On Semantically-Augmented XML-Based P2P Information Systems

Alfredo Cuzzocrea

Department of Electronics, Computer Science, and Systems
University of Calabria, Italy
cuzzocrea@si.deis.unical.it

Abstract. Knowledge representation and extraction techniques can be efficiently used to improve data modeling and IR functionalities of P2P Information Systems, which have recently attracted a lot of attention from industrial and academic researchers. These functionalities can be achieved by pushing semantics in both data and queries, and exploiting the derived expressiveness to improve file sharing primitives and lookup mechanisms made available from first-generation P2P systems. XML-based P2P Information Systems are a more specific and interesting instance of this class of systems, where the overall data domain is composed by very large, Internet-like distributed XML repositories from which users extract useful knowledge manly by means of IR methodologies implemented on the top of XML join queries. This paper focuses on several aspects of XML-based P2P Information Systems, raging from foundations and definitions to knowledge representation and extraction models and algorithms, along with their experimental evaluation. However, the results presented in this paper can also be adapted to deal with any kind of data format (e.g., HTML).

1 Introduction

During the last years, there was a growing interest for *P2P Information Systems* (P2P IS) [1,2], mainly because they fit a wide number of real-life IT applications. Digital libraries are only a significant instance of P2P IS, but it is very easy to foresee how large the impact of P2P IS on innovative and emerging IT scenarios, such as *e*-procurement and *e*-government, will be during the next years.

P2P networks are natively built on the top of a very large repository of data objects (e.g., files), which is intrinsically distributed, fragmented, and partitioned among *participant* peers. P2P Users are usually interested in (*i*) retrieving data objects containing information of interest, like video and audio files, and (*ii*) sharing information with other (participant) users/peers. From the *Information Retrieval* (IR) perspective, P2P users (*i*) typically submit short, loose queries by means of keywords derived from natural language-style questions (e.g., "*find all the music files containing Mozart's compositions*" is posed through the keywords "*compositions*" and "*Mozart*"), and (*ii*), due to resource-sharing purposes, are usually interested in retrieving as result a *set* of data objects rather than only one. Then, well-founded IR methodologies like ranking can be successfully applied on intermediate results (i.e., sets of data objects) to improve system capabilities thus achieving performances better than those of more

H. Legind Larsen et al. (Eds.): FQAS 2006, LNAI 4027, pp. 441 – 457, 2006.

traditional database-like query schemes. Furthermore, the P2P IR mechanism is "self-alimenting" as intermediate results can be then re-used for sharing new information, or for setting and specializing new search/query activities. In other words, from the database perspective, P2P users typically adopt a semi-structured (data) model for querying data objects rather than a structured (data) model. On the other hand, efficiently accessing data in P2P systems, which is an aspect directly related with the above issues, is a relevant and still incompletely solved open research challenge [1].

Basically, P2P IS extend the traditional functionalities of P2P systems (i.e., file sharing primitives and simple lookup mechanisms based on partial- or exact-match of strings), by adding to the primitive of the latter useful (and more complex) knowledge representation and extraction techniques. Achieving the definition of new knowledge delivering paradigms over P2P networks is the main goal of this effort. In fact, the completely decentralized nature of P2P networks, which enable peers and data objects to come and go at will, allows us to (*i*) successfully exploit self-alimenting mechanisms of knowledge production, and (*ii*) take advantage from innovative knowledge representation and extraction models based on semantics, metadata management, probability etc.

Despite more or less advanced query strategies, all today P2P systems are devoted to cover their initial goals, and, as a consequence, there is a strength, effective demand for enriching P2P systems with functionalities that (*i*) are proper of the information systems, such as *Knowledge Discovery* (KD)- and IR-style data object querying, and (*ii*) cannot be supported by the actual data representation and query models of (traditional) P2P systems. Hence, P2P IS represent an attempt to support even complex processes like knowledge representation, discovering, and management over P2P networks. More properly, knowledge representation and management techniques mainly concern with the modeling of P2P IS, whereas knowledge discovering techniques (implemented via IR functionalities) mainly concern with the querying (i.e., knowledge extraction) of P2P IS.

All considering, we can claim that traditional P2P functionalities are inadequate for the innovative requirements of P2P IS, whereas P2P systems relying on XML repositories, whose management has reached a sufficient maturity by now, offer from a side all the typical functionalities supported by a P2P system, such as information sharing, dynamism, and scalability, and from another side native support for real-life IT application scenarios (as XML data are semi-structured by definition). Furthermore, by adopting XML as core data model, it is possible to meaningfully augment *semantics of data*, and supporting advanced KD- and IR-style functionalities. In fact, complex mining/reasoning models, like those based on graphs and trees, can be natively derived from the structure of XML data, which are intrinsically hierarchical, and used to infer knowledge via semantics, also adopting a large set of already-available algorithms and techniques coming from well-known, mature scientific disciplines like *Data Mining* (DM) and *Knowledge Discovery in Databases* (KDD).

As will be evident through the paper, even if our proposal is targeted at XML documents, it can be extended to deal with different kinds of document (e.g., HTML pages). Thus, in the rest of the paper, due to its popularity, we assume of dealing with XML documents as a relevant case of interest.