Peer-to-Peer Paradigm for a Semantic Search Engine

S. Bergamaschi[1,2], F. Guerra[1]
e-mail: {bergamaschi.sonia,guerra.francesco}@unimo.it

[1] Dipartimento di Ingegneria dell'Informazione
Università di Modena e Reggio Emilia
Via Vignolese 905, 41100 Modena
[2] CSITE-CNR Bologna
V.le Risorgimento 2, 40136 Bologna

Abstract. This paper provides, firstly, a general description of the research project SEWASIE and, secondly, a proposal of an architectural evolution of the SEWASIE system in the direction of peer-to-peer paradigm. The SEWASIE project has the aim to design and implement an advanced search engine enabling intelligent access to heterogeneous data sources on the web using community-specific multilingual ontologies. After a presentation of the main features of the system a preliminar proposal of architectural evolutions of the SEWASIE system in the direction of peer-to-peer paradigm is proposed.

1 Introduction

Peer-to-peer (hereafter P2P) consists of an open-ended network of distributed computational *peers*, where each peer can exchange data and services with a set of other peers called *acquaintances*. Peers should be autonomous in choosing their aquaintances. Moreover, it is usually assumed that there is no global control in the form of a global registry, global services, or global resources management nor a global schema or data repository. Gnutella and Napster [8] made the P2P paradigm popular as a version of distributed computing between traditional distributed systems and the web. Very recently a proposal in data management raised by this paradigm has been presented in [4]. In this context, each peer may have data to share with other peers and, in [4], it is assumed that each peer's database is relational and, since the data residing in different databases may have semantic inter-dependencies, peers are allowed to specify *coordination formulas*. Coordination formulas explain how data in one peer must relate to data in an acquientance and may also act as constraints or for propagating updates.

Peer' s need an acquaintance initialization protocol where two peers exchange views of their databases and agree on levels of coordination and the level of coordination should be dynamic, i.e. peers should be able to establish and modify acquintances, with little human intervention. This is a crucial point and introduces a high degree of innovation into the traditional distributed databases and multi-database systems data management approach. The common assumption in this area is, in fact, to have a global database schema, usually obtained by skilled databases designers [13, 9]. In the new dynamic setting of P2P, we cannot assume the existence of a global schema for all databases in a P2P network or even those of the acquianted databases. Nevertheless, as proposed both in [4] and in this paper, the architecture of heterogeneous distributed databases or often called *multi-database systems* e.g. Multibase [11] , Momis [2, 1, 3], Garlic [5], TSIMMIS [6], and Information Manifold [7] is still valid. In most of these systems, a user issues queries to a global schema, and the system (called a *mediator* in [12]) maps the queries to subqueries on the underlying data sources. Each data source has a wrapper able to map subqueries into its native query language. A database designer is responsible for creating the global schema and the mappings with the data sources and for maintaining the schema and mappings with respect to evoution (i. e. data sources entering and leaving the system).

The SEWASIE system, presented in the paper, organizes and manages information in SINodes which follow the architecture of heterogeneous multi-database systems.

After a presentation of the main features of the system, section 2, a preliminar proposal of architectural evolutions of the SEWASIE system in the direction of peer-to peer paradigm is proposed in section 3.

2 The SEWASIE project and architecture

SEWASIE (Semantic Webs and AgentS in Integrated Economies) (IST-2001-34825) is a research project founded by EU on action line Semantic Web (May 2002/April 2005) (http://www.sewasie.org). The goal of the SEWASIE project is to design and implement an advanced search engine enabling intelligent access to heterogeneous data sources on the web via semantic enrichment to provide the basis of structured secure web-based communication. A SEWASIE user has at his disposal a search client with an easy-to-use query interface able to extract the required information from the Internet and to show it in an easily enjoyable format. From an architectural point of view, the SEWASIE prototype will provide a search engine client and indexing servers and ontologies.

