OmniPaper: Towards a Universal Standard Model for Efficient Information Retrieval

Markus Schranz

Distributed Systems Group, Institute of Information Systems,
Technical University of Vienna
Argentinierstrasse 8/184-1, Vienna, Austria
schranz@infosys.tuwien.ac.at

Abstract

Modern Information Society is overwhelmed with steadily raising quantities of information in a continuously integrating digital network involving standard PC, networked mobile devices and individual personal assistants, promising to us the optimum access to all the data we need. On the downside the uncontrolled information amount creates an enormous overload and costs enterprises and individuals money, often in ways that are not easily measured: Costs that result from lowered productivity and from mislead business decisions. To really satisfy user needs and restricted budgets, the myriads of information need to be structured and organized in an intelligent and user-oriented way. Multiple approaches have been followed to integrate heterogeneous information sources and promising research results have been achieved in particular application domains. This paper discusses the area of online news integration by modern service architectures, Web service technologies and the use of artificial intelligence to semantically relate news within an intelligent news retrieval interface engine. OmniPaper has created a multilingual navigation and linking layer on top of distributed information resources to provide a sophisticated way of managing multinational news archives with strong semantic coupling. The research results have been documented in detail in a voluminous project deliverable and the most important findings are outlined in this paper.

1 Introduction

Recent high-tech developments and ongoing research in the area of distributed information retrieval, document management and knowledge management offer us both feasible hardware equipment and intelligent software tools to handle large amounts of information. Yet the identification of the most relevant and appropriate data in a specific field of interest is a task machines can not easily provide for humans. Information often is available from different independent sources and a semantic integration involves several challenging tasks, both of quantitative and qualitative character. Typically one must cope with mutlilinguality, synonym identification, linguistic reasoning, relevance ranking, keyword handling and the identification of primary sources in case of multiple copies.

The OmniPaper project (Bueno, 2002) concentrated on the application field of online news publishing. The multinational project aimed towards the creation of an intelligent multilingual navigation and linking service on top of distributed information resources, thus providing a sophisticated approach to manage multinational news archives with strong semantic coupling. Technically, appropriate architectures to integrate existing archives with an intelligent news retrieval engine have been developed and ongoing research has investigated ways for drastically enhancing access to many different types of distributed information resources. To satisfy an information need, generally 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 semantically joined (Nottelmann 2001). Up to now, these are inefficient manual tasks for which accurate tools are desirable.

Promising research activities in the area of digital libraries (Fuhr 1999) provide end-to-end solutions for federated digital libraries, which cover most of the problematic issues. Information retrieval techniques, retrieval quality and the integration of non-cooperating libraries are the research focus. Especially in digital news archives, the integration of various existing non-standardized services is a demanding challenge to both information architects (Rosenfeld 2002) and system engineers (Schranz 2000).

The herein discussed approach was oriented towards the establishment of modern interaction and interconnection between existing news archives, thus lifting widely distributed digital collections to a higher level, by applying a common multilingual thesaurus superstructure (Kramer 1997) to them, linking them to each other, and enriching their quality and the navigational features through modern user interface management.

This paper presents an overview of the research results and methodologies used in the OmniPaper project. State-of-the art technologies (such as SOAP, RDF and Topic Maps) have been examined, compared and prototyped in

order to find the best ways for creating flexible navigation, filtering of information, cross-lingual and cross-archive information retrieval with most respect to existing research results in the news retrieval area (Mantzaris 2000). Artificial intelligence concepts have been incorporated to automate the creation and maintenance of a powerful knowledge layer in a user-oriented presentation environment. Automatic keyword extraction provides a uniform relevance ranking mechanism across the different searched archives.
We focus on the results of architectural research for digital libraries and the integration of multilingual news archives into an intelligent international news search and retrieval engine for the WWW. The paper explains the research work conducted and is structured as follows. Section 2 defines the basic concepts, requirements and related research areas to the OmniPaper project. Section 3 focuses on the architectural design and the implementation and provides an overview on system components and news archive processes. Section 4 outlines the applicability of research findings and technological results to further application domains and explains the use of AI and usability results for future information retrieval projects. The conclusion sums the project experiences and the use of the project documentation as a universal standard model for distributed information retrieval.

