Semantic Navigation with VIeW*s*

Paul Buitelaar, Thomas Eigner, Stefania Racioppa

DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg 3
66123 Saarbruecken, Germany

paulb@dfki.de

The paper describes VIeWs, a system that combines ontologies, web-based information extraction, and automatic hyperlinking to enrich web documents with additional relevant background information. The central idea behind VIeWs is to demonstrate how web portals can be dynamically tailored to special interest groups by use of corresponding ontologies. As a particular use case we developed an application for the "saarland.de" web portal of the Saarland region in Germany, which we present here in some detail. The paper describes the ideas behind the system and the Saarland.de application and provides an overview of the system architecture and components. Additionally, next to a comparison with related work, also some discussion on end user aspects of the application and its connection to the Semantic Web is given. It is argued that VIeWs is a typical end user application that depends on ontologies as semantic models for different scenarios, but that the need for Semantic Web technology beyond this has not been proven yet.

1 Introduction

The central idea behind VieWs is to demonstrate how web portals can be dynamically tailored to special interest groups by use of corresponding ontologies. For this purpose, the VieWs system combines ontologies, web-based information extraction, and automatic hyperlinking to enrich web documents with additional relevant background information, relative to particular ontologies selected by individual users.

The automatically generated hyperlinks are based on specific ontological "views" on the web portal information, which allow for a high level definition of specific interest topics. As a particular use case we developed an application for the "saarland.de" web portal of the Saarland region in Germany, which we present here in some detail.

The paper is organized as follows: first in section 2 an overview of the VieWs saarland.de application will be described, followed by a brief description of the system architecture and individual components in section 3, and in section 4 by a discussion of end-user issues of the application described here as well as of related work.

2 VieW*s* on saarland.de

The “saarland.de” web portal[1] provides general information on events concerning the local government and institutions. Additionally, sub-sections of the portal include information on various broader topics, such as tourism (“tourismus.saarland.de”), business (“wirtschaft.saarland.de”), etc.

The VIeWs saarland.de application automatically provides users with additional information that is specific to their interests (e.g. hotel information with indication of price and location for tourists or information on the city council, representations of political parties or similar for a local citizen) as derived from the saarland.de portal itself (interlinking portal web pages) or from the web in general (interlinking portal web pages with external information).

2.1 Scenarios

Two application scenarios have been defined and represented in ontologies reflecting the profiles of user groups that correspond to these scenarios:

The **Tourism** scenario reflects a visit of the saarland.de web portal by someone who is interested in tourism options of the Saarland region. The “Tourist” will be interested to know about hotels, restaurants in any city mentioned on the pages of the web portal. In the Tourism ontology this ‘view’ on saarland.de has been defined as follows: a city has Cultural Institutes (Theatre, Cinema), Accommodations (Hotel, “Gasthof”), and Gastronomy (Restaurant, “Konditorei”). These topics are defined in the ontology as classes that are connected over attributes with the class “Stadt” (City). Additionally, every class has attributes such as Location, Number of Rooms, Name, Address and Homepage for the Accommodations class and its subclasses (Hotel, “Gasthof”).

The **Administration (“Verwaltung”)** scenario reflects a visit of the saarland.de web portal by a local citizen who knows the cities in the region but may be interested in specifics, such as administrative offices, political parties, etc. In the Administration ontology this ‘view’ on saarland.de has been defined as follows: a city has a City Administration, Organizations (Political Party, “Wirtschaftsverband”), and Council Offices (“Arbeitsagentur”, ”Standesamt”) In the ontology these topics again are defined as classes that are connected with the class “Stadt”. Additionally, every class has attributes such as Name, Address and Homepage for the Organizations class and its subclasses (Political Party, “Wirtschaftsverband”).

