Agent-Community-based P2P Semantic Web Information Retrieval System Architecture

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 propose a conceptual architecture for a personal semantic Web information retrieval system. It incorporates semantic Web, Web service, P2P and multi-agent technologies to enable not only precise location of Web resources but also the automatic or semi-automatic integration of Web resources delivered through Web contents and Web services. In this architecture, the semantic issues concerning the whole lifecycle of information retrieval were considered consistently and the integration of Web contents and Web services is enabled seamlessly.

1 Introduction

1.1 Motivation

With ever-increasing information overload, Web information retrieval systems are facing new challenges for helping people not only locating relevant information precisely but also accessing and aggregating a variety of information from different resources automatically.

Currently, new technologies for enabling precise and automatic machine processing such as semantic Web and Web services are emerging and have attracted more and more attentions from academia and industry in recent years.

The semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation [4]. Currently, there are a lot of researches such as [10] [9] [24] trying to apply semantic Web technologies to Web information retrieval systems, but they all address only problems concerning certain phases or certain aspects of the total complex issues involved. There isn't any research addressing the semantic issues from the whole life cycle of information retrieval and architecture point of view. However, for the reasons we show below, we argue that it is important to clarify the requirements of a Web information retrieval system architecture to apply semantic web technology to it. First, we need to ensure the semantics are not lost sight of during the whole life cycle of information retrieval, including publishing, querying, accessing, processing, storing and reusing. Hence the interfaces involved in the whole life cycle of information retrieval tasks need to be re-considered. Second, efficient searching for high quality results is based on pertinent matching between well-defined resources and user queries, where the matching reflects user preferences. Therefore the description of Web site capability and the way of submitting queries incorporating user preferences should consistently be considered from an architectural point of view.

Web service mechanisms provide a good solution for application interoperability between heterogeneous environments. It will provide a new way for accessing Web information and play a vital role in Web information retrieval activities in the future. However, the conventional "Web contents" resources target at human consumption but new "Web services" resources target at machine consumption. Thus they have been managed separately for publishing, discovering, accessing, and processing until now. On the other hand, in the semantic Web, contents are given well-defined meaning, and they are becoming such data that can be understood and processed by machine as well. As both Web contents and Web services will be consumed by machines, this introduces the possibility and necessity of managing them together in a personal Web information retrieval system.

In this paper, we propose a conceptual architecture for a personal semantic Web information retrieval system. It incorporates semantic Web, Web services, peer-to-peer and multi-agent technologies to enable not only precise location of Web resources but also the automatic or semi-automatic integration of hybrid semantic information delivered through Web contents and Web services.

1.2 Approach

The conceptual architecture of our semantic Web information retrieval system is constructed based on the following four main ideas.

First, "using peer-to-peer computing architecture with emphasis on efficient method for reducing communication load." As a centralized system has the bottle neck of accessing and its maintenance cost is expensive, scalable and decentralized P2P systems are receiving more and more attention especially in the research and product development for the open and dynamic Web environment. However, due to the decentralization, the performance comes to be a significant concern when a large number of messages are propagated in the network and large amounts of information are being transferred among many peers [14]. Hence, an efficient mechanism for reducing communication load with least loosing 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 method [15]. On the other hand, as we noticed that generally users retrieve and re-use certain amount of information repeatedly for a high percentage, it is essential to store frequently used information locally for the user with an efficient retrieval mechanism. We enable users refining and storing retrieved Web information in their local environment and manage them with semantic Web technology for the later re-use. As the user interested information is limited resource and the storing and retrieving mechanism can be adjusted to the user-specific way, the access time to the most frequently used information will be decreased significantly than searching in the vast open Web. As the possibility of accessing external resources is decreased, the searching time and network transfer time are all saved.

Second, "all participants contribute to the semantic description consistently." The Web information retrieval system concerns three main kinds of participants: the "consumer" which searches for Web resources, the "provider" which holds certain resources, and the "mediator" which enables the communication between the consumer and the provider. In order to guarantee semantic interoperability during the whole life cycle of information retrieval, all participants need to consistently contribute to the semantic description. The provider needs to precisely describe their capabilities and the users need to pertinently describe their requirements as well. The mediator needs to correctly interpret the semantic dimension and to ensure that semantics are not lost sight of during the processing.

