Chapter X

Issues in Personalized Access to Multi-Version XML Documents

Fabio Grandi
Università di Bologna, Italy

Federica Mandreoli
Università di Modena e Reggio Emilia, Italy

Riccardo Martoglia
Università di Modena e Reggio Emilia, Italy

ABSTRACT

In several application fields including legal and medical domains, XML documents are “versioned” along different dimensions of interest, whose nature depends on the application needs such as time, space and security. Specifically, temporal and Semantic versioning is particularly demanding in a broad range of application domains where temporal versioning can be used to maintain histories of the underlying resources along various time dimensions, and Semantic versioning can then be used to model limited applicability of resources to individual cases or contexts. The selection and reconstruction of the version(s) of interest for a user means the retrieval of those fragments of documents that match both the implicit and explicit user needs, which can be formalized as what we call personalization queries.

In this chapter, the authors focus on the design and implementation issues of a personalization query processor. They consider different design options and, among them, they introduce an in-depth study of a native solution by showing, also through experimental evaluation, how some of the best performing technological solutions available today for XML data management can be successfully extended and optimally combined in order to support personalization queries.

OVERVIEW AND MOTIVATION

Nowadays, XML has become ubiquitous with an ever-increasing number of computer applications exchanging and storing information in XML format. In particular, a large number of organizations, including private companies and public institutions, place rich collections of documents...
Issues in Personalized Access to Multi-Version XML Documents

at the disposal of Internet users. Generally, such collections are large XML repositories containing millions of semi-structured documents, each one containing thousands of nodes. Portals and Websites which allow users to access such repositories are usually equipped with classic keyword-based search engines which are not adequate to retrieve all and only the information that is relevant for the user, as the tree structure of documents must also be taken into account. As a consequence, in recent years many research efforts have been expended to support structural querying in XML repositories and discovering the occurrences of labelled trees—or twig query—patterns (Amer-Yahia et al., 2001) has become a core operation for XML query processing.

Moreover, in several application fields including legal and medical domains, management of bills of materials and catalogue data, accounting and finance, XML documents are “versioned” along different dimensions of interest, whose nature depends on the application needs (e.g. time, space, security). In this chapter, we consider time pertinence and applicability as versioning dimensions, which give rise to multidimensional temporal and Semantic versioning. Indeed, temporal and Semantic versioning is particularly demanding in a broad range of application domains where temporal versioning can be used to maintain histories of the underlying resources along various time dimensions, and Semantic versioning can then be used to model limited applicability of resources (or resource portions) to individual cases or contexts. In all these cases, while the most important version is the “current” one with respect to the temporal dimensions (and with generic applicability), past versions are also very important for applications and cannot be discarded.

For instance, in the legal domain, a clear example of such multi-version resources are norm texts, including laws, acts, decrees, provisions, regulations, etc. Norm texts are continually subject to amendments and modifications and multiple temporal versions coexist as a consequence of the dynamics of the legislative activity. In particular, several temporal dimensions are involved in the representation and management of norm texts, including transaction, validity, efficacy, applicability, publication and enactment times (Grandi et al., 2005; Palmirani & Brighi, 2006). The most important version of a norm is the consolidated version, which is the one produced by the application of all the modifications the norm has undergone so far, as it is the one which is currently part of the regulations in force and generically applicable to all citizens. However, past versions (even with limited applicability) are also virtually needed. For instance, considering validity time and Semantic applicability to individual cases, a court might be called to judge a case involving a crime C committed at a time T on the basis of the (versions of the) laws which were valid at time T and applicable to crime C.

Another interesting example is the medical domain, where multi-version resources of interest are, for instance, clinical guidelines, which are definitions of “best practices” encoding and standardizing clinical procedures for a given disease. Clinical guidelines are also subject to continuous development and revision by committees of expert physicians and health authorities, and multiple temporal versions coexist as a consequence of the clinical and healthcare activity. Several temporal dimensions are also involved in the representation and management of clinical guidelines, including valid, transaction, event, availability, proposal and acceptance times (Combi & Montanari, 2001; Terenziani et al., 2005). Also, in the medical domain, past versions continue to be relevant, as a physician might be called upon to justify his/her actions for a given patient P at a time T on the basis on the (versions of the) clinical guidelines which were valid at time T and applicable to the pathology of patient P.

In addition to individual cases (e.g. citizens or diseases), Semantic versioning can also involve more generic applicability contexts (e.g. districts