object Access Protocol) [9] is used to create a uniform access method to these archives. Rich indexing and meta-data structures, such as Topic Maps [7] and RDF, make intelligent search possible. A cross-archive intelligent index (or 'knowledge layer') contains concepts, relationships between them and occurrences in different languages.

OmniPaper is investigating and prototyping both RDF and Topic Maps, allowing an empirical comparison between both techniques. Artificial intelligence is used to enhance the multilinguality and the richness of the knowledge layer. On the one hand, data mining software automatically extracts keywords from news articles; on the other hand, online user behavior enriches the knowledge layer using web mining techniques. The navigational and search behavior of users tells us something about the relationships that the user makes between concepts.

The OmniPaper news prototype will enhance the online news experience using various techniques. Multiple news archives will be made accessible through one multilingual user interface. Project partners provide information in various news management areas: an Austrian online news agency provides access to primary source material, written on daily basis by editorial staff, a Spanish newspaper clipping service with access to Spanish and English contents, and a Belgian newspaper clipping service, providing access to Dutch news for the multinational news archive.

Users will be able to enter search terms in their own language, getting results from news archives in different languages. Users will be able to navigate through the thesaurus (list of topics and subtopics), but also through concepts related to their query. Instead of performing full-text search, the prototype will search in a multi-archive knowledge layer containing highly structured and inter-linked meta-data. OmniPaper introduces the concept of "guided query", a query that can be widened or narrowed down. If a search term is too general, the guided query will

list the possible options, allowing the user to narrow down his search. If a search term is very specific, it can be interesting to let the guided query propose synonyms and related concepts, enabling search results that the user had in mind but did not type in exactly.

It is anticipated that the OmniPaper results will be sufficiently generic so that the technique can be broadened to other application domains such as libraries and museums.

3 OmniPaper – The Smartest European News Finder

The OmniPaper project is investigating ways for drastically enhancing access to many different types of distributed information resources. In addition to combining the multilingual aspects within European countries and the focus on user-centered information retrieval, even further concepts from AI shall be introduced to create an intelligent European news archive.

3.1 OmniPaper Goals and Objectives

The key objective of the OmniPaper project is the creation of a *multilingual navigation and linking layer on top of distributed information resources in a self-learning environment.*

The project aims to:

- Find and test mechanisms for retrieving information from distributed sources in an efficient way.
- Find and test ways for creating a uniform access point to several distributed information sources.
- Make this access point as usable and user-friendly as possible.
- Lift widely distributed digital collections to a higher level.

OmniPaper is a 3-year project with partners from four European countries which will have two main end results:

- A reference guide (Blueprint) for intelligent, efficient and multilingual knowledge retrieval.
- A newspaper prototype that enables users (from the news professional to the occasional user) to have simultaneous and structured access to the articles of a large number of digital European newspapers. One access gate will enable its users to search and navigate through news subjects in their own language.

3.2 Research & Development Challenges

From the scientific point of view, the project lifts widely distributed digital collections to a higher level, by:

- Applying a common multilingual thesaurus superstructure to them;
- Linking them to each other;
- Enriching their quality and the navigational features through learning from user behavior.

By building a multilingual interface to distributed archives, the project will take into account the local aspects of cultural and scientific information provision. The linked keywords that form a navigation layer will be automatically (and depending from their context) translated in the different languages that exist in the various archives. That way, readers can look up news information without having to know anything about the language of each of the archives. Newspaper articles themselves will be in the original newspaper's language, but can be (semi-) automatically translated at the user's request .

By personalizing this common knowledge layer according to the user's desire search results will be even more adequate, which increases the economic prospects of OmniPaper's end results. The knowledge layer will first be profiled to different kinds of users, such as journalists, politicians and the general public. Each user can then make his/her interface even more relevant by (for example) eliminating keywords that are not interesting and adding links between keywords that were previously not included. By building this personalized "map" on top of different information sources, OmniPaper will go much further than any existing search engine.

OmniPaper is not a project about digitization of news, but about bringing digitized news originating from various sources together through a single access gate. Therefore, the project assumes that the source material is already available in a digital form, containing sophisticated meta-data and navigational information. The added value brought by the OmniPaper system resides in the intelligent, self-learning and navigational knowledge superstructure built on top of this already enriched material. By doing this, the project will highly contribute to technical developments in view of a "Semantic Web".

3.3 Information & Content Retrieval Challenges

In most European countries initiatives do exist (or are at least initiated) for newspaper article exchange on a larger scale. These initiatives all share some limitations:

1. They all use a very centralized approach, that is, newspaper-articles are sent in a more or less standard format for check-in in a central database system that resides at a service provider's site.

2. Most of these initiatives do not cross language or country boundaries.

The "news" however is Europeanizing, since an increasing amount of local political, social, economical and cultural events do have their impact on a European level. As a consequence, many high quality newspapers do print an increasing amount of articles borrowed from foreign newspapers, often even in a non-translated form in order to maintain the true nature of the article.