2 Approaches to Information Retrieval

OmniPaper is not a project about digitisation of news, but about bringing digitised 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, multilingual and navigable knowledge superstructure built on top of this already enriched material.
As an inherent part of its conception is the networking of distributed computers, a logical consequence is that a lot of information on the Internet is physically distributed. The distribution itself is not necessarily the problem, since it is the Internet and modern Internet technologies that makes information available everywhere even it is stored on computers far away. The problem is that it becomes hard to find out what information is out there and where to get it (on what addresses). Search engines solve a great deal of this problem and they are as such becoming more and more popular, not only for finding information for which users don't know where to look, but also for getting to information that users have already consulted before—so for navigation instead of pure search.
Even with the help of search engines the main problem still exists: information is spread across the Internet and this makes it very hard to get a complete view on all available information about one topic or to make connections between parts of distributed information.
Information is not only scattered physically, but also in terms of storage formats, environments, hardware, database formats, information structures, etc. This problem adds up to the problem described before: even if the data source is known, this doesn't mean that real information can be extracted from it easily. Heterogeneous environments and information formats make efficient information capturing more difficult.
Additionally, high efforts in the Semantic Web area have been spent in defining metadata standards: standard ways to describe metadata. The idea is that information will get much better accessible if it is well-described. For general metadata Dublin Core is one of the most important standards; for news articles NewsML is widely adopted.
However metadata initiatives in the Semantic Web are becoming more and more important, there still is a lack of good quality metadata. Now the phenomenon can be practically observed, that everybody seams convinced of the advantages of metadata, but nobody really bothers for creating and maintaining them. Authors often see the creation of metadata as a burden, even in the area of news publishing. Journalists and editors are under time pressure all the time for publishing news as soon as possible. As a result creating good quality metadata for their news articles often is ruled out by other priorities.
These observations mean that, if an important tool for improving information retrieval is metadata, this metadata has to be gathered using other means than mere manual creation.

2.1 Approaches for Distributed Information Retrieval and Publishing

Conceptually, distributed information is retrieved by web robots or centralized databases, harvesting as much information as possible for high performance access on distributed heterogeneous databases. Web robots follow the regular "crawling" of the web: systems that scan an entire data collection and that make giant indexes of all found information. Most popular search engines use this method because it allows users to search very fast in huge amounts of information—if the indexes are efficiently constructed. A key feature of this method is that it "pulls" information from all directions into a centralised index.

The centralized database approach is completely opposite from the previous one as it requires complete co-operation of the data sources: they regularly send updates to the central database or index on their own initiative ("push"). An advantage is that the quality of information in the central index can be much better because it stems directly from the original source. Disadvantages of this method are the difficult maintainability of such a central data store, possibility of outdated information and possible problems in information exchange if different data structures are being used.
The OmniPaper service combined features of both information retrieval approaches. The designed architecture proposed the following information retrieval activities and were (partially) implemented in a runnig prototype (see also section 3):

1. The archives co-operate to the centralised search system.
2. All information exchange between the central system and the distributed archives happens through the Simple Object Access Protocol SOAP (Gudgin 2003). This means that it is well-structured and uses XML syntax.
3. Manually created metadata is enriched with automatically extracted keywords.
4. Central database with only metadata (including extracted keywords). News article contents are never stored centrally but reside at their original location at the information provider.
5. Push combined with pull: normally the distributed archives notify the central system if information updates are available, but the central system can also proactively ask the local archives to give a status update. Either ways should make sure that the central system has the metadata of all articles, including the most recent ones.
6. Query and navigation are both methods for searching and should be combined as much as possible. Navigation can provide support to querying and vice versa.

Related Research and Practical Application Areas

The OmniPaper service contains technical components and research approaches from multiple application areas within the field of electronic publishing and information retrieval. Multiple information retrieval domains such as general distributed information retrieval, online news publishing, modern search engines, and even ontology-driven systems, handling categorization of topics to qualitatively enhance the information structuring and retrieval management, are providing services and applications that offer features comparable and/or complementary to OmniPaper's services.

3 Prototype Design and Development

The major technological objective of the OmniPaper project was to create an intelligent uniform entrance gate to a large number of European digital newspapers, allowing readers a more objective view on subjects. This rather general technological objective were split up in three parts which are easier to verify and measure:

1. Find and test mechanisms for retrieving information from distributed sources in an efficient way.
2. Find and test ways for creating a uniform access point to several distributed information sources.
3. Make this access point as usable and user-friendly as possible.

