Chapter II
Engineering Information into Open Documents

Chia-Chu Chiang
University of Arkansas at Little Rock, USA

ABSTRACT

Documents are perfectly suited for information exchange via the Internet. In order to insure that there are no misunderstandings, information embedded in a document needs to be precise and unambiguous. Having a (de facto) standard data model and conceptual information model insures that the involved parties will agree on what the information means. XML (eXtensible Markup Language) has become the de facto standard format for representing information in documents for document exchange. Many techniques have been proposed to create XML documents, including the validation and transformation of XML documents. However, very little is discussed when it comes to extracting information from non-XML documents and engineering the information into XML documents. The extraction process can be a highly labor intensive task if it is done manually. The use of automated tools would make the process more efficient. In this chapter, the author will briefly survey document engineering techniques for XML documents. Then, the author will present two techniques to extract data from Windows documents into XML documents. These two techniques have been successfully applied in two industrial projects. He believes that techniques that automate the extraction of data from non-XML documents into XML formats will definitely enhance the use of XML documents.

INTRODUCTION

Due to the availability of electronic devices and networking systems today, the use of information has shifted mainly to the use of computer-based information systems to collect, store, process, and retrieve information from the Internet. Organizations including companies, governments, and individuals

mainly conduct business functions through the exchanging of documents that carry information via the Internet. A document can therefore be viewed as an electronic file created by a computer application that contains meaningful data.

We can agree that nowadays, we are overwhelmed by the huge volume of information in the documents created every day. Unfortunately, the information stored in the documents is not always represented in the same format; currently, there exists a multitude of incompatible document types. Thus, we are required to use different software tools to access different document types for the information we need. The root of this problem lies in the fact that documents and their corresponding tools have inseparable relationships where each tool creates different representations of documents, and many of the representations are proprietary. Not only do incompatible document types create a problem when accessing documents, but incompatible conceptual data models also describe the problem of aggregating and comparing information in different documents even when they are of the same type. Both the document type and conceptual data model must be defined by the organizations if they wish to agree on the meaning of their documents.

In this chapter, we are particularly interested in techniques that can engineer information into open documents in XML. When we use the term "open documents", we are referring to the documents that are created using a de facto standard document type, such as XML. Although there exists more than one standard for document types, we are not resolving the issues of incompatible standards in this chapter. Existing techniques for processing XML documents will be discussed. Several conferences such as ACM, SIGMOD, and DocEng provide excellent resources for the techniques that will be discussed. A survey (Forward & Lethbridge, 2002) presents the relevance of documentation, tools, and technologies. Afterwards, we will present two techniques used to extract information from non-XML documents into XML documents. Finally, we discuss issues and future trends and conclude the chapter.

The chapter is organized as follows: The background section briefly introduces XML syntax and semantics. This section also presents engineering techniques for XML documents. Some techniques support the syntax and semantics of XML documents. Some techniques target the transformations of XML documents into other document types. In the next section, we will present two techniques that emphasize the creation of XML documents from non-XML documents. We observe that many commercial companies have provided software tools to automate this process. Unfortunately, the technical details of these tools are usually not released to the public. The techniques presented in this section will allow users to convert entire documents to XML and also allow users to extract data of interest to XML. The future trends section will discuss the issues of document engineering and the potential solutions. Finally, the chapter is summarized in the conclusion section.

BACKGROUND

Glushko and McGrath (2005) define documents in a general notion as follows,

"Document in a technology-neutral way as a purposeful and self-contained collection of information."

Organizations should think of documents in an abstract and technology-neutral way. Documents should be flexibly exchanged via the Internet without concern as to how the documents are to be sent