Semantic Blogging Agents: Weblogs and Personalization in the Semantic Web

Wolfgang Woerndl, Georg Groh, Karlheinz Toni

Technische Universitaet Muenchen, Institut fuer Informatik
Boltzmannstr. 3, 85748 Garching b. Muenchen
Germany
{woerndl,grohg,toni}@in.tum.de

Abstract

In this paper we propose an architecture for personalized and context-aware information access in the Semantic Web. We identify different agents and explain how blogs can be utilized to improve information access in this architecture. One ambition is the extraction of semantics from blog entries and other Web resources. We also try to improve trust by exploiting information about user identities.
This paper is intended as a work in progress paper, most of the ideas are still in the early phases.

Areas of interest: [08], [04], [07]

Introduction

Blogging has emerged as a widespread tool for end user publication where people can share their opinions with anyone. Blogs (a.k.a. weblogs) are easy to use and capture (low barrier), and provide a decentralized and distributed form of knowledge management among communities. One interesting idea is to combine blogging with the Semantic Web (Cayzer 2004).

The Semantic Web goal is that agents or other programs understand the "meaning" of information by using ontologies to mark up information sources and items with (semantic) meta data. One application area is the adaptation of information access to a user's needs and context. In the existing Web, users manually search for web sites, e.g. blogs, and have to decide for themselves whether information is relevant and trustworthy. This could be automated or the user could be supported in choosing suitable blogs, at least. Hence, the vision of the Semantic Web is condensed in form of personal agents which assist the user in information retrieval.

So far, some results have been achieved in using ontologies for semantic data integration, for example, but overall, the Semantic Web vision is not a reality yet. This is especially true when it comes to trust in information access. This paper aims to outline some of our ideas, but most of the work is still in progress at this time.

The rest of the paper is organized as follows. First, we present a motivating scenario. Then we characterize Semantic Web information sources. In the next section, we give an overview of our architecture, show how some existing approaches can be used to solve part of the problem and suggest ideas for other parts. We also explain some of the key concepts of our semantic blogging agents and ideas of using identity management to possibly improve trust. We finish with a conclusion where we also outline some related work.

Motivation

Scenario from User's Perspective

In this chapter we outline an end user scenario to further illustrate the mentioned adaptation.

A user searches for information about a journey and needs information such as airline schedules, hotels restaurants at his destination and points of interests. This is one typical Semantic Web scenario, because usually different information sources have to be combined to achieve the task. Blogs could play a bigger role because they not only provide facts about restaurants, for example, but also opinions from user about certain restaurants that may be very valuable for our user. This should be done by matching information sources with user preferences, for example the kind of restaurant a user likes.

Now, if the user is already travelling and is looking for a restaurant to eat right now using a mobile device, the requirements for a personalized search are quite similar but the context is different. Now the query should incorporate different contextual information such as the current location of the user, opening times and directions to the restaurant. It is also important to consider the actual role or identity of the user (e.g. "private" or "work"). For example, points of interests are likely less important on a business trip. In addition, information access has to be adapted to the end device, e.g. desktop PC, PDA or mobile phone.

To summarize, it is not only important to allow for personalized searches but also adapt information access to the current context of the user.

Web Information Sources

Different information sources have to be combined to realize the scenario. In a Semantic Web, Web pages can be enhanced with meta data for better and more personalized

searches. However, "Semantic Web information agents" do not exist yet. We therefore propose "semantic crawling agents" which wrap existing information sources (Woerndl and Groh 2005). The advantage is that existing Web pages can be reused in a Semantic Web and users do not have to manually annotate Web pages with semantics. In addition, it can contribute to initial meta data to generate the Semantic Web network effect ("bootstrapping problem", Cayzer 2004).

