FVC: A Feature-Vector-Based Classification for XML Dissemination

Xiaoling Wang[1], Ester Martin[2], Weining Qian[1], and Aoying Zhou[1]

[1] Shanghai Key Laboratory of Trustworthy Computing, Software Engineering Institute, East China Normal University, Shanghai 200062, China
{xlwang,wnqian,ayzhou}@sei.ecnu.edu.cn
[2] School of Computing Science, Simon Fraser University, Burnaby, BC, Canada
ester@cs.sfu.ca

Abstract. With the adoption of XML in a wide range of applications, efficient XML classification has become an important research topic. In current studies, users' interests are expressed by XPath or XQuery queries. However, such a query is hard to formulate, because it requires a good knowledge of the structure and contents of the documents that will arrive and some knowledge of XQuery which few consumers will have. The query may even be impossible to formulate in cases where the distinction of relevant and irrelevant documents requires the consideration of a large number of features. Traditional classification method can't work well for XML dissemination, because the number of training example is often small. Therefore, this paper introduces a data mining approach to XML dissemination that uses a given document collection of the user to automatically learn a classifier modelling his/her information needs. We present a novel XML classifier taking into account the structure as well as the content of XML documents. Our experimental evaluation on several real XML document sets demonstrates the accuracy and efficiency of the proposed XML classification approach.

1 Introduction

Nowadays, Extensible Mark-up Language (XML) becomes pervasive in more and more applications, such as Digital Library, XML subscribe/publish system, and other XML repositories. With the adoption of XML in a wide range of applications, XML classification has become an important application. There are many"news streams" that can be modeled as XML documents arriving at some server that is responsible for managing these documents and disseminating them to a pool of clients. New CS research papers (DBLP), new webpages (GoogleAlert), new movies all are examples of news streams and often use XML as data exchange format. Clients such as computer scientists, businessmen and consumers want to be alerted of relevant news documents without being overwhelmed by "spam" documents. For this purpose, clients subscribe at the server with some specification of their information needs by some kind of keyword-based XML query. Former methods have been proposed to optimize XML dissemination,

M. Yoshikawa et al. (Eds.): DASFAA 2010, LNCS 6193, pp. 398–409, 2010.

improving the efficiency of the simultaneous execution of very large numbers of XML queries by sharing the processing costs for common subqueries [1,2].

Current approaches [3,4] to XML dissemination require the clients to specify their interests as XPath or XQuery queries. Unfortunately, XML queries are hard to formulate for consumers, because they require a good knowledge of the structure and contents of the documents of the news stream. In the case of XML documents, it would also require familiarity with the XQuery language which few clients will have.

On the other hand, in many scenarios a client would have an initial collection of relevant documents obtained from other sources, e.g. a collection of related research papers, a collection of favorite bookmarks or a collection of owned movies. These document collections do implicitly specify the information needs of the client which motivates a data mining approach to XML dissemination based on classifiers that are learned automatically from the given collections of relevant documents:

1. A training collection of "positive" (relevant) XML documents on each client describes the client's (consumer's) interest. Using a sample of other documents as "negative" training documents, a two-class classifier (relevant / irrelevant) is trained on the client.
2. A client subscribes to the server by sending his classifier to the server.
3. The server applies the classifiers of all clients that have subscribed to the incoming XML documents and forwards a document to all clients whose classifier predicts the document as relevant.
4. The client (consumer) receives a potentially relevant XML document, checks the document for actual relevance and records relevance feedback (is / is not relevant).
5. The client maintains and updates the classifier. If too many documents received from the server are judged as irrelevant by the consumer, according to the client relevance feedback, the client re-trains the classifier on the current document collection and sends a new classifier to the server.

The architecture of a data mining-based XML dissemination system is shown in Figure. 1.

Classification methods for XML documents have recently received some attention in the database and data mining communities [6,9]. The information dissemination scenario, however, creates unique challenges for XML classification that have not yet been addressed in the literature. Compared to the classical scenario of text and XML classification, training datasets are small, i.e. contain only a few hundreds or dozens of documents. This makes classifier construction much more difficult. For example, SVM approach needs more training examples in order to achieve high precision.

To deal with small training sets, we introduce a novel classification method that represents both the structure and the contents of XML documents in a natural way. Our method integrates the XRules approach [9], that takes into account the XML structure only, and the Bag of Words text classification approach [6], that exploits only the contents of XML documents.