Automating Information Aggregation Processes in Information Centres

Markus Bauer[a], Christoph Herzog[a],
Hannes Werthner[b], Birgit Dippelreiter[b]
and Kathrin Prantner[c]

[a] E-Commerce Competence Center EC3, Austria
{markus.bauer, christoph.herzog}@ec3.at

[b] E-Commerce Group,
Vienna University of Technology, Austria
{hannes.werthner, birgit.dippelreiter}@ec.tuwien.ac.at

[c] DERI Innsbruck, Austria
kathrin.prantner@deri.at

Abstract

As tourism is an information business, tourists have to rely on the information available beforehand. One – prominent – source of information are call centres of national or regional tourism organisations, which provide users expert advise and also pre-compiled information documents and brochures. However, such requests are usually answered with on demand off and on-line searches, communicating the required information to the customer via telephone. This is a cumbersome and often low quality process. The project OnTourism aims at improving these processes by automatically extracting (publicly) available information about the requested topics and compiling them into comprehensible documents which can then be sent to the customers. The information access for the call centre's personnel is supported by semantic annotation of the available documents, thus improving the required time to answer customers' requests as well as the service quality.

Keywords: applied research, web extraction, information aggregation, semantic annotation

1 Introduction

Tourism is an information centred business, tourism goods are mainly confidence goods (Werthner 2003). That means that it is usually impossible for the customer to judge the quality of a tourism offer (e.g., a music event or a hotel room) beforehand.

Therefore, customers are heavily dependent on the available information to make their decisions.

One major source of information for tourists are the call centres of regional or national tourist organisations, such as the *Urlaubscenter* of Österreich Werbung. The call centre agents not only provide expert information for the customer, but also send them available brochures or information documents on request. For the latter, they have defined a process, where for major and anticipated user preferences related content objects are assembled, which are then compiled into respective documents. However, since the effort to create these documents is relatively high, only major preferences and related topics are considered.

For even slightly more individual information needs, online searches are conducted in order to provide the customer with the required information. These online searches offer opportunities for improving the call centre's process and the quality of the information provided to the customer. This can be achieved by pre-compiling a larger number of more specific documents. However, the time consumption for manually assembling documents even for only the most important categories of customer requests would be unfeasible.

In the OnTourism project we improve the information aggregation process of Österreich Werbung's call centre by categorising the customer requests which have to be answered through on-demand online searches and analysing which information sources are queried in order to provide the caller with the required information. These processes are then automated, using advanced web extraction techniques to aggregate the information from the identified web sources. These data is further compiled into comprehensible PDF documents, which can be sent to the customer by email. Information about upcoming events (e.g., cultural or sports events) has been selected as use case for our approach.

Furthermore, the call centre agents need to be able to quickly access the compiled information documents. For storing pre-researched information, they have a document repository available, which supports a semantic search functionality for more precise information retrieval. The automatically generated documents are stored within that document repository and are also automatically annotated with a description of the documents' contents in terms of a formal ontology. This enables a more precise search result with better recall through semantic reasoning.

While our solution is based on the specific needs of the call centre scenario, it comprises a toolkit for extracting information from web sources, integrating them in an intelligent manner into comprehensive documents for human consumption and for describing the contents of these documents in terms of a formal ontology, enabling advanced business processes to be triggered and controlled. In the remainder of the paper we focus on the call centre application as a case study or proof of concept of our approach.

The paper is organised as follows: Section 2 gives an overview of the information extraction process and tools. Section 3 details the process of creating comprehensive human-readable documents from the aggregated information and section 4 describes how these documents are semantically annotated and stored in the call centre's document repository. Section 5 details some of the obstacles we faced during the implementation of our solution and section 6 gives examples for further application scenarios of the underlying technological infrastructure.

2 Automatic information extraction

In order to create valuable documents, we have to extract information from a multitude of different websites, producing semi-structured information as the basis for further processing. The extraction engine uses listings of events from all Austrian provinces and extracts additional interesting information concerning these events (e.g., weather forecast, travel information, etc.). Automatic document generation out of aggregated up-to-date information of upcoming events assists the call centre agents in answering specific customer requests in a highly efficient manner, which will be described in the following.

2.1 Document sources

In the call centre, the most common requests are answered by sending manually prepared information documents to the customer. Answering questions regarding up-to-date information, however, was previously only possible with on-demand searches on different websites. In order to speed up this process, requested information types (e.g. events) are assessed, analysing which ones have a high demand for up-to-date information with previously little or no coverage in the knowledge base.