[1] http://www.saarland.de (see http://www.english.saarland.de/ for an English version - only partial)

2.2 Demonstrator

VieWs is a server side application that can be used with a standard web browser[2], which makes it transparent to the normal web user. The user simply browses the saarland.de web portal as normal, but is now being supported by the VieWs system that adds additional information on the basis of a web-based search and from an automatically extracted knowledge base and shows this over generated hyperlink structures. The user can simply decide to follow the regular links or the generated links with added information.

The new links include information from within the saarland.de domain, or also from outside. Depending on the application scenario, this should be set by the user or could be fixed by the system administrator. For instance, in the case of tourism it does make sense to include also external web sites, e.g. hotel home pages. On the other hand, in the case of information for the citizen it may be better to include only 'controlled' information, i.e. only web pages from within the saarland.de domain. The current demonstrator leaves this decision up to the user.

The VieWs entry page[3] for saarland.de enables the user to select their specific interest, currently either "Tourismus" (Tourism) or "Verwaltung" (Administration). By selecting a preference the user automatically enters the VieWs system. From this point on all navigation will be supported by the system according to the selected ontology and, dependent on how the user entered (as a tourist or as a citizen), identified city names will be hyperlinked with additional tourist- or administration-related topics and web-based information.

2.3 VIeWs on Tourism in saarland.de

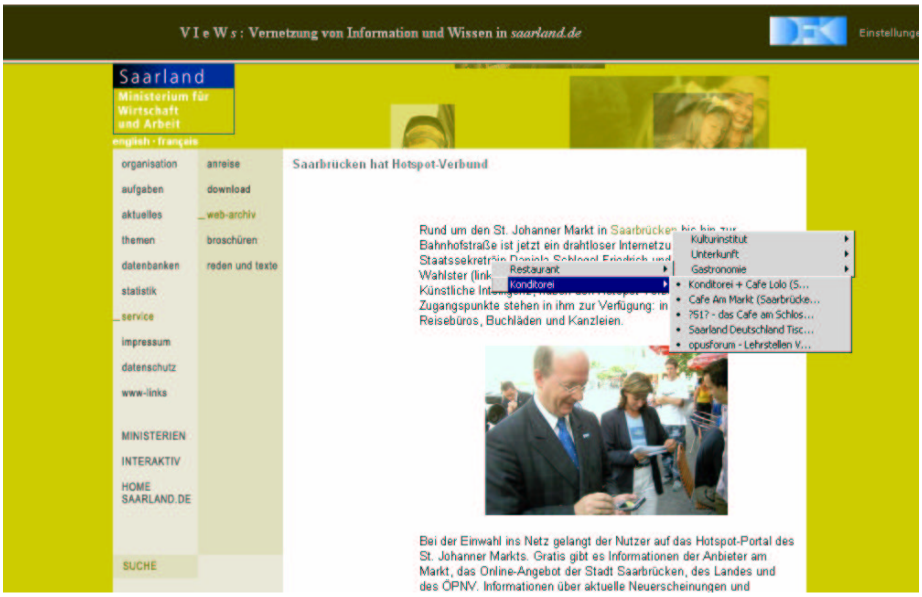
For example, if the user entered as a "Tourist", as shown in Figure 1 below, the generated hyperlink structure shows web links to accommodation (e.g. hotels), dinner options (e.g. restaurants) and cultural institutions (e.g. cinema, theatre) for each identified city name (of the Saarland region) on the page. The added information is included through a Google-based web search for each recognized city name in combination with keywords ("Hotel", "Restaurant", etc.) derived from the ontology class label names.

For selected classes (e.g. hotels) additional information (e.g. address, indication of size, location) is added as shown in Figure 2. This additional information has been previously extracted from retrieved web pages. For this purpose each time a web search has been executed, all retrieved URLs are checked for existence in the knowledge base. If the URL is not in the knowledge base, it will be send to the information extraction component for further extraction of relevant, class-specific information.

The hyperlink structure is generated out of the corresponding ontology, i.e. from the underlying RDF/S file. Over a separate window this structure can be inspected by the user as shown in Figure 3.