The OmniPaper project created a reference document and a prototype software system for improving access to distributed information resources. The latter acts as a uniform, multilingual access system to articles from various European newspapers. This system enables users to search a newspaper article in one language, returning multilingual results originating from different important newspapers.

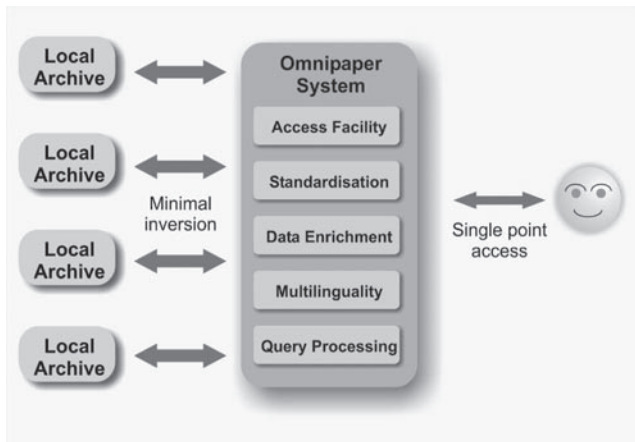

Figure 1: Abstract Architecuture of OmniPaper

The overall system architecture provides an integrative description of the components designed and developed for the creation and practical use of the OmniPaper system prototype. The OmniPaper architecture starts from distributed news archives, all within different operating environments, database formats and indexing mechanisms. Heterogeneity, performance and usability are challenges to the responsible system architects. In a standardization effort, SOAP has been selected to create a uniform access method to the existing archives. In addition to the simple access requirements, the intelligent news archive is required to extract specific contents and create relations between information units. Rich indexing and meta-data structures, such as Topic Maps (Pepper 2001) and RDF are utilized to make intelligent search possible. A cross-archive intelligent index (or 'knowledge layer') contains concepts, relationships between them and occurrences in different languages.
Figure 1 shows an abstract architecture of the OmniPaper system. Distributed source data exists in various standard formats, different languages and with varying depth of available information. The existing local archives are connected to the OmniPaper system using minimal inversion i.e. existing access methods are reused and wrapped to standardized SOAP requests.
From the user interface, the architecture provides access via NL Queries or browsing the content by categories in the user's own language. The provided results are retrieved from all local archives.

3.1 Architectural Design

In order to meet the goals and fulfill the requirements of the multilingual news archive, research on and evaluation of related international projects (Nottelmann 2001), retrieval methodologies and semantic relation approaches has been applied in the field of digital libraries and news archives. A top-level architecture describes an appropriate grouping of system components and gives an overview on two major different usage interfaces.
The top-level system architecture contains a multi-layer view on the technical architecture of a distributed news archive. The architecture is based on existing digital news archives as the bottom layer. The distributed information retrieval layer contains components and control processes that access directly or indirectly the existing archives for news retrieval and metadata management. The overall knowledge layer combines the features of integrating distributed information with the capability of creating semantic coupling of the corresponding content. The multilingual aspect is supported by extracting existing keywords and metadata from the heterogeneous archive information and associating it with existing domain specific thesauri for the relevant language. The overall knowledge layer contains a network of thesauri based on EuroWordNet (Vossen 2001), thus coupling corresponding standardized terms and enabling the intelligent news archive to find corresponding articles in news archives over different countries and languages. Based on these layers, the topmost user interface layer allows journalists and researchers to investigate material on specific topics in a multilingual environment, relying on high result quality and content relevance.
The top-level system architecture from Figure 1 is constructed mainly from the point of view of a news provider for the end-user. Existing archives and news repositories act as data sources for the distributed news archive and provide their contents via standardized SOAP interfaces. The system itself offers simply a data feed interface, acting as black box for the existing archives. To provide openness and facilitate easy archive extension the used set of SOAP queries is limited to seven requests. Detailed definitions of the SOAP requests and relevant parameters are defined with the OmniPaper project documentation and are outside the scope of this paper.
From the user's perspective the distributed news archive offers a completely different interface. Either digital users (applications) or human users may access the archive using natural language queries, set of terms or browse through a web of related terms (concepts). The user interface layer translates the user queries into a manageable format for the distributed news archive.
The user interface is designed in the system architecture as a user-centered Web application. Digital users both can use simple access to Web forms and XML Web Services.

