A Web Information Retrieval System Architecture Based on Semantic MyPortal

Haibo Yu[1], Tsunenori Mine[2], and Makoto Amamiya[2]

Department of Intelligent Systems, {Graduate School[1], Faculty[2]} of Information
Science and Electrical Engineering, Kyushu University
6-1 Kasuga-koen, Kasuga, Fukuoka 816-8580, JAPAN
{yu, mine, amamiya}@al.is.kyushu-u.ac.jp

Abstract. In this paper, we mainly focus on a communication mechanism which enables efficient information publishing and sharing among semantic desktops. We propose MyPortal as a "one stop" for all the information relevant to the user and further propose the conceptual architecture of a P2P community Web information retrieval system based on MyPortal. This architecture enables not only precise location of MyPortal instances and their Web resources but also the automatic or semi-automatic integration of hybrid semantic information delivered through Web content and Web services, and it also ensures that the semantics will not be lost during any part of the lifecycle of the information retrieval process.

1 Introduction

Current Web design targets human consumption, based on keywords for information indexing and searching, which not only gives rise to an enormous number of irrelevant search responses, but is unsuitable for machine processing. In addition, the user's desktop information and the published Web information are managed separately, giving rise not only to a redundancy of information but also creating difficulties in managing the relationship among items of information and applying user personalization.

Currently, there are some research projects, such as Haystack [1] and Gnowsis [3] trying to use semantic Web technology for the management of user personal desktop information. However, they lack the functionality for searching, accessing, aggregating and processing of the Web information on the fly when necessary and a unified interface for managing not only the personal desktop information but also the relevant Web information. And a reasonable architecture and efficient mechanisms for the connecting, discovering, and sharing of the information among semantic information nodes are necessary.

In this paper, we make our main concern on how to connect these information nodes in a robust and efficient way, how to discover and share the information among these information nodes and what functionalities need to be provided in order to realize these targets.

We propose our semantic Web information retrieval system architecture based on the following main ideas.

First, “combining Web portal technology with semantic desktop technology to provide a “one stop” for the user to all his relevant information.” As semantic desktop provides a good solution for managing user personal information but lacks the functionality to search, collect and aggregate information from the Web for the user on the fly. On the other hand, Web portals provide a good solution for collecting relevant information for the user, but lack options for personalization and suffer from the problems of centralized architecture. We make use of the basic mechanisms for semantic personal information management of current semantic desktops and enhance their Web information publishing and sharing functionalities to construct a semantic MyPortal.

Second, “using peer-to-peer computing architecture to connect MyPortals with emphasis on an efficient method for reducing communication load.” Decentralized P2P systems are robust, scalable and cheap to maintain, but tend to have large amounts of information transferred among many peers. Hence, an efficient mechanism for reducing communication loads with least loss of precision and recall is very important in a P2P information retrieval system. We propose our Agent-Community-based Peer-to-Peer information retrieval method called ACP2P to connect and manage the communication among MyPortals.

Third, “ensure that the semantics are not lost sight of during any part of the lifecycle of information retrieval.” In order to enable consumer re-using semantic data, we designed the interfaces and the protocols involved in the whole life cycle of information retrieval tasks with semantic technology.

Fourth, “all participants contribute to the semantic description consistently.” Efficient searching for high quality results is based on pertinent matching between well-defined resources and user queries, where the matching reflects user preferences. We use Web site capability description (WSCD) to describe the capabilities of MyPortal and submit user queries consistently.

Fifth, “integrating Web information delivered through Web contents and Web services.” Conventional Web contents and Web services have been managed separately as they targeted different consumer, we will support the integrated management of semantic Web contents and Web services at different levels in MyPortal.

2 MyPortal

MyPortal is a “one stop” that links the user to all the information s/he needs. It is at the user’s own desktop, which is also a Web server itself and is designed to manage user’s personal information with semantic Web technology in a flexible personalized way. It provides both semantic browser and semantic search engine functionalities and these functions manage not only local user desktop information but also the remote semantic MyPortal information. Its information can be published through Web contents and Web services and shared by others with proper authority.

The structure of MyPortal is shown in Fig 1. It consists of following four components: core component provides basic support for semantic Web technologies and knowledge management, user interface component provides a unified interface for creating, browsing, querying, and managing of the relevant information, desktop information management component manages the conventional personal information such as documents, e-mail, contact information, and communication component which is the delegate of the user for communication with other MyPortals.

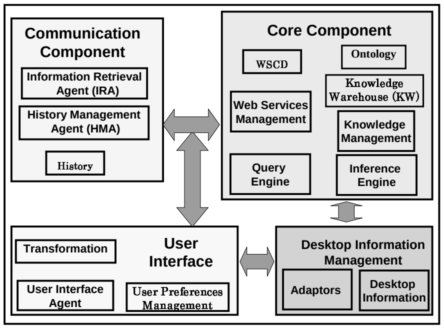

Fig. 1. Structure of MyPortal

One can refer to [4] for a little more detail for MyPortal.

3 Conceptual Architecture of Web Information Retrieval System Based on MyPortal

Our conceptual architecture for a community semantic Web information retrieval system is illustrated in Fig 2.