On the other hand, most newspaper publishers maintain their own electronic archive as a service for their journalists. A centralized service shared between several newspapers, leads inevitably to a physical (be it transformed) duplicate of the existing archive, having as a direct result all sorts of maintenance problems. Further, journalists who are writing on a certain topic are now mostly limited to what their archive offers them. If they can access articles from other newspapers in an easy and efficient way, their understanding of this topic will gain both in width and depth.

As for now, cross-national distributed approaches do not exist. OmniPaper will cross both boundaries, creating multilingual access to the news-content of various European newspapers. Thus, OmniPaper will develop a common set of archiving and indexing tools that can serve as a future standard in this field.

4 Technical Architecture and Intelligent Services

Set up as a "smart information retrieval" project, OmniPaper is structured from a organizational point of view into several work packages. The work package structure reflects the necessary steps for making this smart retrieval possible. As depicted in figure 1, the architecture used in the project makes a distinction between a local ("Distributed Information Retrieval") and an overall layer ("Overall Knowledge Layer"). The three work packages included in Figure 1 (WP2, WP3 and WP5) will each pursue one of the three technological objectives.

On the level of the local layer ways for retrieving information from distributed sources are analyzed and thoroughly tested. When combining different archives into one large information pool, access to them must be possible in a uniform way. Techniques are currently established that allow this uniform access to differently structured archives. The overall knowledge layer will bring the local layers together in a well-structured manner. It will make cross-archive navigation and linking possible in a multilingual environment, resulting in a multi-archive knowledge layer. On top of the overall layer the user interface will provide a user-friendly and interactive presentation of the knowledge layer.

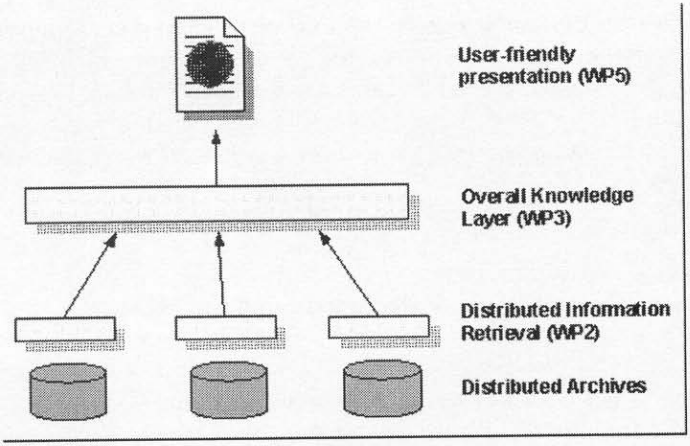

Fig. 1. *System Architecture. This fig. shows the bottom-up approach that will be used in the project. The local layer is constructed by building representative access modules for each type of the distributed archives.*

4.1 Local Layer - Distributed Information Retrieval

In order to access the distributed information archives in a well structured and uniquely defined manner, the local layer have to provide standardized interfaces both to the distributed archives and to the overall knowledge layer. Considering an embedded "OmniPaper system", the interfacing to internal layers is more an organizational challenge, thus defining features and activities which should be supported by interfaces to method calls and date exchanges. A technical challenge is the interface to the distributed archives, since these systems are considered legacy applications to the OmniPaper system.

The main reasons to focus on modern standardized interfacing technologies like SOAP are the platform-independent software availability and the weak influence that the OmniPaper system can have to existing news archives. Only standardized retrieval methods can provide a unique access to multiple heterogeneous archives and thus make use of the rich content provided within this digital libraries.

The main task for the local layer modules is to provide standardized access to the information which is available in the distributed archives. This is either achieved by transferring the access queries in the standardized format to the corresponding archive or reformulating the query for the existing legacy interface of the archive. The former variant can only be used, if the archive is able to provide the standardized interface required by the OmniPaper system.

The interface is currently defined by SOAP queries, containing data structures to retrieve newspaper information from distributed archives. The queries focus

on article selection by search criteria and keyword extraction. High level content organization and semantic coupling of information is handled by meta data services in the overall knowledge layer. AI methodologies like multilingual searches and user behavior evaluation for query improvements are managed at higher levels, yet the relevant queries are transformed to the above-mentioned simple SOAP accesses to the distributed archives.

4.2 Overall Knowledge Layer - AI, Multilinguality, and Knowledge Management

The results from the local layer queries will constitute the input for research on the overall layer. On both levels, new techniques are analyzed and compared. The resulting prototypes are planned to be rather small. Cross-testing of the prototypes allows conclusions on both levels, even the integration of different platforms for the software at the local layer by utilizing SOAP as the standardized query interface. These conclusions will be summarized in the "Blueprint for smart distributed information retrieval". Prerequisite of the prototypes is that the local archives do already exist in a digital form. This means that OmniPaper is not a project about digitization of news, but about bringing digitized news originating from various sources (and in various formats) together.