The project will develop an agent-based secure, scalable and distributed system architecture for semantic search using community-specific multilingual ontologies; ontologies will be equipped with an inference layer grounded in WC3 standards. The developed system have to meet the needs of SMEs in a EU context.

The SEWASIE vision helps European enterprises to compete in a global market and to form strategic alliances at a European level by providing a sophisticated retrieval, brokering and communication service on basis of the semantic web technology. In particular, SEWASIE has to help European SMEs to find the right strategic information at the right time in a multinational environment; provide advanced and novel services for monitoring and linking information in the context of risk management and competitor analysis; provide ontology-based communication mechanisms for negotiation in multi-language environments; ease the use of complex cross-language retrieval and data condensation tools by providing intuitive interfaces.

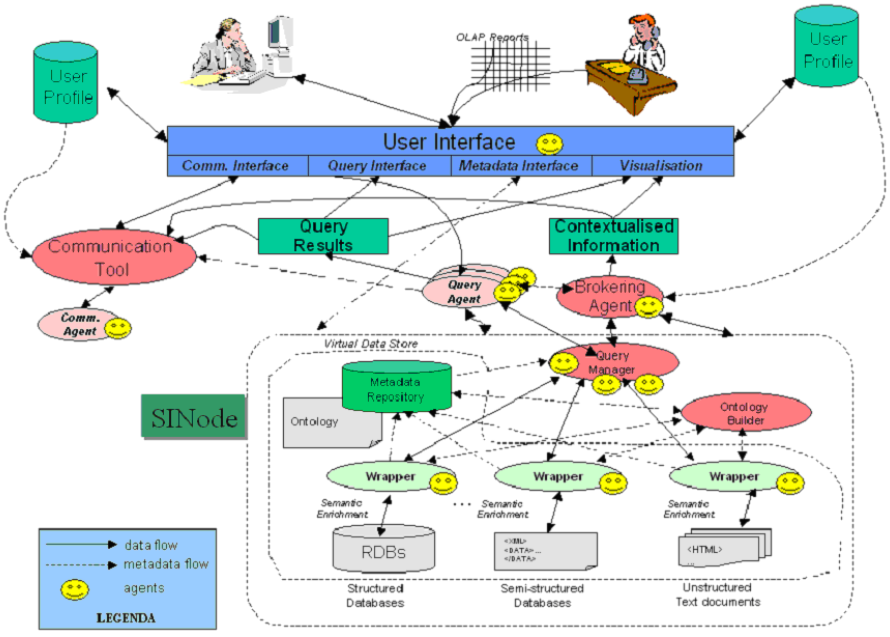

Fig. 1. The Sewasie Architecture

From an architectural point of view (see figure 1) , the SEWASIE system will realise a virtual network, SEWASIE Virtual Network (SVN) whose nodes are SEWASIE Information Nodes (SINode):

- SINodes are mediator-based systems, each including a Virtual Data Store, an Ontology Builder, and a Query Manager;
- The managed Information Sources are heterogeneous collections of structured, semi-structured, or unstructured data, e.g. relational databases, XML or HTML documents;
- Ontologies are multilingual;
- In query solving phase, starting from a specified SINode, a Query Agent accesses other SINodes and thus collects partial answers;
- A Query Agent communicates with the Brokering Agent to acquire useful SINodes.
- The Brokering Agent maintains the knowledge related to the SEWASIE Virtual Network and the user profiles;
- The Brokering Agent classifies SINodes, it is responsible for handling the acquisition of a new SINode and for consequently updating of the SEWASIE Virtual Network.

3 SEWASIE in a P2P architecture

In the general case, a P2P system has no centralized schema and no central administration. In the SEWASIE architecture we rely on two centralized aspects: the brokering agent (global control) that holds the knowledge of the overall network and the global schema or data repository of the network. Furthermore, SINodes are passive elements whose data and metadata are extracted by query agents and the brokering agent. How can we change the SEWASIE architecture in order to evolve towards a peer-to-peer paradigm?

