Chapter XIV

XML Benchmarking:
The State of the Art and Possible Enhancements

Irena Mlynkova
Charles University, Czech Republic

ABSTRACT

Since XML technologies have become a standard for data representation, numerous methods for processing XML data emerge every day. Consequently, it is necessary to compare the newly proposed methods with the existing ones, as well as analyze the effect of a particular method when applied to various types of data. In this chapter, the auhtors provide an overview of existing approaches to XML benchmarking from the perspective of various applications and show that to date the problem has been highly marginalized. Therefore, in the second part of the chapter they discuss persisting open issues and their possible solutions.

INTRODUCTION

Since XML (Bray et al., 2006) became a de-facto standard for data representation and manipulation, numerous methods have been proposed for efficiently managing, processing, exchanging, querying, updating and compressing XML documents. And new proposals emerge every day. Naturally, each author performs various experimental tests using the newly proposed method and describes its advantages and disadvantages. But, it can be very difficult for a future user to decide which of the existing approaches is the most suitable for his/her particular requirements on the basis of the descriptions of methods. The problem is that various methods are usually tested on different data sets derived from diverse sources which either do not yet exist or which were created only for
the testing purposes, with special requirements of particular applications etc.

An author of a new method will encounter a similar problem whenever he/she wants to compare the new proposal with an existing one. This is possible only if the source or executable files of the existing method or, at least, identical testing data sets are available. But, too often it is impossible to have access to this information. In addition, in the latter case, the performance evaluation is limited by the testing set whose characteristics are often unknown. Hence, a reader finds it difficult to obtain a clear notion of the analyzed situation.

An analogous problem occurs if we want to test the behaviour of a particular method on various types of data, or determine the correlation between the efficiency of the method and changing complexity of the input data. Not even the process of gathering the testing data sets is simple. Firstly, the real-world XML data usually contain a huge number of errors (Mlynkova et al., 2006) which need to be corrected. And what is worse, the real-world data sets are usually surprisingly simple and do not cover all constructs allowed by XML specifications.

Currently, there exist several projects which provide a set of testing XML data collections (usually together with a set of testing XML operations) that are publicly available and well-described. We can find either fixed (or gradually extended) databases of real-world XML data (e.g. project INEX (INEX, 2007)) or projects which enable us to generate synthetic XML data on the basis of user-specified characteristics (e.g. project XMark (Busse, 2003)). But, in the former case, we are limited by the characteristics of the testing set; whereas, in the latter case, the characteristics of the generated data that can be specified are trivial (such as the amount and size of the data).

Goals of the Chapter

The first aim of this chapter is to provide an overview of existing XML benchmarking projects, i.e. projects which provide a set of testing XML data collections, XML operations/scenarios etc. We will discuss their main characteristics and in particular the issues related to their versatility. We will show that the problem of sophisticated XML benchmarking has been so far highly marginalized and the number of possibilities for acquiring at least a reasonable testing set of XML data is surprisingly low.

Since the key operations of XML processing are undoubtedly parsing, validating and querying, most of the existing benchmarking projects focus mainly on them. But, there are also other popular and useful XML technologies or operations with XML data and, hence, there also exist benchmarks determined for other purposes. Nevertheless, their number is surprisingly low or the existing representatives are already obsolete.

The next aim of the chapter is to identify the most significant related open issues and unsolved problems. In particular, we will deal with a system which is able to generate synthetic XML data on the basis of a wide range of user-specified characteristics. We will focus on three aspects of the problem—automatic generation of synthetic XML documents, automatic generation of their XML schema, and automatic generation of respective XML queries. The main idea is that the author of a new method will be able to test its behaviour on any kind of data that can be described using this set of characteristics. On the other hand, any other author can use the same setting-up of characteristics, repeat the generation of the testing data sets and compare a new method with the existing results.

In general, we will describe and discuss such a system in full and focus on the related open problems as well as possible solutions. The particular implementation can then focus only on selected aspects appropriate for concrete exploitation.