The architecture consists of three main components: a “consumer” which searches for Web resources, a “provider” which holds certain resources, and a mediator which enables the communication between the consumer and the provider. In our architecture, the providers and consumers are all MyPortal. Each provider describes its capabilities in what we call a WSCD (Web site capability description), and each consumer will submit relevant queries based on user requirements when a Web search is necessary. The mediator is comprised of agents assigned to the consumer and providers using an Agent-Community-based P2P information retrieval method to fulfill the search and access tasks.

3.1 Connecting MyPortals with ACP2P method

The communication between consumer and providers is based on an Agent-Community-based Peer-to-Peer information retrieval method called ACP2P method[2],

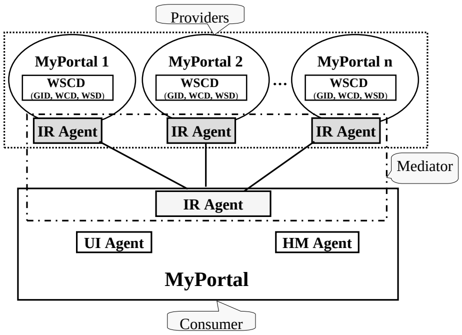

Fig. 2. A Conceptual Architecture

which uses agent communities to manage and look up information related to a user query.

In order to retrieve information relevant to a user query, an agent uses two histories: a query/retrieved document history (Q/RDH for short) and a query/sender agent history (Q/SAH for short). Making use of the Q/SAH is expected to have a collaborative filtering effect, which gradually creates virtual agent communities, where agents with the same interests stay together.

The ACP2P method employs three types of agents: user interface (UI) agent, information retrieval (IR) agent and history management (HM) agent. A set of three agents (UI agent, IR agent, HM agent) is assigned to each user. Although a UI agent and an HM agent communicate only with the IR agent of their user, an IR agent communicates with other users' IR agents to search for information relevant to its user's query. A pair of Q/RDH and Q/SAH histories and retrieved content files are managed by the HM agent.

The ACP2P method is implemented with Multi-Agent Kodama (Kyushu university Open & Distributed Autonomous Multi-Agent) [6]. Kodama comprises hierarchical structured agent communities based on a portal-agent model. A portal agent is the representative of all member agents in a community and allows the community to be treated as one normal agent outside the community.

We are currently planning to use SPARQL RDF query language and SPARQL protocol as our semantic communication interfaces between providers and consumers.

3.2 Web site capability description (WSCD)

Resource location is based on matching between user requirements and Web site capabilities, hence a capability description of MyPortal is necessary. We describe the layered capabilities of MyPortal by layers.

First, we semantically describe the general capabilities of the Web site, and we call this a "general information description (GID)." The GID gives an explicit overview of the Web site capabilities such as their category, topic, and can be used as the initial filter for judging congruence with user preferences. Second, we give the Web content capability description (WCD), it is the metadata of Web contents and is composed of knowledge bases of all domains involved. Third, we give the Web service capability description (WSD) which is further expressed by two layers: “a semantic Web service description (SWSD)” and “a concrete Web service description (CWSD).” This hierarchical capability-describing mechanism enables semantic and non-semantic Web service capability-describing and matchmaking for different levels.

For the details of our Web site capability description mechanism, one can refer to document [5].

4 Conclusion

In this paper, we addressed our main ideas on constructing a P2P community semantic Web information retrieval system based on MyPortal, mainly focused on how to connect MyPortals to enable automatic and efficient information sharing and what functionalities are necessary when constructing a MyPortal. In the future, we will realize a prototype of MyPortal and a P2P community Web information retrieval system based on MyPortal, and evaluate the effectiveness of our approaches. Experiments in using the ACP2P method for semantic Web data retrieval in a dynamic multiple community environment will also be carried out.

References

1. D. Huynh, D. Karger, and D. Quan. Haystack: A Platform for Creating, Organizing and Visualizing Information Using RDF. In *Proceedings of the International Workshop on the Semantic Web (at WWW2002)*, 2002. http://semanticweb2002.aifb.uni-karlsruhe.de/proceedings/Research/huynh.pdf.
2. T. Mine, D. Matsuno, A. Kogo, and M. Amamiya. Design and implementation of agent community based peer-to-peer information retrieval method. In *Proc. of Eighth Int. Workshop CIA-2004 on Cooperative Information Agents (CIA 2004), LNAI 3191*, pages 31–46, 9 2004.
3. L. Sauermann. The Gnowsis Semantic Desktop for Information Integration. In *IOA Workshop of the VM2005 Conference*, 2005.
4. H. Yu, T. Mine, and M. Amamiya. Towards a Semantic MyPortal. In *The 3rd International Semantic Web Conference (ISWC 2004) Poster Abstracts*, pages 95–96, 2004.
5. H. Yu, T. Mine, and M. Amamiya. Towards Automatic Discovery of Web Portals -Semantic Description of Web Portal Capabilities-. In *Semantic Web Services and Web Process Composition: First International Workshop, SWSWPC 2004, LNCS 3387/2005*, pages 124–136, 2005.
6. G. Zhong, S. Amamiya, K. Takahashi, T. Mine, and M. Amamiya. The Design and Implementation of KODAMA System. *IEICE Transactions on Information and Systems*, E85-D(4):637–646, April, 2002.