Third, "integrating Web contents with Web services." As we mentioned earlier, Web services will provide a new way for retrieving Web information. In fact, Web users do not care about how the system discovers, accesses and retrieves information from what kind of resources, they only care about the final results which can be used by them conveniently. Hence, the particular characteristics and the concrete realization details of both Web services and Web contents need to be hidden from users as much as possible. Therefore an integrated or unified management of Web contents and Web services needs to be carried out through different levels including the description of capabilities and requirements, querying, discovering, selection and aggregation.

Fourth, "providing a gateway to all the information that the user is interested in." A user generally concern two types of information: local and remote information. The local information such as documents, emails, contacts and schedule are stored at user's desktop and managed by various of applications, and the remote information such as Web information which are published by its Web site manager and can be searched and accessed through Web applications. Since the user needs to access and process all these local and remote information, a gateway providing an unified interface to all relevant information is necessary. We propose a personalized "Myportal [26]" to satisfy all the information requirements of a user.

The rest of the paper is organized as follows: Section 2 outlines our conceptual architecture, the components and their communication mechanism of a personal semantic Web information retrieval system. The process flow of an information retrieval system will be described in section 3. In Section 4, we will discuss related work and the concluding remarks will be summarized in section 5.

2 A Conceptual Architecture

Our conceptual architecture for a personal semantic Web information retrieval system is illustrated in figure 1.

The architecture consists of three main components: "consumer" which searches for Web resources, "provider" which holds certain resources, and mediator which enables the communication between the consumer and the provider. In our archi-

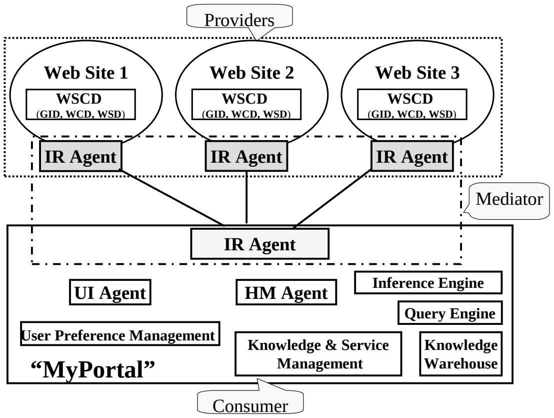

Fig. 1. A Conceptual Architecture

tecture, each provider describes its capabilities in what we call a WSCD (Web site capability description), each consumer is constructed as a "Myportal" providing a gateway to the information relevant to the user. The mediator is conformed by agents assigned to the consumer and providers using Agent-Community-based P2P information retrieval method to fullfill the searching and accessing tasks. We will describe each component of our architecture in a little more detail in following sub-sections.

2.1 Web site capability description (WSCD)

Resource location is based on matching between user requirements and Web site capabilities, so a capability description of Web sites is necessary. We describe the layered capabilities of a Web site as shown in figure 2.

First, we semantically describe the general capabilities of the Web site, and we call this a "general information description (GID)." We argue that some explicit general ideas about a Web site are strongly required in order to precisely locate Web resources based on user preferences. Therefore a brief general information description of the Web site is defined at the top level. The GID gives an explicit overview of the Web site capabilities, and can be used as the initial filter for judging congruence with user preferences.

Second, we give the Web content capability description (WCD) and there is a link from GID to WCD for using semantic Web contents. The WCD is the metadata of Web contents and is composed of knowledge bases of all domains involved. We use OWL [13] to describe domain ontologies and the metadata will be described in RDF [12].

Third, we give Web service capability description (WSD) and there is a link from GID to WSD for facilitating the further matching and use of Web services. In order to semantically describe the capabilities and support the concrete

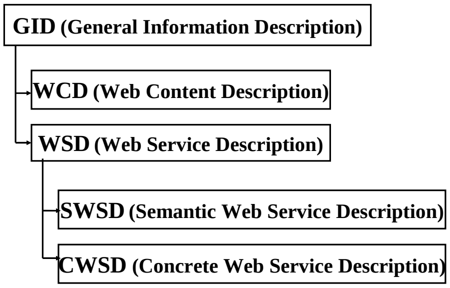
GID (General Information Description)
WCD (Web Content Description)
WSD (Web Service Description)
SWSD (Semantic Web Service Description)
CWSD (Concrete Web Service Description)

Fig. 2. Web Site Capability Description

realization of services, we express the service capability description in two layers: "semantic Web service description (SWSD)" and "concrete Web service description (CWSD)." This hierarchical capability-describing mechanism enables semantic and non-semantic Web service capability-describing and matchmaking for different levels. we use WSDL [7] for the concrete Web service description and OWL-S [6] to express the semantic Web service description.

