ToXgene: A template-based data generator for XML

Denilson Barbosa Alberto Mendelzon
Department of Computer Science
University of Toronto
{dmb,mendel}@db.toronto.edu

John Keenleyside Kelly Lyons
IBM Toronto Lab
{keenley,klyons}@ca.ibm.com

Synthetic collections of XML documents can be useful in many applications, such as benchmarking (e.g., Xmark [4], XOO7 [2]) and algorithm testing and evaluation. We present ToXgene, a template-based tool for facilitating the generation of large, consistent collections of synthetic XML documents.

ToXgene was designed with the following requirements in mind: it should be declarative, to speed the data generation up; it should be general enough to generate fairly complex XML content and it should be powerful enough to capture the most common kinds of constraints in popular benchmarks. Preliminary experimental results show that our tool can closely reproduce the data sets for the Xmark and the TPC-H benchmarks [6].

The ToXgene Template Specification Language (TSL) is a subset of the XML Schema notation augmented with annotations for specifying certain properties of the intended data, such as probability distributions, the vocabulary for `CDATA` content, etc. We use XML Schema as the basis for TSL not only because it is a W3C standard, but also because it provides a more detailed description of XML documents than DTDs; in particular, it allows the specification of datatypes. We note that our tool gives the user total control over the data to be generated; thus, it is intended for the cases when the user *knows* the structure of the data she wants and *requires* the data to conform to this structure (however, we note that the structure does *not* have to be regular). The main features of our tool are:

Generation of complex XML content: our tool supports all XML element content models (`CDATA`, element and mixed) and allows the generation of attributes as well. `CDATA` values are generated according to a type declaration; various string, numeric and date types are supported.

Skewed probability distributions: the user can specify probability distributions to determine the number of occurrences for elements, as well as to control the generation of `CDATA` literals (e.g., the length of string values). ToXgene supports the uniform, exponential, normal, log-normal, geometric, and arbitrary discrete distributions.

Element sharing: ToXgene allows different elements (or attributes) to share `CDATA` literals, thus allowing the generation of references among elements in the same (or in different) documents. This enables the generation of collections of correlated documents (i.e., documents that can be joined by value).

Integrity constraints: element sharing in ToXgene is achieved by generating the shared content prior to the actual documents, and storing this data in what we call tox-lists. Our tool allows the specification of most common integrity constraints (e.g., uniqueness) over the data in such lists; thus, one can generate consistent `ID`, `IDREF` and `IDREFS` attributes. One can also specify integrity constraints over elements (or attributes) in different documents, which allows the generation of consistent single or multiple document data sets.

Reuse of existing data: our tool allows the user to load existing data into tox-lists; such data is treated as any other shared data in the generation process. This allows the mixing of real and synthetic data, often required in common benchmarks (e.g., names of countries) and also the growing of existing collections of documents without having to start from scratch again.

Extensibility: ToXgene was developed in Java 2, and has very simple interfaces for plugging in new CDATA generators (e.g., DNA sequence data), and new random number generators, according to other probability distributions.

ToXgene is part of ToX – the Toronto XML Engine – project [5]; more information on ToXgene, including a comparison of our tool with other XML generators ([3, 1]) can be found in [6].

ACM SIGMOD '2002 June 4-6, Madison, Wisconsin, USA

1. REFERENCES

[1] A. Aboulnaga, J. Naughton and C. Zhang. Generating Synthetic Complex-strutcured XML Data. *WebDB'00.*
[2] S. Bressan et al. XOO7: Applying OO7 Benchmark to XML Query Processing Tools. *CIKM'01.*
[3] IBM XML Generator. `http://www.alphaworks.ibm.com/tech/xmlgenerator`.
[4] The XML Benchmark Project. `http://monetdb.cwi.nl/xml/`.
[5] ToX - the Toronto XML Engine project. `http://www.cs.toronto.edu/tox`.
[6] ToXgene. `http://www.cs.toronto.edu/tox/toxgene`.