Source selection. This assessment is based on data about the tourist requests collected at Österreich Werbung's call centre. These data were collected by the call centre agents, summarizing the requests over the period of a year. The survey result shows that the major part of requests came from German speaking countries and was concerned with the topics of city/culture as well as with upcoming events. The topic of events, however, was previously not covered with up-to-date information, mainly due to the fact that the call centre personnel couldn't afford the time needed to assemble documents about a large number of events each week.

In order to generate additional value by automatically aggregating up-to-date information, the essential sources were found from which the necessary information can be extracted (i.e., the challenge was to identify structured and stable websites with enough information for an automatic extraction).

OnTourism uses listings of events from the web pages of each of the Austrian provinces as a basis for further extraction. Then for each event additional interesting information from other sources is extracted. The integrated document sources are:

- event calendar of each province – these deliver important event information
- location information – deals with the location of the event and a short description, including pictures
- arrival and mobility – shows how to arrive by car and other information about mobility in the region
- connection to public transport – details the arrival by train or other public transport means
- current weather – shows the weather forecast for the next week
- obstacles on the road – advise about possible obstacles on arrival (blocked roads)
- press releases – offer current press reports about the event

2.2 Information extraction

Information extraction is "the automatic identification of selected types of entities, relations, or events in free text" (Grisham, 2003), an emerging technology whose function is to process natural language text in order to locate specific pieces of information, or facts, within the text. These facts are then further transformed into structured or semi-structured representation formats. The aim of a so-called wrapper program is to locate relevant information in well and clearly structured websites and to put the information into a self-describing representation format for further processing. Hence wrappers should offer the following fundamental qualities (Gilleron, 2005):

- **expressiveness** to cover various data formats and heterogeneous sources
- **efficiency** of the generated wrappers and of the generation process
- **robustness** in missing values and changes in the layout
- **maintainability** to support the generated wrapper a long time

Toolkits for generation wrappers can be differentiated in a number of ways. They can be categorised by their output methods, web crawling capability, interface type, use of a graphical user interface (GUI) and several other characteristics. Laender et al. categorise a number of toolkits based on the methods used for generation wrappers (Laender, Ribeiro-Neto, Silva & Teixeira, 2002). These methods include specially designed wrapper development languages and algorithms based on HTML awareness, induction modelling, ontology and natural language processing. For the OnTourism project a wrapper has to fulfil some necessary conditions:

- A data integration module for generated wrappers is essential to allow the wrapper designers to integrate external web data sources and extract the information with web crawling support.

- An important feature is the possible output format in which the extracted data can be exported. XML allows for the most flexible handling and easy processing of the data into comprehensive, human readable PDF documents.

As a state of the art product, the tools should support a GUI and an editor for visually designing the wrapper programs and for setting up and executing the information extraction as well as the aggregation process.

The *Lixto Visual Developer* fulfils these requirements as a state-of-the art tool for supervised wrapper generation and automated web information extraction. Furthermore, the *Lixto Transformation Server* provides a platform for the subsequent processing of the aggregated information. Therefore, the Lixto Suite[1] has been selected as a means for implementing the outlined call centre application.

In the first step of a wrapper design process, extraction rules have to be semi-automatically and visually defined in an iterative process to build extraction patterns (Baumgartner, Fröhlich & Gottlob 2007). Figure 1 depicts this process. Information extraction systems rely on a set of extraction patterns that they use in order to retrieve the relevant information from each source. The distinctive features of Lixto are mainly (Gottlob et Al. 2004):

- **High productivity**: Lixto includes a fully visual wrapper specification process to allow for a steep learning curve and high productivity in creating wrappers.

- **Expressivity and Stability**: The Lixto extraction engine is based on an internal logic-based language similar to Datalog called *Elog.* This language allows for the extraction of target patterns based on surrounding landmarks, on the content itself,

[1] see *http://www.lixto.com/*

on HTML attributes, on the order of appearance and on semantic and syntactic concepts. Elog in its core fragment captures precisely the expressiveness of monadic second-order logic (MSO) over trees (Gottlob & Koch 2004) and has been proposed as a yardstick for evaluating and comparing wrappers.

Each document source is covered by a single wrapper program. XSLT composition files, managed by an admin interface, are used to aggregate the information produced by these wrappers into a unified XML document. The admin interface is an application that modifies the global XSLT composition by changing its defined structure. This way the information available for further processing can be controlled.