[2] The demonstrator has been optimized for Internet Explorer 6.x.

[3] http://views.dfki.de

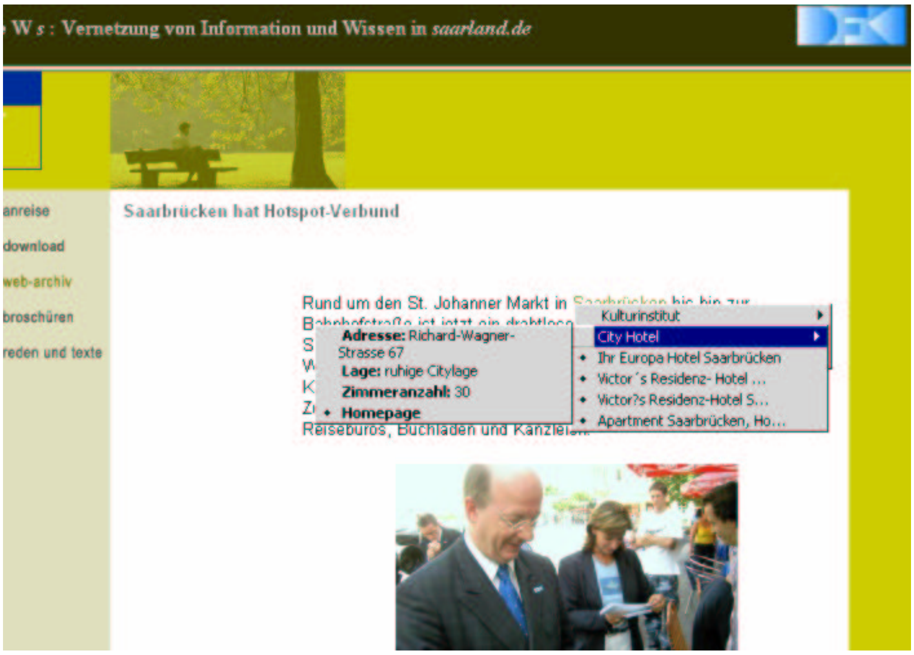


Saarbrücken hat Hotspot-Verbund

Rund um den St. Johanner Markt in Saarbrücken
Bahnhofstraße ist jetzt ein drahtloser Internetzu
Staatssekretärin
Wahlster (link
Künstliche Int
Zugangspunkte stehen in ihm zur Verfügung: in
Reisebüros, Buchläden und Kanzleien.

Bei der Einwahl ins Netz gelangt der Nutzer auf das Hotspot-Portal des St. Johanner Markts. Gratis gibt es Informationen der Anbieter am Markt, das Online-Angebot der Stadt Saarbrücken, des Landes und des ÖPNV. Informationen über aktuelle Neuerscheinungen und

Figure 1: VIeWs with the Tourism Ontology


Ws: Vernetzung von Information und Wissen in saarland.de
anreise
download
web-archiv
broschüren
reden und texte
Kulturinstitut
City Hotel
Ihr Europa Hotel Saarbrücken
Victor´s Residenz- Hotel ...
Victor?s Residenz-Hotel S...
Apartment Saarbrücken, Ho...
Homepage

Saarbrücken hat Hotspot-Verbund

Rund um den St. Johanner Markt in
Reisebüros, Buchläden und Kanzleien.