We can define two alternative P2P networks (see figure 2):

- **Brokering Agents Network.** We can devise more brokering agents, one for each SINode, holding both SINode knowledge and coordination knowledge. Furthermore, within the Brokering Agents Network, each Brokering Agent communicates with other peers in order to have information about other information nodes.
- **INTER SINodes Network.** A SINode provides to other SINodes the knowledge about the involved information sources. It is possible to specify coordination formulas that explain how the data in one peer must relate data in an acquaintance.

The Brokering Agents P2P network generates a distributed knowledge about the involved information sources and may provide a support for generating coordination formulas (e.g. by using schema matching [10, 2] , by deriving relations among the peers using inference techniques).

The Brokering Agents network supports the generation of the query plan in order to identify which SINodes have to be queried. In particular, the Network may:

- Generate interest groups with nodes that have similar content.
- Help the query optimization, by giving information to the query agent about the "data placement". A peer knows how data are distributed across the nodes and the query plan may take into account the existing resources and bandwidth constraints.

SINodes network is an alternative approach where the P2P pardigm is directly supported by SINodes. In this case, we may maintain a single brokering agent, holding the knowledge of the network topology and we need a P2P layer in each SINode holding the following functionalities:

- the P2P layer of a SINode needs a protocol for establishing an acquaintance dinamically;
- the P2P layer of a SINode could offer semi-automated support for generating coordination formulas, e.g. by using schema matching [10, 2];
- the P2P layer of a SINode can use approaches for query processing of multidatabase systems but also needs a policy to propagate subqueries through chains of P2P connections (i..e. the path for the query agents);
- the P2P layer of a SINode should be able to advertise its ontology presumably using a directory service (this service could be held/exported towards a brokering agent); this information is useful to create acquaintances and individuate other nodes with similar contents.

4 Conclusion

In this paper we provided, firstly, a general description of the research project SEWASIE and, secondly, a proposal of an architectural evolution of the SEWASIE system in the direction of peer-to-peer paradigm. The SEWASIE project has the aim to design and implement an advanced search engine enabling intelligent access to heterogeneous data sources on the web using community-specific multilingual ontologies.

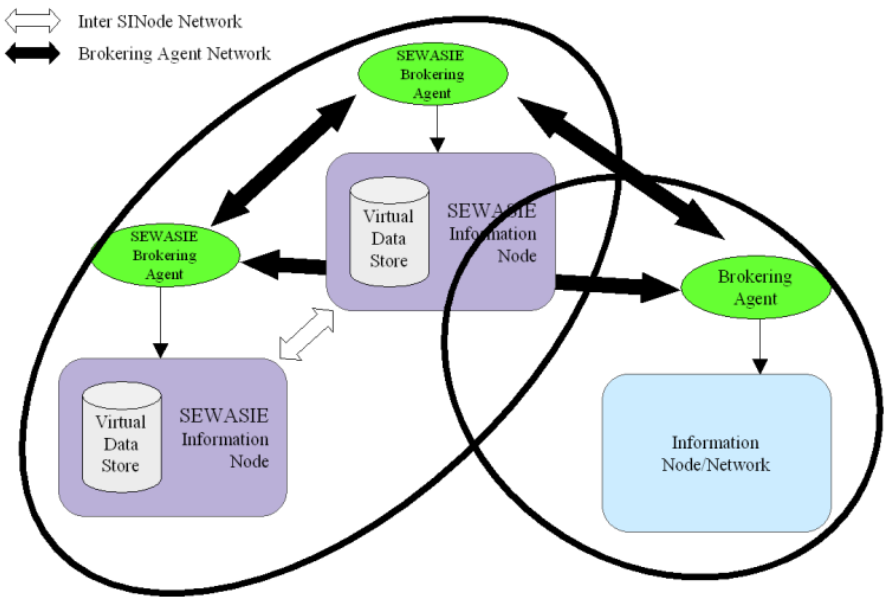

Fig. 2. The two alternative P2P networks

After a presentation of the main features of the SEWASIE architecture system a preliminar proposal of architectural evolutions of the SEWASIE system in the direction of peer-to-peer paradigm have been proposed.

