Semantic Query Routing in SenPeer, a P2P Data Management System

David Faye[1], Gilles Nachouki[2], and Patrick Valduriez[2]

[1] LANI, Université Gaston Berger de Saint-Louis, Sénégal
David.Faye@univ-nantes.fr
[2] LINA, Université de Nantes, France
Gilles.Nachouki@univ-nantes.fr,
Patrick.Valduriez@inria.fr

Abstract. A challenging problem in a schema-based peer-to-peer (P2P) system is how to locate peers that are relevant with respect to a given query. In this paper, we propose a new semantic routing mechanism in the context of the SenPeer *P2P Data Management System* (PDMS). SenPeer is an unstructured P2P system based on an organization of peers around super-peers according to their semantic domains. Our proposal is based on the use of a distributed data structure, called expertise table, maintained by the super-peers and describing data at the neighboring peers. This table, combined with matching techniques, is the basis of a semantic overlay network. Semantic links are exploited for efficient query propagation towards peers that may have relevant data. We give a performance evaluation of our semantic query routing with respect to important criteria such as precision, recall and number of messages. The results show that our algorithm significantly outperforms a baseline algorithm without semantics.

1 Introduction

PDMS aim at overcoming the scalability problems of data integration systems by combining P2P and distributed database techniques. Peers join the system by providing their own schemas and matching their respective schemas to discover their acquaintances for effective data sharing. In such a setting, a major problem is efficient query routing across peer data sources, i.e. deciding to which peers the query must be sent for efficiency and effectiveness. To motivate our discussion, let us consider a PDMS[6] where experts and scientists share data about the development of the Senegal river. The data are covered by topics such as hydraulics, agronomic research, health, etc. For these multidisciplinary applications, schema sharing and efficient query routing are crucial in order to enable users discovering remote data sources.

Unstructured P2P systems typically employ flooding and random walk approaches to locate files, which result in much network traffic and low recall/ precision. To improve performance, peer clustering and indexing allow peers to select from an index the relevant peers to send a query to. In some PDMS with

T. Enokido, L. Barolli, and M. Takizawa (Eds.): NBiS 2007, LNCS 4658, pp. 365–374, 2007.

complex queries facilities, peer selection relies on the use of semantic descriptions of peers. We observe that they are essentially based on statistical observations and exploit, in some cases, a shared ontology, or a taxonomy. In some cases, flooding is unavoidable.

In this paper, we propose a new semantic routing mechanism in the context of SenPeer, a PDMS with various data models and ontologies. We assume an unstructured P2P system where peers are connected to super-peers according to their semantic domains. In SenPeer, the knowledge of a peer is represented through a semantic network (called *sGraph*) which can represent data conforming to various data models. To interact, super-peers maintain expertise tables describing data at the semantically linked (super)-peers and define semantic mappings between their content descriptions. These mappings are the basis of semantic overlay network where peers having similar schemas form a semantic neighborhood. This semantic overlay is exploited further to address query propagation. Our routing technique is applied in the presence of several data models provided that there are wrappers which undertake the transformation of the query to a suitable query language for each data source. We make the following contributions :

(i) We introduce the SQUEL language, for exchanging queries between peers.
(ii) We propose a distributed data structure called *expertise table* that is used for advertising content shared by a peer.
(iii) we propose a semantic routing algorithm of queries toward relevant peers.
(iv) We provide an experimental validation. The results show that our routing algorithm significantly outperforms a baseline algorithm without semantics.

The remainder of this paper is organized as follows. Section 2, introduces our semantic overlay. Section 3 presents our semantic query routing algorithm. Section 4 gives an experimental validation. Section 5 discusses relevant work. Finally, Section 6 concludes and discuss our future work .

2 Semantic Overlay

In this section, we present the techniques for the semantic overlay formation.

2.1 Basic Context

Assume a PDMS in which each peer hosts a data source. Our main goal is the efficient search across the PDMS by routing queries only to relevant peers. To support peer autonomy, network self-organization, we choose an unstructured super-peer network. We adopt a super-peer network topology that combines the efficiency of centralized search with the autonomy, load balancing and robustness of distributed search. Peers are attached to super-peers according to their topic of interest or semantic domain. Each super-peer SP_j responsible of a semantic domain D_j suggests a schema SS_j for that domain. However, each peer describes its data with its own schema S_i. Moreover, each peer P_i submits queries with its