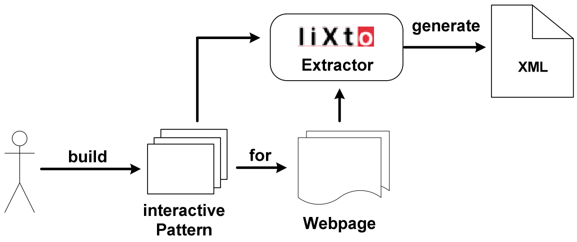

Fig. 1. Data Extraction

3 Document generation and semantic description

Following the data extraction step the XML data generated by wrappers are processed within the transformation server, which schedules the extraction process on one hand and transforms the generated data into the required output format on the other one.

3.1 Information transformation

The desired output of the information aggregation process is an easy to read and information rich PDF document, which the call centre agents can email to the customers on request. The aggregated information, available after the extraction process, needs to be compiled into a human readable and well structured document. Furthermore, in order to ensure that the documents can be easily found by the call centre personnel, a semantic annotation in the terms of a formal ontology is added. The ontology serves as a formal conceptualization of the tourism domain, allowing for semantic reasoning and formulating precise search queries. The OnTourism ontology is expressed in OWL[2], an RDF[3] based ontology language.

Figure 2 shows the whole workflow of building a PDF file and the semantic description represented as an OWL document from the extracted XML data. The system features an administrator interface for selecting the information sources which serve as basis for building the individual documents. This gives the call centre agent the possibility to modify the document content in order to adapt it to specific needs.

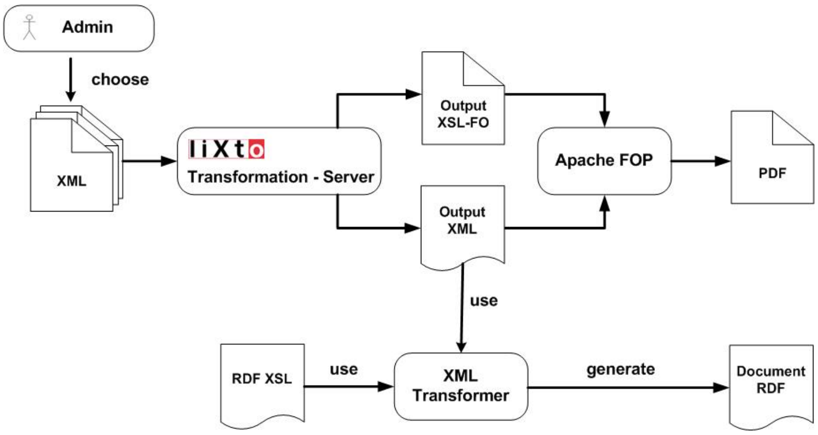

Fig. 2. Document Generation

Based on the selection of information sources, the transformation server merges the extracted information and compiles it into an output XSL-FO document and an output XML file. Both are subsequently needed to generate a PDF file and the OWL annotation from the extracted information.

Schedule. In order to capture up-to-date information, the extraction process is performed at least once a week. Regarding the time required for extracting and processing the information, as well as the rate at which new events are published at the monitored portals, a shorter extraction cycle does not seem to be need.

[2] Web Ontology Language, see *http://www.w3.org/2004/OWL/*
[3] Resource Definition Framework, see *http://www.w3.org/RDF/*

3.2 Semantic annotation

The OnTourism ontology aims to map from real-life tourism concepts and relations into an appropriate vocabulary, paying special attention to the needs of the environment of the call centre. An essential part of the ontology design process was the consideration of reusing existing ontologies in the tourism domain (Prantner, Ding, Luger, Yan & Herzog 2007). Since the semantic metadata annotation will be used mainly to enhance the search process in the call centre (Herzog, Luger & Herzog, 2007), the metadata structure is developed in close collaboration with the call centre agents and adopted to their special needs.

Ontology modelling. The respective OnTourism ontology is represented in OWL Lite, the least expressive and most efficient subclass of OWL languages (Zhanova & Keller 2005). However, this language already requires reasoning with equality, which significantly increases computational complexity. In OWL based ontologies, information is annotated by adding statements to an arbitrarily resource characterized by a unique reference. Instances can be derived from the basic classes to represent a concrete document or event and can be further related to class instances representing metadata such as a place or month. Such relations can be easily added to document, allowing for efficient handling of the metadata. Figure 3 shows a small part of the OnTourism ontology, which contains its major characteristics. The basic structure, founded by the classes "document" and "metadata", is connected using transitive relations and offers the possibility of adding information to each document in the repository. Therefore, new events are connected with other metadata, expanding the ontology.