For the details of our Web site capability description mechanism, one can refer to document [27].

2.2 "Myportal"

"Myportal" is a "one stop" that links the user to all the information s/he needs. It resites on the user's own desktop which is also a Web server itself and is designed to satisfy user's personal information requirements and to be mastered freely by the user her/himself. It provides both semantic browser and semantic search engine functionalities, and these functions manage not only local user information but also the other Web sites as conventional browser. The information can be shared by others with proper authority. The structure of "Myportal" is shown in figure 3.

"Myportal" is composed of three types of main functional components: core component, consumer component and provider component.

The core component provides basic support for semantic technologies and information management. It consists of "Knowledge Warehouse (KW)," "Knowledge Management," "Query Engine (QE)" and "Inference Engine (IE)." As a consumer, it will bring together a variety of necessary information from different resources automatically or semi-automatically for the user. It is assigned agents to fulfill the information retrieval tasks through the communication with provider agents. As a provider, the contents and services of "Myportal" can be consumed by humans as well as machines. The human can be the user or other permitted persons, and the machine can be local or remote. A unified inter-

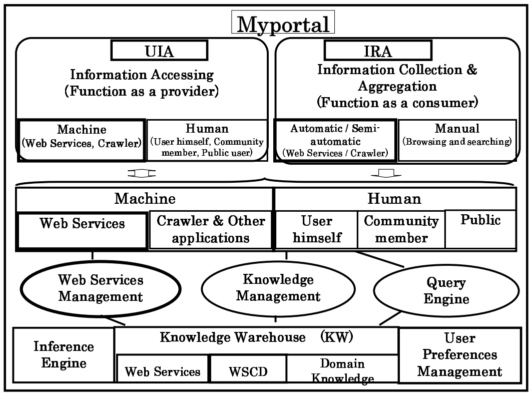

Fig. 3. Structure of "Myportal"

face for browsing, searching and facilitating Web contents and services will be provided. We described "Myportal" in a little more detail in document [26].

2.3 Mediator

The communication between consumer and providers is based on an Agent-Community-based Peer-to-Peer information retrieval method called ACP2P method, 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. We have demonstrated through several experiments that the method reduced communication loads much more than other methods which do not employ Q/SAH to look up a target agent, and was useful for creating a "give and take" effect, i.e., as an agent receives more queries, it acquires more links to new knowledge[16].

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.

The UI agent receives requirements from the user, factors in missing or inherent information based on user preferences, breaks and transforms the requirements into formal queries and sends them to the IR agent.

When receiving a query from a UI agent, an IR agent asks an HM agent to look up target agents with its history or asks a portal agent to do it using a query multicasting technique. When receiving a query from other IR agents, an IR agent looks up the information relevant to the query from its original content

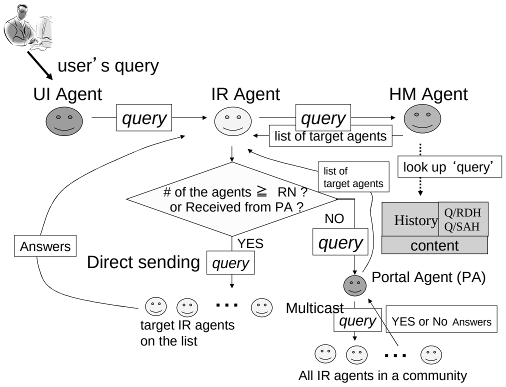

Fig. 4. Actions for Sending a Query

and retrieved content files, sends an answer to the query-sender IR agent, and also sends a pair of the query and the address of the query-sender IR agent to an HM agent so that it can update Q/SAH.

The returned answer is either a pair of a 'Yes' message and retrieved information or a 'No' message indicating that there is no relevant information. When receiving answers with a 'Yes' message from other IR agents, the IR agent sends them to a UI agent, and sends them with a pair of a query and the addresses of answer sender IR agents to an HM agent.

The ACP2P method is implemented with Multi-Agent Kodama (Kyushu university Open & Distributed Autonomous Multi-Agent) [28]. 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. A portal agent has its role limited in a community, and the portal agent itself may be managed by another higher-level portal agent. A portal agent manages all member agents in its community and can multicast a message to them. Any member agent in a community can ask the portal agent to multicast its message. Fig.5 shows the agent community structure which the ACP2P method is based on.