3.2 Prototype Implementation

Within OmniPaper, a standard format for the news contents has been defined. These formats are used by all prototypes. The content of the system is actually the metadata of the articles; it is written in XML, which helps its interaction with standards like RDF and/or XTM. The connection with the content providers is via SOAP. Requests are used to retrieve documents for processing on uploading and to show the contents to the user on downloading, once the system has determined, based on the abovementioned metadata, that a news article fulfils a user's request.
The standard format for the metadata has been defined following widely accepted standards, in particular the Dublin Core Metadata Element Set (DCMES), and the News Industry Text Format (NITF), but also NewsML. The format describes twenty-three basic elements, grouped under these categories: Identification, Ownership, Location, Relevance, Classification, and LinkInfo.

During the service development the consortium has built several different prototypes to prove individual concepts and to finally integrate the bottom up developments to an intelligent news retrieval service. The prototypes document the progress in the implementation of the architectural work and the consecutive steps to achieve a highly functional distributed information retrieval service.
In the OmniPaper project a bottom-up approach was used for prototype development. Several smaller prototypes have been developed in order to cross-test different technologies and to limit the project development risks. These smaller prototypes have been combined into the Distributed Information Retrieval Prototype.

Distibuted Information Retrieval Prototype

The goal of this prototype is to provide a smart search and navigation layer on top of one news archive. This "Local Knowledge Layer" allows semantic-based search, navigation and filtering of English newspaper articles. This prototype is the combined end result of the earlier project work in the prototypes using SOAP, XTM, RDF and automatic keyword extraction. The task of the prototype is to enhance the user experience in finding online news of interest. The prototype obtains its data from online news sources managed by a number of news providers and tries to build an intelligent top layer on the data that consists of metadata and an intelligent search interface. In order to achieve this goal the available metadata must be stored in a structured way and a number of dedicated query and navigation mechanisms to access this data must be designed.

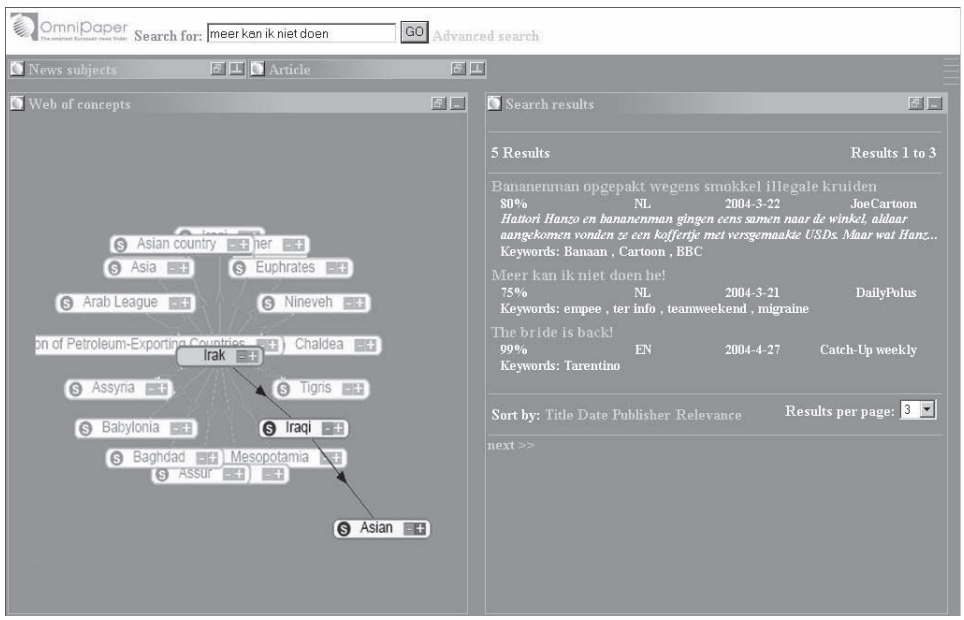

Figure 2: OmniPaper final prototype retrieval result screenshot

In this prototype, querying and navigation are considered as alternative methods to find relevant information. Both interact with each other and together they produce a combined user experience that can be expressed as "find what you were looking for and then browse away from it". The prototype considers both querying and navigation as a kind of search action. The only difference is that in navigation the user follows predefined paths, whereas in querying the user is free in what he or she submits as a query.
This prototype implements four kinds of query and navigation. A first method allows users to navigate through news subjects (categories) in a traditional, hierarchical way. A more sophisticated tool is the relational navigation, where users can browse through a "web of concepts". The starting point for relational navigation (the "focus concept" in OmniPaper terminology) is the result of the last navigation or query action and the predefined paths that can be followed are paths to concepts that are related to the focus concept in the knowledge map. Finally smart querying is enabled using a "knowledge map" of semantically related keywords and concepts.