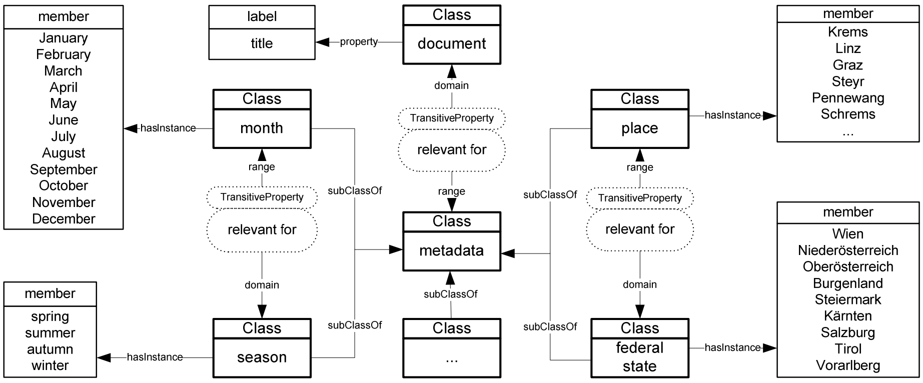

Fig. 3. Ontology Model

3.3 Document generation

A scheduled and automatically controlled application imports the XSL-FO[4] result document provided by the transformation process and uses it to generate the final PDF document. The implemented transformation server is able to activate automatically this application after completing the extraction and transformation processes. With XSL-FO, which is an XML based print page description language, a direct transformation from XML to PDF is possible. The extracted event information is converted into a XSL-FO representation. The topical content and the detailed design is governed by predefined rules regarding the available information items and their properties. Apache FOP, a print formatter driven by XSL formatting objects (Apache software foundation, 2007), is used to generate a PDF document from the XSL-FO specification.

4 Document storage

The generated PDF documents are automatically imported into the repository (a Microsoft SharePoint server). Figure 4 depicts this process. After an automatically generated document has been added to the repository, the likewise generated OWL document, containing the ontology representation of the newly added document's description, is sent to the repository's semantic subsystem and is further used for search and retrieval. Towards this end, a semantic search application allows the user to refine search queries by selecting objects from the underlying semantic data model (Herzog, Luger & Herzog, 2007). Since the semantic data model allows for semantic reasoning processes, this search method also finds documents related to but not directly annotated with the specified search terms (higher *recall*). For example, documents annotated with “Inntalkette” (a mountain range in Tyrolia) can also be found with the terms “Mountain” or “Tyrolia”.

[4] XSL-FO is part of the Extensible Stylesheet Language (XSL) specification, see *http://www.w3.org/TR/xsl/*

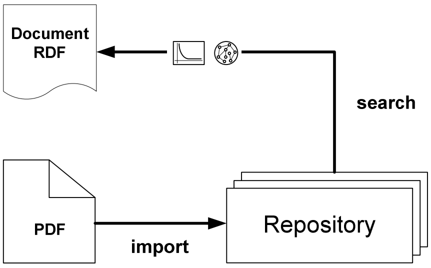

Fig. 4. Document Import & Search

5 Obstacles faced

The following key areas were identified in the course of the project.

Usability-Engineering. One of the most common causes for project failures is the lack of user integration during the design and implementation phase as stipulated in DIN EN ISO[5]. Hence, the members of the call centre where integrated in the design process to adopt the system to their special needs. This close cooperation ensured the design of a system which, while as a framework applicable to a larger set of usage scenarios, meets the demands of these users.

Ontology mapping. Creating a metadata description based on the OnTourism Ontology was a major challenge during the design process. Since there is no way to foresee all possible combinations of available metadata that can be extracted from the web sources, a framework was required in which mappings between the extracted metadata and the ontology terms have to be hypothesized dynamically as the terms are encountered after extraction. This mapping needs to be controlled and adapted when the extracted meta information is significantly modified (e.g., should a new categorization of events be introduced by a website listing events). The challenge of this mapping was closely related to the challenge of building a stable ontology structure which can accommodate the information required to describe the automatically created documents. A crucial step in this process was the incorporation of existing ontologies, enabling a wide coverage of the tourism domain (Prantner, Ding, Luger, Yan & Herzog 2007).

