ABSTRACT

Knowledge representation and management 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 both industrial and academic research communities. 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 by first-generation P2P systems. XML-based P2P Information Systems are a more specific 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 by means of IR methods implemented on top of XML join queries against the repositories. In this chapter, we first focus our attention on the definition and the formalization of the XML-based P2P Information Systems class, also deriving interesting properties on such systems, and then we present a knowledge-representation-and-management-based framework, enriched via semantics, that allows us to efficiently process knowledge and support advanced IR techniques in XML-based P2P Information Systems, thus achieving the definition of the so-called Semantically-Augmented XML-based P2P Information Systems. Also, we complete our analytical contribution with an experimental evaluation of our framework against state-of-the-art IR techniques for P2P networks, and its theoretical analysis in comparison with other similar semantics-based proposals.
INTRODUCTION

Motivations
During the last years, there has been a growing interest for P2P Information Systems (IS$_{p2p}$) (Aberer, 2001; Aberer & Despotovic, 2001), mainly because they fit a large number of real-life IT applications. Digital libraries over P2P networks are only a significant instance of IS$_{p2p}$, but it is very easy to foresee how large the impact of IS$_{p2p}$ on innovative and emerging IT scenarios, such as e-government and e-procurement, will be in next years.

P2P networks are natively built on 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 by means of the keywords “compositions” and “Mozart”), and, due to resource sharing purposes, (ii) are usually interested in retrieving as result a set of data objects rather than a specific one. As a consequence, well-founded IR methodologies (e.g., ranking), which have already reached a significant degree of maturity, can successfully be applied in the context of P2P systems in order to improve the capabilities of these systems in retrieving useful information (i.e., knowledge), and achieve performance better than that of more traditional database-like query schemes. On the other hand, the latter schemes are quite inadequate in the absence of fixed, rigorously structured data schemas, as happens in P2P networks.

Furthermore, the consolidate IR mechanism naturally supports the self-alimenting nature of P2P systems, as in such a mechanism intermediate results can then be (re-)used to share new information, or to set and specialize new search activities. As regards schemas, from the database perspective, P2P users typically adopt a semi-structured (data) model to query data objects rather than a structured (data) model. This feature also poses unrecognized problems concerning the issue of integrating heterogeneous data sources over P2P networks. In addition to this, efficiently access data in P2P systems, which is another interesting aspect directly related to our work, is still a research challenge (Aberer et al., 2002).

Basically, P2P IR techniques extend traditional functionalities of P2P systems (i.e., file sharing primitives and simple lookup mechanisms based on partial- or exact-match of strings), by enhancing the latter via useful (and more complex) knowledge extraction features. Accomplishment of the definition and development of innovative knowledge delivery paradigms over P2P networks is the goal that underlies the idea of integrating IR techniques inside core layers of P2P networks. In fact, P2P networks meaningfully marry with the IR philosophy, thus allowing 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. Therefore, without loss of generality, we can claim that IR techniques can be effectively used to support even complex processes like knowledge representation, discovery, and management over P2P networks, being the retrieval of information in the vest of appropriate sets of data objects the basic issue to be faced-off.

Nevertheless, several characteristics of P2P networks pose important limitations to the accomplishment of this goal. Among these, we recall: (i) the completely decentralized nature of P2P networks, which enable peers and data objects to come and go at will; (ii) the absence of global or mediate schemas of data sources, which