4 Modern Approaches in Distributed Information Retrieval

Besides the technological development in the OmniPaper project, the research experiences reach far beyond an implementation model for distributed information retrieval. Our findings in the area of semantic information coupling and end-user integration advance the documentation of the project to a universal standard model for distributed information retrieval.

Automated Content Relation Management

Among the available functionalities provided by the OmniPaper service there is an on-line indexing function based on vector space technology, devoted to the reordering of the articles retrieved for a user search according to the likelihood of satisfying the query stated by the user. As shown in Figure 3 the retrieval results contain ranking numbers (percentage measures) to compare the matching quality of the displayed news articles. Modern multilingual vector space models based on automated keyword extraction support the creation of a multilingual neutral ranking mechanism that provides the users with the most appropriate search results for their OmniPaper queries. Details on the utilization of AI methodologies are beyond the scope of this paper and are presented in detail in the projects blueprint documentations, publicly available at the European Commission.

Usability Findings

The prototype described in the design and development section was created according to the guidelines defined in the usability research for the OmniPaper project. We have identified user interface guidelines for integrating the Omnipaper multilingual search into (general) web interfaces. These guidelines help developers to design effective search interfaces that are useful and easy to use. The blueprint documentation details different preconditions and requirements when integrating the search interface into existing or planned systems and user interfaces with different archives and design constraints.

5 Conclusion

Finding accurate but widely dispersed information is highly important for newspapers, whose success strongly depends on their speed of providing news. By building a multilingual interface to distributed archives, the OmniPaper project's approach allows to take into account the local aspects of cultural and scientific information provision. Queries are automatically 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.

Developed for very large-scale distributed collections, the described service is targeted to serve systems that improve access to cultural and scientific knowledge sources. Access to digital news services will not only be improved quantitatively by combining a large number of digital newspapers in one system. The architecture also improves the quality of access by supporting the building of personalized, cross-lingual and self-learning interfaces to the distributed collections.

The paper summarizes the research work and the project results the consortium has performed and achieved during the three years of project duration. The purpose of the paper is to outline major findings that have lead to guidelines for efficient information retrieval in a distributed and heterogeneous environment. Using this paper, future research and practical efforts by subsequent organizations and projects shall be started from a well researched and clearly described position that reflects the consortiums experiences, major findings and developments and thoroughly researched component evaluations towards a standard architecture and process management for similar projects.

Acknowledgements

This work was partially funded by the EU 5th Framework project OmniPaper (IST-2001-32174).

Referentes

Bueno, F. et. al. (2002), OmniPaper – Smart Access to European Newspapers, EU project IST 2001-32174, http://www.omnipaper.org/, Jan 2002.

Fuhr, N. (1999), Towards data abstraction in networked information retrieval systems. Information Processing and Management, 35(2), p. 101-119.

Gudgin M. et. al. (2003), SOAP Version 1.2 Part 1: Messaging Framework, http://www.w3.org/TR/SOAP/, 24 April 2005

Kramer R. et al. (1997) Thesaurus federations : loosely integrated thesauri for document retrieval in networks based on Internet technologies*, International Journal on Digital Libraries* 1(2) pp 122-131, June 1997.

Mantzaris S.L. et al. (2000), Integrated search tools for newspaper digital libraries, *in proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR 2000*, July 24-28 Athens, Greece, 2000.

Nottelmann, H. and Fuhr, N. (2001), 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.*

Pepper S. and Moore G. (2001), XML Topic Maps (XTM) 1.0, http://www.topicmaps.org/xtm/1.0/, April 2005.

Rosenfeld, L. and Morville, P. (2002), Information Architecture for the World Wide Web, O'Reilly & Associates.

Schranz, M. et. al. (2000), Engineering Complex World Wide Web Services with JESSICA and UML. *In proceedings (ISBN 0-7695-0493-0) of the 'Hawaii International Conference On System Sciences HICSS-33'*, Maui, Hawaii, USA, Jan 4-7, Jan 2000, p. 167.

Vossen, P., EuroWordNet, (2001), EU-funded project, LE-4 8328, http://www.illc.uva.nl/EuroWordNet/ April 2005.