The query language and protocol communicated between IR agents need to be defined. Since the semantic Web information is commonly based on RDF which is a recommendation of W3C, a standard interface for querying and accessing RDF data is ideal for the interoperability between heterogeneous semantic Web information environments. The W3C RDF Data Accessing Working Group (DAWG) has published their working drafts of RDF Query Language SPARQL [19] and SPARQL protocol that are expected to be standards in this field. Although our architecture is designed for any reasonable communication interfaces, we are currently planning to use SPARQL RDF query language and

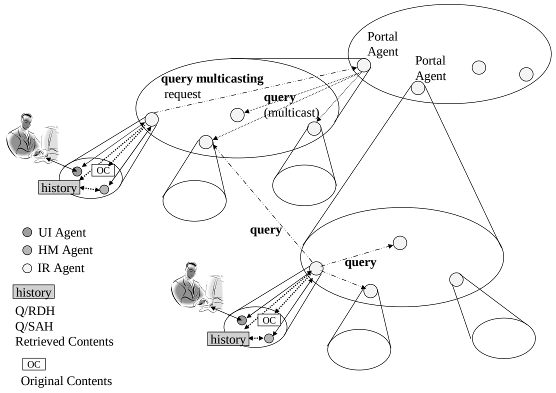

Fig. 5. Agents and their Community Structure

SPARQL protocol as our semantic communication interfaces between providers and consumers.

3 Process Flow of Information Retrieval

The total process flow of the P2P Web information retrieval system can be illustrated as shown in figure 6.

At the beginning, the consumer (the user) will edit his profile and preferences, and providers will describe their capabilities using WSCD which we described in section 2.1. When the user wants to search for the information, a search interface need to be provided. Although various kinds of user interface, such as natural language, template-style or formal equation, can be considered, taking user convenience and reality in account, we provide a template-style search interface, enabling users to input or select their preferences as well as query items from recommendation lists. The missing or inherent information will be inferred based on the user profile and preferences, and the requirements will be broken down and transformed into formal queries. The formal query is composed of three types of element fields: user preferences (UP), content query (CQ) and Web service query (SQ). The search will be carried out first inside the MyPortal knowledge warehouse, and only when we cannot find satisfactory information from MyPortal, will we extend the search to the other providers and the request will be sent to the candidate information retrieval agents on provider side (IRA-P for short) through the information retrieval agent on consumer side (IRA-C for short). The information discovery on the provider side is based on matching between user requirements and provider capabilities. We do matching at three levels. First, we do matching of Web site general description (GID) against user preferences to see whether they match at the overview level or not. Second, we do matching of Web contents, and finally do the matching of Web services. A matching score

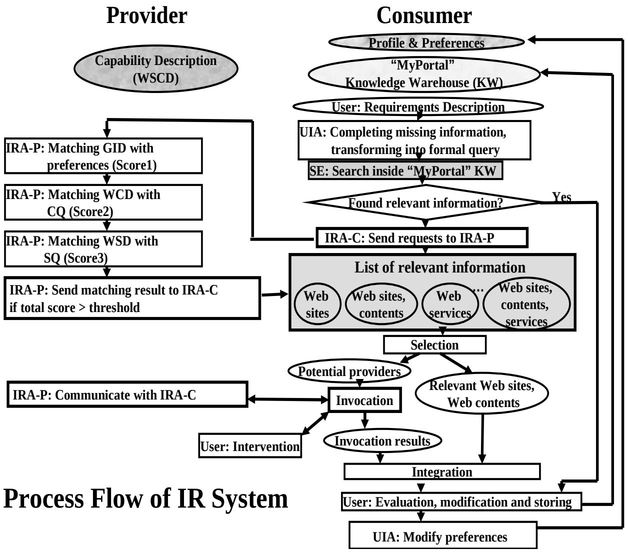
Provider
Consumer
Profile & Preferences
Capability Description (WSCD)
"MyPortal" Knowledge Warehouse (KW)
User: Requirements Description
UIA: Completing missing information, transforming into formal query
SE: Search inside "MyPortal" KW
Found relevant information?
Yes
IRA-P: Matching GID with preferences (Score1)
IRA-P: Matching WCD with CQ (Score2)
IRA-P: Matching WSD with SQ (Score3)
IRA-C: Send requests to IRA-P
IRA-P: Send matching result to IRA-C if total score > threshold
List of relevant information
Web sites
Web sites, contents
Web services
...
Web sites, contents, services
Selection
Potential providers
Relevant Web sites, Web contents
IRA-P: Communicate with IRA-C
Invocation
User: Intervention
Invocation results
Integration
Process Flow of IR System
User: Evaluation, modification and storing
UIA: Modify preferences

