Chapter LXXXV
A Query-Strategy-Focused Taxonomy of P2P IR Techniques

Alfredo Cuzzocrea
University of Calabria, Italy

INTRODUCTION

During the last years, there was a growing interest in peer-to-peer (P2P) systems, mainly because they fit a wide number of real-life ICT applications. Digital libraries are only a significant instance of P2P systems, but it is very easy to foresee how large the impact of P2P systems on innovative and emerging ICT scenarios, such as e-government and e-procurement, will be during the next years.

P2P networks are natively built on top of a very large repository of data objects (e.g., files) that is intrinsically distributed, fragmented, and partitioned among participant peers. P2P users are usually interested in (a) retrieving data objects containing information of interest, like video and audio files, and (b) sharing information with other (participant) users or peers. From the information retrieval (IR) perspective, P2P users (a) 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 (b), due to resource-sharing purposes, are usually interested in retrieving as a result a set of data objects rather than only one. Based on such set of items, well-founded IR methodologies like ranking can be successfully applied to improve system query capabilities, thus achieving performance better than that of more traditional database-like query schemes. Furthermore, the above-described P2P IR mechanism is self-alimenting as intermediate results can be then reused to share new information, or to set and specialize new search and query activities. In other words, from the database perspective, P2P users typically adopt a semistructured (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 to the above issues, is a relevant and still incompletely solved open research challenge.

Traditional functionalities of first-generation P2P systems are currently being extended by adding to their native capabilities (i.e., file sharing primitives and simple lookup mechanisms based on partial or exact match of search strings) useful (and more complex) knowledge representation and extraction techniques. Achieving the defini-

tion of new knowledge delivery paradigms over P2P networks is the underlying 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 (a) successfully exploit self-alimenting mechanisms of knowledge production, and (b) take advantages from innovative knowledge representation and extraction models based on semantics, metadata management, probability, and so forth. All considering, we can claim that, presently, there is a strong, effective demand for enriching P2P systems with functionalities that are proper of IS, such as knowledge discovery (KD) and IR-style data object querying, and cannot be supported by the actual data representation and query models of traditional P2P systems. More specifically, knowledge representation and management techniques mainly concern the modeling of P2P systems, whereas knowledge discovery techniques (implemented via IR functionalities) mainly concern the querying (i.e., knowledge extraction) of P2P systems.

Following this trend, a plethora of P2P IR techniques have been proposed recently, each of them focused on covering a particular or specific aspect of the KD phase. A meaningful way of studying P2P IR techniques under a common plan is looking at their query strategies used to retrieve information and knowledge. In fact, despite the implementation and architectural details, the underlying query strategy is the most relevant characteristic of any P2P IR technique, mainly from the database research perspective.

According to these considerations, in this article we provide a taxonomy of state-of-the-art P2P IR techniques, which emphasize the query strategy used to retrieve information and knowledge from peers, and put in evidence similarities and differences among the investigated techniques. This taxonomy helps us to keep track of the large number of proposals that have come up in the last years, and to support future research in this leading area.

BACKGROUND

The first experiences of P2P systems, such as Gnutella (*The Gnutella File Sharing System*, 2006), KaZaA (*The KaZaA File Sharing System*, 2006), and *Napster* (*The Napster File Sharing System*, 2006), which mainly focused on data management issues on P2P networks, were oriented toward designing techniques for which sharing data objects and generating large communities of participant peers were the most relevant goals. Under this assumption, two reference architectures have gained a leading role for P2P systems, each of them addressing two different ways of retrieving data objects by querying: unstructured P2P systems and structured P2P systems.

As regards the unstructured P2P systems, there are three main variants. In the first one (e.g., *Napster*, 2006), a centralized index storing a directory of all data objects currently available on the P2P network is located in a certain peer, whose identity is known to all the peers. When a participant peer p_i receives a request for a missing data object, it (a) performs a query against the peer containing the centralized index for retrieving the name of the (participant) peer p_j where the required data object is stored, and (b) redirects the request toward p_j. In the second variant (e.g., *Gnutella*, 2006), there is no centralized index as it can be the source of failures; each participant peer needs to maintain only information about his or her own data for supporting data object lookups, and information about neighboring peers for routing requests coming from other peers. Given such a scheme, a request for a missing data object is flooded from a peer p_i toward other peers via the neighboring peers of p_i. In the last variant (e.g., *KaZaA,* 2006), peers connect to a super-peer who builds an index over the data objects shared by his or her set of peers. In addition to this, each super-peer keeps information about neighboring super-peers in the system, and queries are routed among super-peers. Scalability is the most important drawback for unstructured P2P systems: In fact, when the number of participant peers grows, the described query mechanism can become very