Adresse: Richard-Wagner-Strasse 67
Lage: ruhige Citylage
Zimmeranzahl: 30

Figure 2: Detailed information on hotels from the Knowledge Base

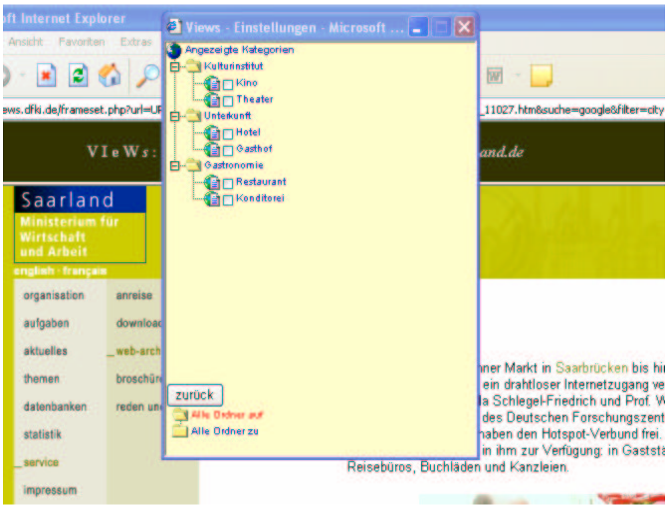


Figure 3: User interaction with the Tourism Ontology

3 The VIeW*s* System

VieWs is implemented as a web-based system and consists of several components as shown in Figure 4 below. The user activates the system over the VieWs web interface as discussed in section 2. The accessed web page is processed by extracting text segments and sending these to the named-entity recognition component for the identification and markup of relevant hyperlink anchors (e.g. city names). For each combination of city name and keyword (“hotel”, “restaurant”, etc.) derived from the ontology, a Google-based web search is started. The results of the web search and information already in the knowledge base is shown in the form of generated hyperlink menus on each of the identified city names. Additionally, an information extraction process is started in the background over the retrieved documents to extract additional relevant information that will be stored in the knowledge base for future access.

The online part of the VIeWs system is written entirely in Java and consists of a hyperlinking component (for generating hyperlink menus in JavaScript), the Google API (for web search with ontology-based keywords), a web service interface with the named-entity recognition component, a database connection with the knowledge base and a crawling component (for downloading the web pages that were retrieved by the web Google API).

The offline part of VIeWs consists of an independently developed information extraction system (the same as used for the online named-entity recognition) and the knowledge base.

The ontologies are an additional static resource that are used online (in building up the hyperlinking menus) and offline (in information extraction).

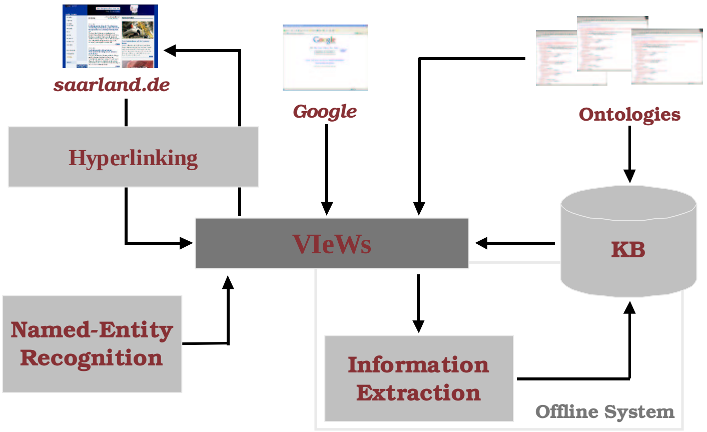


Figure 4: VIeW*s* System Overview

3.1 Hyperlinking

The hyperlinking component takes the accessed web page and regenerates it with the addition of JavaScript hyperlink menus for all identified anchors. The hyperlink structure shows the five best results from Google for each ontology-based keyword (i.e. ontology class name) with stored facts if available.

In this process the following information is integrated:

- Identified hyperlink anchors – named-entity recognition with SProUT
- Ontology structure – ontologies are parsed with Jena
- Results of web search with Google – accessed with Google API
- Stored facts from the knowledge base

3.2 Named-Entity Recognition

The named-entity component is based on SProUT[4], a type-driven information extraction tool that was developed at DFKI [Drozdzynski et al., 2004]. Anchors, e.g. city names, are recognized on the basis of gazetteers and extraction rules over shallow linguistic information (part-of-speech, morphological analysis). A rule in SproUT consists of a regular expression over typed feature structures representing the recognition pattern, and a typed feature structure on the right-hand side that specifies positions and attributes of identified entities in an XML format (see also [Busemann et al., 2003]).