We can distinguish between several types of information sources to be analyzed by the semantic crawling agents. Dimension of Web information sources include:

- structured (e.g. Amazon product catalog) vs. semi-structured (e.g. blog or wiki) vs. natural language text (e.g. arbitrary Web page in English): structured sources can be processed rather easily (e.g. existing web services), while natural language text is hard to analyze and interpret
- highly inter-linked (e.g. wikis) vs. "Web islands" (e.g. most corporate web sites): links are used to improve search results and derive relationships between Web resources
- unpersonal (Web page by anonymous authors) vs. personal (e.g. blog entry): more and more Web resources such as blog and wiki contributions are often written by one identifiable person
- semantic vs. keyword-based (e.g. Google) queries: semantic queries might also better facilitate "browsing" instead of "searching". Thus users can browse related concepts, even if they only have a rather vague idea what they are looking for

Blogs are semi-structured, highly linked and personal. Our goal is to move blog query possibilities from keyword-based to semantic and also making more use if the personal character of blogs. The semi-structuredness of blogs provides a minimum of structure that can be utilized in extracting semantics. For example, headlines and links to other Web resources can be identified. Information about authors could be used to improve the trustworthiness of information access and search results.

Architecture and Blogging Agents

Overview of Proposed Architecture

With respect to the classification from above, we propose different types of semantic crawling and information agents with common (semantic) interfaces to be able to search and browse global information space. In this section, we give an overview of our proposed architecture (Fig. 1). The goal is to combine and extend existing solutions to move a step closer towards the vision in the scenario.

The intended procedure is as follows: the user makes a query or uses a client side application for information access. First, the adaptation agent tries to acquire context and user profile information that can be used in the query. The query is then passed on to search agents that find suitable services. One type of service is a "semantic blogging agent" which is explained in more detail in the next section. Services respond to the query, the answer is adapted to the user context (e.g. capabilities of current device) and returned to the user.

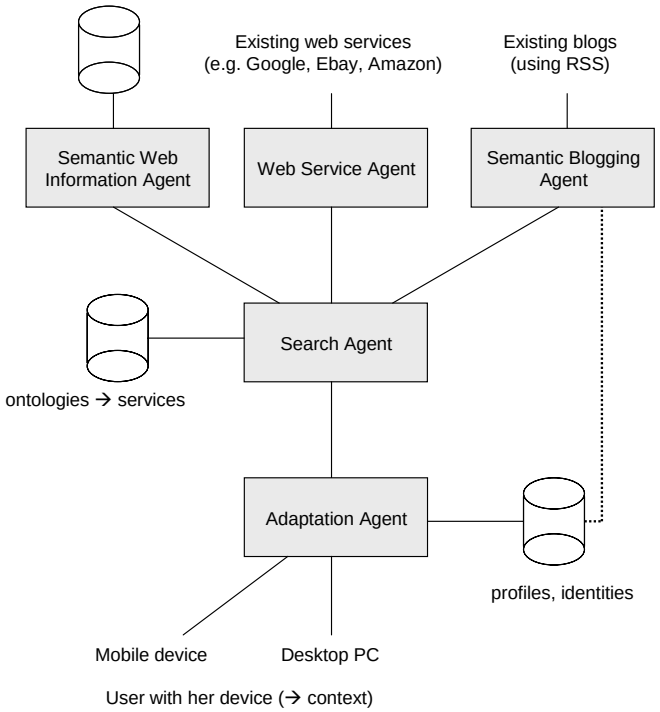

Fig. 1 Architecture

We think it is useful to distinguish between context and profile in this scenario. A user profile contains a user's interests, past transactions, ratings and other personal information. This profile information is rather static and does not change that often. The profile contains different identities such as "private" and "work" roles. Part of the profile is dependent on the identity (e.g. private vs. professional interests), while other parts (e.g. birth date) are not.

The user's context on the other hand includes very dynamic and transient information such as her current physical location, her last search request, her current devices (e.g. if she is currently using a laptop computer or a PDA), possibly her "virtual location" (e.g. reading in a certain blog) and so on. One context attribute is also the current identity of the user.

Adaptation, Search and Information Agents

The adaptation agent enriches the query with relevant profile and context information (e.g. "AsianCuisine" could be added in a restaurant search) and sends it to a search agent. The search agent then retrieves a suitable ontology

for this request. Searching for ontologies can be done by using Swoogle (http://swoogle.umbc.edu/). Swoogle is a indexing and retrieval system that crawls Semantic Web documents (Ding et.al. 2004). Semantic Web Documents are Web resources with meta data (instance data) or ontologies. For example, searching for "restaurants" returns ontologies which feature concepts such as "AsianCuisine".

In addition to finding ontologies, search agents have to maintain a list of services and also ontologies that these services can handle. After the search agent has figured out appropriate services (information agents) with regard to the user query it sends out the query to the information agents (at the top in Fig. 1).