The overall knowledge layer combines the features of integrating distributed information with the capability of creating semantic coupling of the corresponding content. Meta data techniques are evaluated to act as a knowledge management platform in order to improve content research and multilingual information retrieval. The multilingual aspect is supported by extracting existing keywords and meta data from the heterogeneous archive information and associate it with existing domain specific thesauri for the relevant language. The overall knowledge layer contains a network of thesauri, thus coupling corresponding standardized terms and enabling the intelligent news archive to find corresponding articles in news archives over different countries and languages. This allows journalists and researchers to investigate material on specific topics in an multilingual environment, relying on high result quality and content relevance.

4.3 Usage Layer - Presentation Interface

A user-friendly presentation of the system, based on current HCI understandings [1], will be set up above the overall knowledge layer. Efforts in the corresponding work package concentrate on the news search engine, the display of the overall knowledge layer, and of newspaper articles and cross-links.

Obeying current technological developments, the user interface is planned to be developed on modern web service and web browser technology [4]. Nevertheless,

intelligent features from the overall knowledge layer need to be presented to the user in order to excel current information retrieval approaches.

- Special retrieval interfaces are developed to utilize the multilingual search results in order to provide the user with multinational contents or enable query precision through narrowing the search space in an interactive manner.
- User behavior is captured by the interface in a personalized way in order to provide relevant information to specific groups of users. The meta information collected and handled in the overall knowledge layer provide multidimensional filtering of the information. Typical characteristics within these dimensions can be profiled for the user groups and utilized for supplying relevant information to specific users.
- Modern agent technology is used to defined typical archive usage patterns. OmniPaper users are for example able to set *future queries*, i.e. profiles of typical news search attributes that can either be queried on the existing multilingual data set or postponed to detect and report future occurrences of news that match the stored patterns. Thus *future queries* keep the user informed, even if the event or information has not been published yet.

5 Conclusion and Future Work

Through the construction of a blueprint on intelligent access to distributed digital collections, OmniPaper will contribute to the long term development of cultural and scientific content. Focus will go to the advanced networking of archival resources in a self-learning environment. This way, digital collections will be made dynamic, interoperable and thematically contextualized. Thus the project contributes to a secondary action line, the "IST2001-II.1.2. Knowledge Management" of Key Action II (New Methods of Work and Electronic Commerce).

The prototype will provide access to personalized and context-specific content, and will organize heterogeneous information sources using ontology and semantic cross-lingual search. Since a self-learning knowledge layer will be built, the project's end results will dynamically provide relevant knowledge to the reader. The self-learning aspect of the prototype system includes the analysis of user behavior and the enriching of the knowledge layer with lessons learned from this behavior. That way, people can share knowledge without even realizing it.

Prototypes of distributed information retrieval and automatic metadata generation will be available for testing at the end of 2002. The project plan is to cross-test several approaches and elect the most promising techniques in order to build a framework and a blueprint as a guideline for future intelligent news archives.

References

1. Thomas Grechenig, Manfred Tscheligi: Human Computer Interaction, Vienna Conference, VCHCI '93, Fin de Siècle, Vienna, Austria, September 20-22, 1993, Proceedings. Springer 1993
2. Louis Rosenfeld, Peter Morville, Information Architecture for the World Wide Web, O'Reilly & Associates, Feb 1998.
3. Norbert Fuhr, Towards data abstraction in networked information retrieval systems. Information Processing and Management, 35(2):101-119, 1999.
4. Markus W. Schranz, Johannes Weidl, Karl M. Göschka, Stefan Zechmeister: Engineering Complex World Wide Web Services with JESSICA and UML. HICSS 2000.
5. Alexander Linden, Jim Jacobs. How to Swim, not Sink, in the Information-Flood. Gartner Research Online, research note, COM-13-4082, `http://www.gartner.com/resources/98300/98359/98359.pdf`, May 2001.
6. Henrik Nottelmann, Norbert Fuhr, MIND: An architecture for multimedia information retrieval in federated digital libraries, *Proceedings of the DELOS-Workshop on Interoperability in Digital Libraries.* DELOS-Network of Excellence on Digital Libraries., 2001
7. Steve Pepper, Graham Moore. XML Topic Maps (XTM) 1.0, `http://www.topicmaps.org/xtm/1.0/`, August 2001.
8. Karl M. Göschka, Markus W. Schranz: Client and Legacy Integration in Object-Oriented Web Engineering. IEEE MultiMedia 8(1): 32-41 (2001)
9. Martin Gudgin, Marc Hadley, Jean-Jacques Moreau, Henrik F. Nielsen, SOAP Version 1.2, W3C Working Draft, `http://www.w3.com/2002/ws` Dec 2001.
10. Henrik Nottelmann, Norbert Fuhr, Resource selection framework and methods, Deliverable D 3.1, EU project MIND (IST 2000-26061), `http://ls6-www.cs.uni-dortmund.de/ir/projects/mind/d31.pdf`, Feb 25, 2002.
11. OmniPaper - Smart Access to European Newspapers, EU project IST 2001-32174, `http://www.omnipaper.org/`, Jan 2002.