Chapter 12
XP2P:
A Framework for Fragmenting and Managing XML Data over Structured Peer-to-Peer Networks

Angela Bonifati
ICAR-CNR and University of Basilicata, Italy

Alfredo Cuzzocrea
ICAR-CNR and University of Calabria, Italy

ABSTRACT

This chapter presents XP2P (XPath for P2P), a framework for fragmenting and managing XML data over structured peer-to-peer networks. XP2P is characterized by an innovative mechanism for fragmenting XML documents based on meaningful XPath queries, and novel fingerprinting techniques for indexing and looking-up distributed fragments based on Chord's DHT. Efficient algorithms for querying distributed fragments over peer-to-peer networks are also presented and experimentally assessed against both synthetic and real XML data sets. A comprehensive analysis of future research directions on XML data management over peer-to-peer networks completes the contribution of the chapter.

INTRODUCTION

XML is a data format available in the Internet and, with increasing popularity, in P2P networks. Here, the nature of XML data, which are intrinsically *semi-structured*, naturally couples with the topology and structure of P2P networks, which are usually wide and loosely connected. Numerous are the scenarios in which XML and P2P ties together such as *Knowledge P2P Management Systems and Advanced P2P Information Retrieval Systems*, to mention a few examples.

However, there are still many challenges to investigate in order to realize a full-fledged *P2P XML Data Management System*, among which query performance and support for complex XML queries are the most relevant ones. A solution to these issues could be considering new data models and storage schemes for XML data over P2P networks, along with highly-efficient algorithms for retrieving useful knowledge from large XML repositories across the network.

Inspired by these considerations, in this chapter we address the problem of storing and retrieving XML data in a DHT-based P2P network. DHTs are

DOI: 10.4018/978-1-61520-727-5.ch012

widely known because of their accuracy, logarithmic efficiency and greater scalability, as discussed by (Aberer et al., 2003). In light of this, DHTs are starting to be considered as the foundational indexes for data management applications on top of P2P networks. However, the kinds of queries so far allowed in such DHT-based architectures are mainly lookup queries, i.e. queries that return singleton items.

First, we focus on identifying the pieces of data which are of interest to a peer. Indeed, *for either space or relevance reasons*, a peer is not interested to store an XML document as a whole, but to store a subpart of it, namely a *fragment*. Space constraints are relevant for any distributed system, and become crucial for DHT-based systems, which heavily rely on load balancing. Relevance metrics should be attentively considered in a network of peers sharing a large XML document. As an example, consider a P2P data repository sharing the DB research data (may be DBLP or other DB research data, such as that employed in the *Piazza* system (Halevy et al., 2003)). It is conceivable that a DB lab peer working on "streaming" only stores locally the XML data of other DB labs working on the same topic, while still wishing to fetch data on other topics/labs whenever needed. Another example is XML biological data, such as *SwissProt* and *Protein Sequence Data Base* (PSDB), which is of interest to several biological peer databases. None of the peers is willing to locally store such large XML documents, but only to hold part of them, depending on their current interest. For instance, SwissProt contains the description of proteins and genes, their features and the papers in which they were first studied. A peer would be interested to store the genes and their citations locally, while others would be interested to store their characteristics and still keeping links to their citations. A similar behaviour would occur with the PSDB, which has a full description of each protein.

Secondly, we study how the querying mechanism is affected by the presence and availability of XML fragments. Lookup queries in the most traditional sense cannot be adopted in such a case, and need to be properly re-defined. We have devised a system, called *XP2P* (*XPath for P2P*), first presented in (Bonifati et al., 2004), to *share XML documents in a P2P environment such that the sharing is kept transparent to queries*. More precisely, we focus on *efficiently storing and retrieving XML data within an arbitrarily large DHT-based P2P network*. To this purpose, we have designed a fragmentation and replication model for global XML documents, which allows them to remain re-buildable and queryable. The whole path leading to a given tag needs to be used to identify XML data, as indeed the only tag names are not sufficient. Thus, we enable a *path-based identification mechanism* for XML fragments, which couples well with the DHT of structured P2P networks. We do not assume a global mediated schema, which would not be conceivable in a P2P setting, but instead build a decentralized catalog that relies on a few *path expressions*. Similarly to XP2P, (Galanis et al., 2003) proposes to use the tag names to build the DHT keys and guarantee efficient lookup of XML data. However, their approach is found on maintaining a global catalog of data, which is feasible for small communities of peers but unfeasible for large scalable networks.

XP2P OVERVIEW

XP2P is an extension of *Chord* (Stoica et al., 2001) that uses Rabin's fingerprints (Broder, 1993; Rabin, 1981) instead of hash keys (Stoica et al., 2001). The former ones have a remarkable software efficiency, a well-understood probability of collisions (also verified in our system) and nice algebraic properties. We refer the reader to the fourth Section for details.

One of the most relevant goals in our work has been that of handling non-conventional XML queries. In XP2P, we are able to address full *XPath 1.0* (XPath, 2006) queries on the XML data ob-