The next step of research is the design of the SEWASIE system architecture following the P2P paradigm. This step includes the study and development of a coordination laanguage and of a negotiation protocol among peers.

Acknowledgements

This work is supported in part by the 5th Framework IST programme of the European Community through project SEWASIE within the Semantic Web Action Line. The SEWASIE consortium comprises in addition to the author' organization (Sonia Bergamaschi is the coordinator of the project), the Universities of Aachen RWTH (M. Jarke), Roma La Sapienza (M. Lenzerini, T. Catarci), Bolzano (E. Franconi), as well as IBM Italia, Thinking Networks AG and CNA as user organisation.

References

1. D. Beneventano, S. Bergamaschi, S. Castano, A. Corni, R. Guidetti, G. Malvezzi, M. Melchiori, and M. Vincini. Information integration: The momis project demon-

stration. In *VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases, September, 2000, Cairo, Egypt*, pages 611–614. Morgan Kaufmann, 2000.

2. S. Bergamaschi, S. Castano, D. Beneventano, and M. Vincini. Semantic integration of heterogenous information sources. *Journal of Data and Knowledge Engineering*, 36(3):215–249, 2001.
3. S. Bergamaschi, S. Castano, and M. Vincini. Semantic integration of semistructured and structured data sources. *SIGMOD Records*, 28(1), March 1999.
4. P. A. Bernstein, F. Giunchiglia, A. Kementsietsidis, J. M., L. Serafini, and I. Zaihrayeu. Data management for peer-to-peer computing: A vision. In *Proceedings of Fifth International Workshop on the Web and Databases(WebDB2002) Madison, Wisconsin*, Jun 2002.
5. M.J. Carey, L.M. Haas, P.M. Schwarz, M. Arya, W.F. Cody, R. Fagin, M. Flickner, A. Luniewski, W. Niblack, D. Petkovic, J. Thomas II, J.H. Williams, and E.L. Wimmers. Towards heterogeneous multimedia information systems: The garlic approach. In *RIDE-DOM*, pages 124–131, 1985.
6. S. S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. D. Ullman, and J. Widom. The tsimmis project: Integration of heterogeneous information sources. In *Proceedings of the 10th Meeting of the Information Processing Society of Japan, October 1994*, pages 7–18, 1994.
7. A. Y. Levy, A. Rajaraman, and J. J. Ordille. Querying heterogeneous information sources using source descriptions. In *Proc. of VLDB 1996*, pages 251–262, 1996.
8. A. Oram. *Peer-to-Peer: Harnessing the Power of Disruptive Technologies.* O'reilly, 2002.
9. M. T. Ozsu and P. Valduriez, editors. *Principles of Distributed Database Systems.* Prentice Hall, 1999.
10. E. Rahm and P. A. Bernstein. A survey of approaches to automatic schema matching. In *The VLDB Journal: Volume 10 Issue 4 (2001)*, pages 334–350, 2001.
11. J.M. Smith, P.A. Bernstein, U. Dayal, N. Goodman, T. Landers, K.W.T. Lin, and E. Wong. Multibase – integrating heterogeneous distributed database systems. In *Proceedings of 1981 National Computer Conference*, pages 487–499. AFIPS Press, 1981.
12. G. Wiederhold. Mediators in the architecture of future information systems. *IEEE Computer*, 25:38–49, 1992.
13. W.Litwin, L. Mark, and N. Roussopoulos. Interoperability of multiple autonomous databases. *ACM Computing Surveys*, 22(3):267–293, 1990.