Fig. 6. Process Flow

will be given from the matching of each level and they will be used for the final judgment of relevance of Web contents and Web services. IRA-Ps will send back their matching scores to the IRA-C, and the IRA-C will judge and select the most relevant Web services and Web contents based on a total consideration of those matching scores. After selecting the most relevant Web services, the IRA-C will invoke those services. If the input information is not sufficient to trigger invocation, the IRA-C will request the user to provide necessary information through the UIA. The results from different Web service invocations as well as the Web contents results will be aggregated by the IRA-C into a refined, final result based on user preferences and be sent to the user through the UIA. This result can be evaluated, modified and stored in the user's MyPortal knowledge warehouse for future reuse. The integration of different Web service invocation results and Web contents is based on their common RDF data model.

4 Related work

In this section, we discuss some related work that is directly or indirectly of interest to our research work.

Francisco et al. [18] presented an architecture for an infrastructure to provide interoperability using trusted portals and implemented such an infrastructure based on Thematic Portals. The searching portals use semantic access points based on metadata for more precise searching of the resources associated with the potential sources of information. The proposed architecture supports specific and cross domain searching, but only provides semantic representation for the capabilities of Web contents not for their services as far as we understand. Our semantic Web site capability description and pertinent user requirements and preferences description provide interoperability for both Web contents and Web services.

RSS [21] and Atom [17] are lightweight multipurpose extensible metadata descriptions and syndication formats. FOAF vocabulary [5] provides a collection of basic terms that can be used in machine-readable Web homepages for people, groups, companies and so on. RSS, Atom and FOAF vocabulary all focus on certain kinds of Web contents description such as news, Web blog or people, they do not include Web services as we proposed. Our Web site capability description describes not only Web contents but also Web services, so the resources of the portal can not only be located but also used as a computational part of the information retrieval system. RSS, Atom and FOAF can be used for the Web contents capability description which is a part of our Web site capability description.

There are Web portals based on Semantic Web technology, such as KA2 [1] and SEAL [24], but they target uniform access by large numbers of people for human navigation and searching. SEAL provided an interface for a software agent but only for a crawler. None of them supports Web services for information aggregation and publishing at present, as far as we know. Our "Myportal" is a personalized gateway to all user-relevant information and it not only aggregates Web information but also shares its information through Web services.

Haystack [10] and Gnowsis [22] are semantic Web enhanced desktop environment for personal information management. The main purpose of them is to semantically manage user's local information enabling an individual to flexibly manipulate his/her information with a personalized way. They are not constructed from the Web portal point of view and doesn't emphasize the support of machine interoperability between users enabling Web service functionalities. We refer to their ideas of personalization on information management and the integration of existing desktop applications, construct our semantic personal information system as a fully personalized Web portal to provide a gateway to access to not only the local personal information but also the Web information. The "Myportal" acts as both a consumer and a provider to form a basic unit of a P2P information retrieval system. The Web services will be used not only for information retrieval but also for information delivery.

There is lots of work related to the ACP2P method. Structured P2P networks associate each data item with a key and distribute keys among directory services using a Distributed Hash Table (DHT)[23, 20, 25]. Hierarchical P2P networks use the top-layer of directory services to serve regions of the bottom-layer of leave nodes, and directory services work collectively to cover the whole network[8, 2, 3, 11]. The common characterlistics of both approaches are the construction of an overlay network to organize the nodes that provide directory services for efficient query routing. The ACP2P also employes a hierarchical P2P computing architecture based on the Multi-Agent Kodama [28] framework. Unlike other work, the ACP2P makes use of only local information : retrieved documents and two histories, Q/RDH and Q/SAH for query routing. Especially, the Q/SAH is an important clue to search for the relevant information. Further the characteritics using only local information makes the ACP2P responsive to the dynamic environment where the global information is not available.