[5] Standard of “human-centred design processes for interactive systems”, see *http://www.iso.org/*

Information extraction. A further challenge faced appeared during the development phase when a significant structural changes of a web source required the adaptation of an existing wrapper program.. However, by using an extraction and transformation tool with a stable and expressive data language, like Lixto's Elog, wrapper programs can be designed to be able to cope with slight to medium changes in a source web site's structure.

6 Further application scenarios

Apart from the application in the call centre for pre-assembling information in order to provide better answers to customer requests, the described mechanism offers a generic solution for extracting and assembling data from web sources and turning them into PDF documents for human consumption. The automatically generated semantic description offers advanced means for automatic classification through semantic reasoning, or for triggering decision processes. We shortly outline two of the many possible scenarios for utilizing the project results

Tourist information. A possible refinement of the call centre scenario would be the automatic aggregation of information about cities' sights. With location information and current weather reports and forecasts, a tourist can plan ahead for his or her stay in a given city. Should the weather change the tourist can then request an alternative program for the current location on demand (e.g., foul weather program or outdoor activities).

Hotel Leaflets. Hotels often provide leaflets with tips for nearby sights or activities to their guests. Through automatically created documents, these could incorporate up-to-date information (e.g., including current events) and the suggestion can be based on customer preferences and weather conditions.

7 Conclusion

Tourists are highly dependent on information available beforehand to judge the attractiveness of a tourism offer. Call centers of national or regional tourism organizations provide expert advice and are a valuable source of information for the customers. Sending informative documents per email, the customers quickly get the required information for consumption at their own leisure. However, the number of up-to-date documents available to the call centre agents is very limited, since the effort of researching and compiling such documents makes the preparation of documents about information with a short life cycle (e.g., cultural, sports or music events) unfeasible.

In the OnTourism project we implemented a system for automatically aggregating up-to-date information about upcoming events and compiling them into informative and comprehensible PDF documents which can be sent on request to the customers. By pre-assembling the information from public web sources, we can improve the speed and quality of the process and of the answer the customer receives (i.e., a comprehensible and printable document). Furthermore, with semantic annotation of the documents' contents in terms of a formal ontology, the call centre's semantic search application can be used to make the document quickly accessible. The benefit of our solution is, however, not limited to the call centre scenario at hand, but can be used wherever publicly available information needs to be compiled into printable documents for human consumption in an intelligent manner, adding a semantic description that makes it possible to steer business processes.

References

Apache software foundation (2007). *Apache FOP.* Retrieved September 17, 2007, from http://xmlgraphics.apache.org/fop

Baumgartner, R., Frölich, O., Gottlob, G. (2007). The Lixto systems applications in Business Intelligence and Semantic Web. *4th European Semantic Web Conference (2007).*

Gilleron, R. (2005). *Machine Learning and Information Extraction*. Retrieved September 18, 2007, from http://www.grappa.univ-lille3.fr

Grisham, R. (2003). Information Extraction. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (545-559). Oxford: University Press

Gottlob, G., Koch, C. (2004). Monadic Datalog and the Expressive Power of Web Information Extraction. *Journal of the ACM 51* (74-113). New York (2004)

Gottlob, G., Koch, C., Baumgartner, R., Herzog, M., Flesca, S. (2004). The Lixto Data Extraction Project – Back and Forth between Theory and Practice. *PODS 2004.* Paris, France. (June 2004)

Herzog, C., Luger, M., Herzog, M. (2007). Combining Social and Semantic Metadata for Search in a Document Repository. *4th European Semantic Web Conference (2007).*

Laender, A., Ribeiro-Neto, B., Silva, A., & Teixeira, J. (2002). A Brief Survey of Web Data Extraction Tools. *Sigmod Record (Vol. 31)*

Prantner, K., Ding Y., Luger M., Yan, Z., Herzog, C. (2007). Tourism Ontology and Semantic Management System: State-of-the-arts Analysis. *IADIS WWW/Internet 2007 conference*

Werthner, H. (2003). Intelligent Systems in Travel and Tourism. *18th International Joint Conference on Artificial Intelligence.* Acapulco, Mexico. (August 2003)

Zhdanova, A., Keller, U. (2005). Choosing an Ontology Language. *In Proceedings of the Second World Enformatika Congress* (47-50). Istanbul (2005)

Acknowledgements

OnTourism is a cooperative research project partially funded by the Austrian Research Promotion Agency (FFG).