Information Systems 36 (2011) 174–191

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/infosys

A unified multimedia and semantic perspective for data retrieval in the semantic web ☆

Claudio Gennaro [a], Rita Lenzi [b], Federica Mandreoli [b,d], Riccardo Martoglia [b,*], Matteo Mordacchini [a], Wilma Penzo [c,d], Simona Sassatelli [b]

[a] ISTI - CNR, Pisa, Italy
[b] DII - University of Modena e Reggio Emilia, Italy
[c] DEIS - University of Bologna, Italy
[d] IEIIT - BO/CNR, Bologna, Italy

ARTICLE INFO

Keywords:
PDMS
Network organization
Query routing
Semantics
Multimedia data

ABSTRACT

In recent years, the emerging diffusion of peer-to-peer networks is going beyond the single-domain paradigm like, for instance, the mono-thematic file sharing one (e.g. Napster for music). Peers are more and more heterogeneous data sources which need to share data with commercial, educational, and/or collaboration purposes, just to mention a few. Moreover, in current information processing applications data cannot be meaningfully searched by precise database queries that would return exact matches (e.g. when dealing with multimedia, proteomic, statistical data).

In this paper we move a step towards multi-domain multi-type data sharing systems by introducing an advanced technological infrastructure which enables users to meet these new emerging needs.

A fundamental issue in this context is data heterogeneity, which is pervasive and intrinsically present both at intensional level where, due to peers' autonomy, different semantic descriptions of the available information are provided, and at extensional level, where multiple data types can coexist, also including content-based searchable data types such as multimedia data.

Our proposal relies on a Peer Data Management Systems (PDMS) framework to present innovative *network organization* and *query routing* mechanisms which exploit both peers' data description and data content to achieve effective and efficient network management and data retrieval in such a context. The validity of our proposal is demonstrated by an absolutely satisfactory experimental evaluation on a real setting.

1. Introduction

The ever-growing and widespread availability of data from Internet information sources has placed great interest on the potential of information sharing. This awareness has led to the flourishing of Peer-to-Peer systems, i.e. systems with distributed computing capabilities, where each peer exchanges information and services directly with other peers.

Nevertheless, in recent years these systems have demonstrated to be too much restrictive as to the

☆ This work has been partially supported by the Italian National Research Council in the context of the NeP4B Project and of the research initiative "Ricerca a tema libero" and by the IST FP7 European Project S-Cube (Grant Agreement no. 215483).
* Corresponding author.
E-mail addresses: claudio.gennaro@isti.cnr.it (C. Gennaro), rita.lenzi@unimo.it (R. Lenzi), federica.mandreoli@unimo.it (F. Mandreoli), riccardo.martoglia@unimo.it (R. Martoglia), matteo.mordacchini@isti.cnr.it (M. Mordacchini), wilma.penzo@unibo.it (W. Penzo), simona.sassatelli@unimo.it (S. Sassatelli).

0306-4379/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.is.2010.07.002

potentialities they offer for data retrieval. More precisely, they do not appropriately support new information sharing needs which are going beyond the single-domain paradigm like, for instance, the mono-thematic file sharing one (e.g. Napster for music).

On the other hand, the Web fosters the vision of an Internet-based global agora where peers are more and more heterogeneous data sources which need to share various data types with commercial, educational, and/or collaboration purposes, just to mention a few. Moreover, in current information processing applications data cannot be meaningfully searched by precise database queries that would return exact matches [1]. This is particularly true for the multimedia domain which is par excellence the one where data are commonly searched by similarity on its content. However, several other data domains show the same needs. We refer to biochemical entities, medical observations, weather forecast, and so on.

In this paper we move a step towards multi-domain multi-type data sharing systems by introducing an advanced technological infrastructure which enables users to meet these new emerging needs.

In the scenario we consider, peers are autonomous entities which expose the data they want to share with each other for interoperability purposes. A fundamental issue in this context is data heterogeneity, which is pervasive and intrinsically present both at intensional level where, due to peers' autonomy, different semantic descriptions of the available information are provided, and at extensional level, where multiple data types can coexist, also including content-based searchable data types.

