Abstract

In this paper, we present the results of an ongoing research involving the design and implementation, in an eGovernment scenario, of a semantic-aware system supporting efficient and personalized access to a multi-version repository of normative texts. The research activity is entitled “Semantic web techniques for the management of digital identity and the access to norms”. In the context of a complete and modular infrastructure, we defined a multi-version XML data model and developed a temporal and semantical XML query processor supporting both temporal versioning –essential in normative systems– and semantic versioning. Semantic versioning is based on the applicability of different norm parts to different classes of citizens and allows users to retrieve personalized norm versions only containing provisions which are applicable to their personal case. The whole infrastructure, which we plan to complete in the near future, will integrate the querying component with several auxiliary services, including automatic citizen identification and classification and assisted update of the repository data.

Introduction

Nowadays we are witnessing a strong institutional push towards the implementation of eGovernment support services, aimed at a higher level of integration and involvement of the citizens in the Public Administration (PA) activities that concern them. In this framework, collections of norm texts and legal information presented to citizens are made available in XML format (XML 2005) and are becoming popular on the Internet. The offering of personalized versions is aimed at improving and optimizing the involvement of citizens in the eGovernance process. In existing systems, personalization is either absent (e.g. (NIR 2005)) or predefined by human experts and hardwired in the repository structure (e.g. (EGov.it 2005)), whereas flexible and on-demand personalization services are lacking.

In this challenging scenario, takes its place the research activity entitled “Semantic web techniques for the management of digital identity and the access to norms”, which we are carrying out as part of the PRIN Italian project “European Citizen in eGovernance: legal-philosophical, legal, computer science and economical aspects” (PRIN Project 2003). One of the main objectives of such activity is the development of a system allowing an effective and efficient access to multi-version norm repositories supporting temporal queries and personalization facilities. Personalization is based on possible limitations in the applicability of norm provisions and takes into account the identity of the user as a citizen subject to norms. With the term “norm” we mean documents that articulate some law or policy. Indeed, the fast dynamics involved in normative systems implies the coexistence of multiple temporal versions of the normative texts stored in a repository, since laws are continually subject to amendments and modifications (for instance, it is crucial to reconstruct the consolidated version of a norm as produced by the application of all the modifications it underwent so far). Moreover, another kind of versioning plays an important role, because some norms or some of their parts have or acquire a limited applicability. For example, a given norm may contain some articles which are only applicable to particular classes of citizens (e.g. public employees). Hence, a citizen accessing the repository may be interested in finding a personalized version of the norm, that is a version only containing articles which are applicable to his/her personal case.

Fig. 1 shows the complete infrastructure, along with its different components, we envisioned in order to achieve the above mentioned goals (Grandi et al. 2004). Firstly, in order to obtain a personalized access, a secure authentication is required for a citizen accessing the infrastructure. This is performed through a simple elaboration unit, also acting as user interface, which processes the citizen’s requests and manages the results. Then, in order to enhance the participation of the citizens to an eGovernance procedure of interest, their automatic and accurate positioning within the reference legal framework is needed. To solve this problem we employ Semantic Web techniques and introduce a civic ontology ($O_C$ in Fig. 1), which corresponds to a classification of citizens based on the distinctions introduced by subsequent norms which imply some limitation (total or partial) in their applicability. In the following, we refer to such norms as founding acts. Moreover, we define the citizen’s digital identity as the total amount of information concerning him/her –necessary for the sake of classification with respect to the ontology– which is available online (Rodota 2001). Such information must be retrievable in an automatic, se-
For the specification of the identification, classification and creation/update services, we plan to adopt a standard declarative formalism (e.g. based on XML/SOAP (SOAP 2005)).

This section introduces the current achievements of our research activity and presents in detail the complete infrastructure of our system. The section which follows investigates the aspects of selective and personalized access to multi-version documents. The section coming next is devoted to the implementation and assessment of a prototype of the document repository and query processor. Finally, we conclude the paper with a discussion on achievements and future work.

**Personalized access to norms**

Up to now, our research has been mainly focused on the querying phase. In particular, we defined an XML data model which combines semantic annotations with temporal versioning and we defined efficient techniques for querying repositories storing legal documents supporting temporal and semantic versioning.