The information agents provide query access using Semantic Web query languages such as SPARQL (Prud'hommeaux and Seaborne 2005). As explained earlier, different information agents are used in our scenario. Web service agents wrap existing Web services and enrich queries with ontologies.

Finally, the search agent analyzes the responses from the other agents and returns the results to the adaptation agent. The adaptation agent additionally personalizes the results according to profile and context information and the user's current device. For example, information items in a language that the user does not understand are dismissed. Content adaptation to different (mobile) end user devices can be done using standards currently under development, for example the Mobile Web Initiative at the World Wide Web Consortium (W3C) (http://www.w3.org/Mobile/).

We are using a software agent based approach because the components on our scenario are very autonomous. The components in our implementation interact with FIPA (Foundation of Intelligent Physical Agents, http://www.fipa.org) messages using the JADE framework (Bellifemine et.al. 2005).

Semantic Blogging Agents

Fig. 2 gives a general idea of the proposed semantic blogging agents. One blogging agent represents one or possibly more blog(s) from one user.

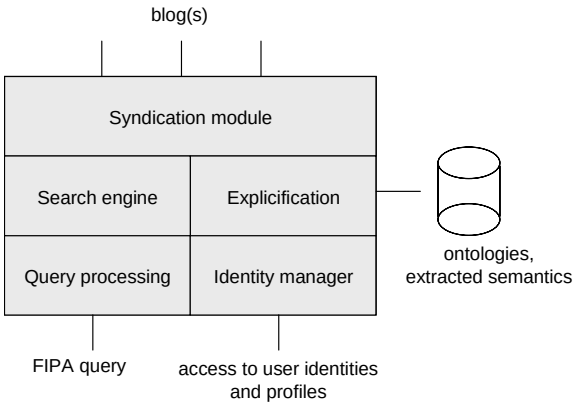

Fig. 2 Components of a semantic blogging agent

The query processing component receives FIPA requests from a search agent (Fig. 2). In addition to the query itself, the request contains a reference to the relevant ontology as specified by the search agent and information about the current identity of the requesting user. The query already contains relevant profile attributes from the requesting user.

The task of the identity manager is to relate the identity of the blog writer to other users and identities. This is done to improve the trustworthiness of information access by taking advantage of the personal nature of blogs. Some corresponding ideas are outlined in the next chapter.

To communicate with the actual blog(s), the blogging agent uses syndications formats such as RSS (see http://web.resource.org/rss/1.0 for RSS 1.0) or Atom (http://www.ietf.org/html.charters/atompub-charter.html).

The semantic blogging agent does not store the blog entries themselves but references to blog entries with additional semantic meta data in a repository. For example, if the explification module discovers that the user is writing about Asian restaurants, this knowledge is added to the repository. In addition to extracted meta data, the repository manages the ontologies the agents understand. Initially, the user could specify relevant ontologies or ontologies could be generated by the explification component. The blogging agent then registers with search agents.

Finally, the task of the search engine module (Fig. 2) is to evaluate information from the other components to generate responses to the query.

Improving Trust through Identity Management

As already mentioned, one advantage of weblogs is that blogs entries can often be associated with actual users. Thereby, the real identity does not have to be revealed, pseudonyms can be used instead. Since we already manage different user identities to customize queries, these user identities can on the other hand utilized to improve trust in our approach.

Information about user relationships can be evaluated using FOAF (Friend-of-a-Friend) (http://www.foaf-project.org). FOAF provides a vocabulary to describe personal information that is associated with Web resources. Thereby, an implicit trust network between users is built. The semantic blogging agents can exploit the FOAF network and compute and evaluate relationships between user identities.

In addition, user profile information can be used to find similar users and rank search results accordingly. If two users share common interests, it may be likely that blog entries are relevant and interesting for the other user. Queries are also be enhanced with user profile attributes to narrow the search.

Finally, users can provide explicit rankings and annotation to blogs and other information sources. This can be done by using a trust ontology we are currently developing. For example, if one user annotates another user's blog with "trustworthy" or "I believe the author is

right", these annotations can be exploited by the semantic crawling agents. Rankings and annotations are stored with a user's profile.