[4] More information on SProUT is available at http://sprout.dfki.de/

3.3 Ontologies

Ontologies are defined using Protégé with export in RDF/S, which is accessed and processed by the VIeWs system to generate a corresponding hyperlink menu in Javascript. As described in section 2 above, each ontology defines a particular user scenario that is organized around a central object class (e.g. cities), over which more specific information objects are defined (e.g. city institutes or organizations). The information structure that is defined in an ontology also guides the information extraction process for filling out the corresponding knowledge base (see also below).

3.4 Web Search

The VieWs system is a hyperlinking application that integrates information on one web page with information from other web pages. For this purpose, a web crawler is included that searches for relevant web pages, given a set of keywords that can be derived from the ontology. The web crawler that we currently use is the Google API, but as it is rather slow and not always reliable in terms of precision we are considering the integration of other search engines (such as Yahoo) or the implementation of a dedicated crawler for the Saarland region.

3.5 Information Extraction

The information extraction component is also based on SProUT and is used offline to derive class-specific information from web pages. For instance, the address, location description (e.g. “central”, “no traffic”, “near railway station”) or the number of rooms for a hotel could be extracted from the hotel home page. The extracted information is stored in the knowledge base and accessed if the corresponding URL of the web page has been retrieved by the web search component. In this way, stored information is only shown if the corresponding web page is still regarded as ‘relevant’ by the web search component (i.e. Google currently).

3.6 Adapting VIeW*s* to Other Domains and Applications

The VieWs system has been designed to be adaptable to other scenarios, either within the saarland.de application or in a completely new application context[5]. For this purpose the following components should be adapted: an ontology should be defined for the new scenario; a corresponding information extraction grammar should be defined; additionally, if the ontology is defined around a different central object class (i.e. different from “cities” in the current implementation) then also the named-entity recognition component should be adapted accordingly.

[5] For instance, we are currently working on an application of VieWs for http://www.dfki.de

4 End User Issues and the Semantic Web

As shown by the examples in this paper, the VieWs system is a typical end user application, in which any level of technological complexity should be kept fully transparent. In this respect it is also irrelevant if the technology used in VieWs is based on Semantic Web technology or not. The main goal is to satisfy user needs in accessing relevant information at the right moment and in the right context.

Nevertheless, exactly this context is the central aspect of the VieWs application that can be expressed by use of available Semantic Web standards and technology. The user context, i.e. a user group profile such as “those web portal visitors interested in tourism”, can be captured in an ontology defined for instance in RDF/S or OWL. Extracted information can be stored in and accessed from a corresponding knowledge base that can be based on Semantic Web technology, such as SESAME, Jena, etc. Reasoning facilities can then also be easily added to the application, e.g. to integrate class-specific semantic web services [Dzbor et al., 2004] or to derive further knowledge by use of rules or axioms.

On the other hand, it is also true that VIeWs in its current form can be implemented without a complete use of Semantic Web standards and tools. Relational databases and other standard technology are equally capable of providing the current functionality of the VIeWs application. Although reasoning capabilities cannot be offered, the use case for these has not been established yet. At the same time, web-based search is central to VIeWs which obviously is also not Semantic Web based.

In summary, semantic context models such as user profiles and associated knowledge bases seem to provide an application scenario for Semantic Web standards and technologies in the VIeWs context, but the use case for this needs still to be proven.

5 Related Work

Related work to VIeWs exists in various respects, i.e. on the level of semantic-based indexing and hyperlinking (e.g. [Pustejovsky et al., 1997], [Carr et al., 2001], [Dill et al., 2003]), information extraction and hyperlinking (e.g. [Busemann et al., 2003], [Popov et al., 2003], [Basili et al., 2004]), and ontologies as user models - in hyperlinking (e.g. [Maedche et al., 2002]).