A recent framework which well models peer interoperability in this scenario is represented by Peer Data Management Systems (PDMSs). Indeed, PDMSs have been addressed in [2,3] as one powerful means for large-scale data sharing among semantic peers, as they ensure information exchange and cooperation, while also completely preserving full peer autonomy, network scalability and dynamism. In a PDMS, peers are heterogeneous data sources, each having its own content modeled upon a local schema that represents the peer's domain of interests. Because of the absence of a common understanding of the vocabularies used at each peer's schema, semantic relationships, called mappings, are established locally between pairs of peers, thus implementing a decentralized schema mediation. Hence, joining a PDMS is inherently a more heavyweight operation than joining a P2P file-sharing system, since some semantic relationships need to be specified. This effort is definitely payed back in contexts like the one considered, where, for instance, the membership changes infrequently or it is restricted due to security or consistency requirements [3]. As an example, this is the case of companies which organize and coordinate themselves in order to develop common and shared opportunities, while respecting their own autonomy and heterogeneity (e.g. enterprises belonging to the same holding). In these contexts, peers are likely to stay available the majority of the time, though being able to join (or add new data) very easily. Thus, the essentially stable nature of a PDMS makes the system quite different from a traditional P2P setting where peers are online very shortly.

As far as the language for data representation is considered, we rely on OWL as Semantic Web standard for data sharing. A sample OWL-based scenario of a PDMS concerning data about floor covering is shown in Fig. 1. The boxes on the right-hand part of the figure depict excerpts of the OWL ontologies exposed by peers `Peer1` and `Peer2`. Data sources include content-based searchable data, such as images of products offered by peers. This information is represented by means of appropriate OWL properties in the peers' ontologies (e.g. `image` in Peer1's ontology of Fig. 1).

The presence of this kind of data types gives rise to the need of a powerful query language which enables the users to also express content-based, aka similarity, predicates. As an example, let us consider the query shown in the shaded box on the left-hand side of Fig. 1. The query is formulated by using a SPARQL-like language properly extended to support similarity predicates on multimedia data, and asks Peer1 for "Companies that sell Italian tiles similar to the represented one and that cost less than 30 euros": the `LIKE` operator in the `FILTER` clause is used to search for images whose similarity on color[1] with the image provided as argument is greater than 0.3. Moreover, the query also specifies the number of expected results in the `LIMIT` clause.

In a PDMS, queries are answered globally: starting from the querying peer p, a query q is first evaluated against the p's local repository, then p forwards q towards its neighborhood and each receiving peer behaves in the same way. Therefore, answers to q can come from any peer p' of the network that is connected to p through a semantic path of mappings. Since a query is formulated on p's ontology, in order to be answered by p', the query has to be reformulated according to its own ontology. To this purpose, semantic mappings are used in a multi-step reformulation of q along the semantic path connecting p to p' [4,5]. For instance, in Fig. 1 the concept `origin` may translate to `country` when the query is forwarded to Peer2.

In this context, query answers are inevitably approximated. On the one hand, each reformulation step may lead to some semantic approximation and, consequently, the returned data may not exactly fit the data description specified in the original query. On the other hand, when a similarity predicate is evaluated at a peer, the returned data are intrinsically approximated as to its content.

This paper leverages such a query answering misalignment as a means for efficiently and effectively supporting query processing in our context. To this purpose, we move along two main directions.

First, we introduce an innovative *network organization* technique which exploits peers' data description and data content to instantiate the network topology according to the linkage closeness of similar peers. This gets two main advantages: (1) it reduces the information loss that it is likely to be along long paths in the PDMS because of missing (or incomplete) mappings [4,6], and (2) it can save computational efforts due to the evaluation of similarity predicates, which is known to be a very costly operation [1], on irrelevant peers, by favoring the execution of

[1] `SC` is a keyword to denote the descriptor `scalable color` for image color according to the MPEG-7 standard.