Extracting Semantics from Blog Entries

We are also working on extracting semantic information from natural language text. Blogs represent a good environment for the application of these ideas, because they provide certain structural elements (such as an ordering over time in diary-style blogs) which aid in constructing "explicifying" agents that transform natural language texts into usable Semantic Web structures.

The intention is not to try to understand all nuances of natural language, of course, but to acquire some more rather simple semantic information that can be exploited afterwards. For example, a semantic blogging agent could determine the kind of restaurant a user is writing about in her blog (e.g. "AsianCuisine"). This information can then be matched with users' preferences to improve search results as outlined above.

Our model is based on Description Logics (DL) (Baader et.al. 2003). One goal is to extract facts (A-Box elements in DL). Here we construct a semantic graph model of the text with help of natural language processing (parsing and tagging). On the basis of this model we are able to deduce e.g. RDF Triples using stochastic mappings. Another goal is to extract ontological knowledge (T-Box elements) in case the other agents can not find a relevant ontology for a query.

For the retrieval of existing ontologies and in order to facilitate not only the construction of an "intra-blog" semantic model but also the construction of "inter-blog" semantics, we also need to address the problem of ontology mapping (Lacher and Groh 2001). A "query" ontology has to be compared with a repository of existing ontologies.

Conclusion

The main ideas of our approach can be summarized as follows:

- Components for personalized and context-aware and information in the Semantic Web
- Separate management and modelling of identities, profiles, context and trust
- Extracting semantics from blog entries and other text with natural language processing methods
- Improving trust through identity management

Existing work on semantic blogging focuses on providing a blogging environment on the Semantic Web and utilizing ontologies (e.g. Cayzer 2004, Karger and Quan 2005). Our goal is to use blogs and other Web resources as information provider for the Semantic Web. Other semantic blogging agents could be integrated in our architecture.

There are several existing services that provide personalized information access, for example Google Personalized Search (http://www.google.com/psearch). Google Personalized Search and other similar services can be compared to our approach from a user's perspective but only offer adaptation for one particular task. Collected information about the user cannot be reused for other services, because it is stored and managed prorietarily at the service. But again, Google Personalization Search could be wrapped by a semantic crawling agent and provide a Web information source for our agents.

With most personalization services, the user has little or no control what information about her is collected and her and why (privacy). In our approach, the personal information is managed by the adaptation agent and not by the service. The adaptation agent provides an interface for the user to manage her information and an authorization mechanism can control the access of user profile attributes (Woerndl 2004).

At this time, our work is still in the early phases. We are currently working on implementing and testing our ideas. Activities include designing and implementing the explained semantic blogging agents, as well as Semantic Web information agents that access the web service API's of Ebay, Amazon and Google. The goal is to provide a uniform, semantic interface to very different information sources as explained above. Also, we are currently modelling context, profile and trust to have one or more ontologies that all our agents share. One application area we are particularly interested in is mobile applications.

References

Baader et al. eds. 2003 *Description Logics Handbook.* Cambridge: Cambridge University Press.

Bellifemine, F. et. al. 2003. JADE - A White Paper. http://jade.tilab.com/papers/WhitePaperJADEEXP.pdf.

Cayzer, S. 2004. Semantic blogging and decentralized knowledge management. *Communications of the ACM* 47(12): 47-52.

Ding, L. et. al. 2004. Swoogle: A Search and Metadata Engine for the Semantic Web. In *Proc. Thirteenth ACM Conference on Information and Knowledge Management (CIKM'04).* Washington DC: ACM Press.

Karger, D.; Quan, D. 2005. What Would It Mean to Blog on the Semantic Web? *Journal of Web Semantics* 2(3).

Lacher, S; Groh, G. 2001. Facilitating the exchange of explixit knowledge through ontology mappings. In Proc. 14th International FLAIRS conference. Key West, FL: AAAI Press

Prud'hommeaux, E. ; Seaborne, A. 2005. SPARQL Query Language for RDF. W3C Working Draft.

Woerndl, W. 2004. Authorization of User Profile Access in Identity Management. In Proc. *IADIS International Conference WWW / Internet 2004*. Madrid, Spain: IADIS.

Woerndl, W.; Groh, G. 2005. A proposal for an agent-based architecture for context-aware personalization in the Semantic Web. In *Proc. IJCAI Workshop Multi-Agent Information Retrieval and Recommender Systems*. Edinburgh, UK: IJCAI.