5 Conclusion

In this paper, we addressed the main aspects of a semantic Web information retrieval system architecture trying to answer the requirements of next-generation semantic Web users. In the architecture, the semantic issues and the integration of Web contents and Web services are considered for the whole lifecycle of information retrieval. Our "Myportal" aims at constructing a fully personalized user's local Web portal, which is adapted to user preferences and satisfies all the requirements of a user's local and Web information usage. We use the ACP2P method for the communication between consumer and providers, which uses agent communities to manage and look up information related to a user's query in order to reduce communication loads in a P2P computing architecture.

In the future, we will realize a prototype of an agent community based P2P personal semantic Web information retrieval system, and evaluate the effectiveness of our proposed architecture based on it. Further experiments for the ACP2P method on semantic Web data retrieval need to be done to see its effectiveness.

References

1. KA2 Portal. http://ka2portal.aifb.uni-karlsruhe.de/.
2. Kazaa v3.0. http://www.kazaa.com/.
3. M. Bawa, G. S. Manku, and P. Raghavan. Sets: search enhanced by topic segmentation. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 306 – 313, 2003.
4. T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, May, 2001.
5. D. Brickley and L. Miller. FOAF Vocabulary Specification. Sept., 2004.
6. David Martin et al. OWL-S 1.1 Release, November, 2004. http://www.daml.org/services/owl-s/1.1/.
7. Erik Christensen et al. Web Services Description Language (WSDL) 1.1, March 15, 2001. http://www.w3.org/TR/wsdl.
8. Gnutella. http://gnutella.wego.com/, 2000.
9. R. Guha, R. McCool, and E. Miller. Semantic Search. In *Proceedings of WWW2003*, pages 700–709, 2003.
10. 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.
11. J. Lu and J. Callan. Content-based retrieval in hybrid peer-to-peer networks. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 199–206, 2003.

12. F. Manola and E. Miller. RDF Primer, February 10, 2004. http://www.w3.org/TR/rdf-primer/.
13. D. L. McGuinness and F. van Harmelen. OWL Web Ontology Language Overview, February 10, 2004. http://www.w3.org/TR/2004/REC-owl-features-20040210/.
14. D. S. Milojicic, V. Kalogeraki, R. Lukose, K. Nagaraja, J. Pruyne, B. Richard, S. Rollins, and Z. Xu. Peer-to-Peer Computing. Technical report, HP, 2002. http://www.cs.wpi.edu/ claypool/courses/4513-B03/papers/p2p/p2p-tutorial.pdf.
15. T. Mine, D. Matsuno, A. Kogo, and M. Amamiya. Acp2p : Agent community based peer-to-peer information retrieval. In *Proc. of Third Int. Workshop on Agents and Peer-to-Peer Computing (AP2PC 2004)*, pages 50–61, 7 2004.
16. 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.
17. M. Nottingham. The Atom Syndication Format 0.3 (pre-draft), December, 2003. http://www.mnot.net/drafts/draft-nottingham-atom-format-02.html.
18. F. Pinto, C. Baptista, and N. Ryan. Using Semantic Searching for Web Portal Interoperability. In *International Workshop on Information Integration on the Web - Technologies and Applications, April 9-11, Rio de Janeiro - Brazil*, April 2001.
19. E. Prud'hommeaux and A. Seaborne. SPARQL Query Language for RDF, April 19, 2005. http://www.w3.org/TR/rdf-sparql-query/.
20. S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker. A scalable content-addressable network. In *SIGCOMM*, pages 161–172, 2001.
21. RSS-DEV Working Group. RDF Site Summary (RSS)1.0, 2000-12-06. http://web.resource.org/rss/1.0/.
22. L. Sauermann. The Gnowsys Semantic Desktop for Information Integration. In *IOA Workshop of the VM2005 Conference*, 2005.
23. I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan. Chord: A scalable peer-to-peer lookup service for internet applications. In *Proceedings of the 2001 conference on applications, technologies, architectures, and protocols for computer communications*, pages 149–160, 2001.
24. N. Stojanovie, A. Maedche, S. Staab, R. Studer, and Y. Sure. SEAL – a framework for developing SEmantic PortALs. In *Proceedings of the International Conference on Knowledge Capture*, pages 155–162, 2001.
25. C. Tang, Z. Xu, and S. Dwarkadas. Peer-to-peer information retrieval using self-organizing semantic overlay networks. In *SIGCOMM*, 2003.
26. 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.
27. 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.
28. 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.