Chapter 11
Content-Based XML Data Dissemination

Guoli Li
University of Toronto, Canada

Shuang Hou
University of Toronto, Canada

Hans-Arno Jacobsen
University of Toronto, Canada

ABSTRACT

XML-based data dissemination networks are rapidly gaining momentum. In these networks XML content is routed from data producers to data consumers throughout an overlay network of content-based routers. Routing decisions are based on XPath expressions (XPEs) stored at each router. To enable efficient routing, while keeping the routing state small, we introduce advertisement-based routing algorithms for XML content, present a novel data structure for managing XPEs, especially apt for the hierarchical nature of XPEs and XML, and develop several optimizations for reducing the number of XPEs required to manage the routing state. The experimental evaluation shows that our algorithms and optimizations reduce the routing table size by up to 90%, improve the routing time by roughly 85%, and reduce overall network traffic by about 35%. Experiments running on PlanetLab show the scalability of our approach.

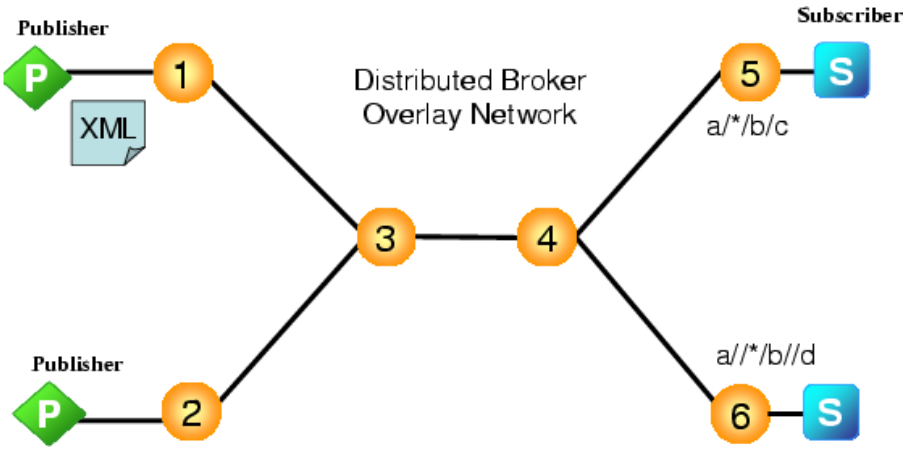
INTRODUCTION

Over the past decade, XML has rapidly evolved as the standard for data representation and exchange. XML marked-up message traffic in intranets and on the Internet ranges from insurance claims, healthcare requests, corporate memos, online ads to news items and entertainment information. The standardization of the mark-up language, the wide range of related standards, and the wide-spread adoption of this technology are further amplifying the network externalities created by this technology.

XML-based data dissemination networks are starting to become a reality. In a dissemination network, data messages, marked-up in XML, are routed based on filter expressions stored at intermediate nodes that indicate where the XML message is to be routed to. These routing nodes are often referred to as *content-based routers* or *brokers*, as they route messages to interested recipients by

DOI: 10.4018/978-1-61520-727-5.ch011

Figure 1. Distributed dissemination network



inspecting the message content. Filter expressions, often expressed as XPath expressions (XPEs), are submitted by data consumers who express interest in receiving certain kinds of documents. This architecture is depicted in Figure 1. For instance, a globally operating insurance company with many branch offices distributed world-wide is linked by an overlay network of content-based routers that comprise the XML dissemination network. An insurance claim, an insurance bid, or a request for proposal can be submitted anywhere into the overlay network (e.g., by a third party insurance broker or an online client) and be routed toward a currently online, specific expert employee, speaking the same language as the requester. Note, the latter constraints are expressed as XPE filter expressions against which the XML document is evaluated in transit. This design fully decouples information requesters and information providers, avoids a single point of control and a single point of failure, and increases scalability due to decentralization and distribution.

The fundamental concepts underlying the content-based dissemination of XML messages are the algorithms and data structures for matching and routing of XML documents against XPEs, the definition of advertisements that efficiently summarize the kind of XML documents that data producers will publish, the interpretation and the development of algorithms for intersecting advertisements with subscriptions, and the definition of covering and merging of XPEs.

This chapter addresses the XML/XPath routing problem. More specifically, this chapter focuses on the problem of efficiently routing an XML document emitted from a data producer at one point in the network to a set of data consumers located anywhere throughout the network. Prior to receiving XML documents, consumers must have expressed interest in receiving XML documents by registering XPEs with the network. This problem statement is akin to the well-known publish/subscribe matching problem. However, the main difference here is that in the case of data dissemination networks, there exists no one single centralized publish/subscribe system, but a network of content-based routers (i.e., a network or federation of publish/subscribe systems.) In the dissemination network, XML documents are routed based on their content and not based on IP address information, which is, due to the completely decoupled design, not available – all routing decisions are exclusively based on content information. Figure 2 provides an overview of the dissemination network this chapter assumes. In the overlay network depicted in Figure 2 each