In general however, VIeWs is most similar to Magpie [Dzbor et al., 2003] although it seems also complementary in some respect. In particular, VieWs integrates an online web search functionality, which makes it very flexible in the kind of information it is able to show. Magpie on the other hand has access only to an underlying static knowledge base. Secondly, VieWs includes an information extraction component that is fully integrated in the automatic processing of retrieved web pages and knowledge base extension and updating. It is not clear if information extraction has been similarly completely integrated with Magpie. Finally, VIeWs can handle most web page formats and seems therefore more robust in real-life applications than Magpie.

6 Conclusions

We presented the VieWs system and its application in the context of the saarland.de web portal. The system consists of clearly defined and efficiently integrated components for web search, information extraction and hyperlinking and has been designed in such a way that it can be readily adapted to other application scenarios and domains.

VieWs can be seen as a Semantic Web application as it uses related standards such as RDF/S and tools such as Jena. On the other hand, the core functionality of Semantic Web applications, reasoning and inference, has not been integrated as the use case for this functionality has not been proven yet. In future work, we will concentrate on identifying the use case for reasoning and inference in the context of real-life applications, such as the saarland.de scenarios described here.

Acknowledgements

This research has been supported by a grant for the project VIeWs by the Saarland Ministry of Economic Affairs.

References

Roberto Basili, Maria Teresa Pazienza, Fabio Massimo Zanzotto *Inducing hyperlinking rules in text collections* In: Proceedings of RANLP2004, John Benjamins, Amsterdam/Philadelphia, 2004.

Stephan Busemann, Witold Drozdzynski Hans-Ulrich Krieger, Jakub Piskorski, Ulrich Schäfer, Hans Uszkoreit, Feiyu Xu *Integrating Information Extraction and Automatic Hyperlinking*. In Proceedings of the ACL-2003 demo session, Sapporo, Japan, 2003.

Leslie Carr, Sean Bechhofer, Carole Goble, Wendy Hall. *Conceptual Linking: Ontology-based Open Hypermedia*. WWW10, Tenth World Wide Web Conference, Hong Kong, May 2001.

Dill S., N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J. A. Tomlin, and J. Y. Zien, *SemTag and Seeker: Bootstrapping the semantic Web via automated semantic annotation*, The Twelfth International World Wide Web Conference Budapest, Hungary, 2003.

W. Drozdzynski, H.-U. Krieger, J. Piskorski, U. Schäfer, F. Xu. *Shallow Processing with Unification and Typed Feature Structures - Foundations and Applications.* In Künstliche Intelligenz, 1/2004.

M. Dzbor, J.B. Domingue, E. Motta *Magpie - towards a semantic web browser*. 2nd International Semantic Web Conference, October 2003, Florida, USA.

M. Dzbor, E. Motta, J.B. Domingue *Opening Up Magpie via Semantic Services*. Proc. of 3rd International Semantic Web Conference (ISWC04). November 2004. Japan.

A. Maedche, S. Staab, R. Studer, Y. Sure and R. Volz. *SEAL - Tying Up Information Integration and Web Site Management by Ontologies*. In: IEEE Computer Society Data Engineering Bulletin, Special issue on "Organizing and Discovering the Semantic Web", Vol. 25, No. 1, pp. 10-17, March 2002.

B. Popov, A. Kiryakov, D. Ognyanoff, D. Manov, A. Kirilov, M. Goranov *Towards Semantic Web Information Extraction.* Human Language Technologies Workshop at the 2nd International Semantic Web Conference (ISWC2003), 20 October 2003, Florida, USA.

J. Pustejovsky, B. Boguraev, M. Verhagen, P.P. Buitelaar and M. Johnston *Semantic Indexing and Typed Hyperlinking* In: Proceedings of AAAI Spring 1997 Workshop on Natural Language Processing for the World Wibe Web, Stanford University, March 1997.