Temporal concerns are widespread in the eGovernment domain and a legal information system should be able to retrieve or reconstruct on demand any version of a given document to meet common application requirements. In fact, whereas it is crucial to reconstruct the current (consolidated) version of a norm as it is the one that currently belongs to the regulations and must be enforced today, also past versions are still important, not only for historical reasons. For example, if a Court has to pass judgment today on some fact committed in the past, the version of norms which must be applied to the case is the one that was in force then. We then extend the temporal framework with semantic versioning in order to provide personalized access to normative texts. Semantic versioning also plays an important role, due to the limited applicability that norms or some of their parts have or acquire. Hence, it is crucial to maintain the mapping between each portion of a norm and the maximal class
Figure 2: An example of civic ontology, where each class has a name and is associated to a (pre,post) pair, and a fragment of an XML norm containing applicability annotations.

of citizens it applies to in order to support an effective personalization service. The norms are translated in our multiversion XML data model through a semi-automated process involving a human expert by means of an “intelligent” interactive editor, to be used for marking-up norms with temporal and semantic attributes and recording the new laws in legal databases (Palmirani & Brighi 2002). Temporal and semantic versioning are detailed in the next subsections.

**Temporal Versioning**

We first focused on the temporal aspects and on the effective and efficient management of time-varying normative texts. Our work on these aspects is based on our previous research experiences (Grandi et al. 2003; Grandi, Mandreoli, & Tiberio 2005). To this purpose, we developed a temporal XML data model which uses four time dimensions to correctly represent the evolution of norms in time and their resulting versioning. The considered dimensions are:

**Validity time.** It is the time the norm is in force. It has the same semantics of valid time as in temporal databases (Jensen, Dyreson, & others 1998), since it represents the time the norm actually belongs to the regulations in the real world.

**Efficacy time.** It is the time the norm can be applied to concrete cases. While such cases do exist, the norm continues its efficacy even if no longer in force. It also has a semantics of valid time although it is independent from validity time.

**Transaction time.** It is the time the norm is stored in a computer system. It has the same semantics of transaction time as in temporal databases (Jensen, Dyreson, & others 1998).

**Publication time.** It is the time of publication of the norm on the Official Journal. It has the same semantics as event time in temporal databases (Kim & Chakravarthy 1993). As a global and unchangeable norm property, it is not used as a versioning dimension.

The data model was defined via an XML Schema (XML Schema 2005), where the structure of norms is defined by means of a contents-section-article-paragraph hierarchy and multiple content versions can be defined at each level of the hierarchy. Each version is characterized by timestamp attributes defining its temporal pertinence with respect to each of the validity, efficacy and transaction time dimensions. The model is also equipped with two basic operators for the management of norm modifications: one is devoted to change the textual content of a norm portion and the other allows modifications to the temporal pertinence of a given version. The former can be used for deletion of (a part of) the norm (abrogation), or the introduction of a new part of the norm (integration), or the replacement of (a part of) the norm (substitution). The latter can be used to deal with the time extension or the suspension of (part of) the norm. Details can be found in (Grandi et al. 2003; Grandi, Mandreoli, & Tiberio 2005).

Legal text repositories are usually managed by traditional information retrieval systems where users are allowed to access their contents by means of keyword-based queries expressing the subjects they are interested in. Advanced systems (e.g. Norma-system (Palmirani & Brighi 2002)) allow users to reconstruct a consistent version of the retrieved norm with respect to validity (consolidated act). We extended such a framework by offering users the possibility of expressing temporal specifications for the reconstruction of a consistent version of the retrieved normative acts with respect to all time dimensions.

Starting from a first implementation of our temporal model, which is described in (Grandi et al. 2003; Grandi, Mandreoli, & Tiberio 2005), we deeply redesigned the overall system architecture, the document storage scheme and the query processing methods in order to improve efficiency. The redesign also took into account the new problems arising from the extension to support semantic versioning. The new system organization, introduced in (Grandi et al. 2005c; 2005a), will be described in the next section, whereas a detailed comparison between the two architectures can be
The tree-like civic ontology is sufficient to satisfy relationships. The tree-like civic ontology is based on a taxonomy induced by IS-A relationships. For instance, the whole article in the Figure is applicable to civic class “3” (attribute applies to) and by default to all its descendants. However, its first paragraph is applicable to class “4”, which is a restriction, whereas the second one is applicable to class “8” (attribute applies also), which is an extension. The reconstruction of pertinent versions of the norm based on its applicability annotations is very important in an eGovernment scenario. The representation of extensions and restrictions gives rise to high expressiveness and flexibility in such a context.

Accessing the right version for personalization

The queries that can be submitted to the system can contain four types of constraints: temporal, structural, textual and applicability. Such constraints are completely orthogonal and allow users to perform very specific searches in the XML norm repository. Let us focus first on the applicability constraint. Consider again the ontology and norm fragment in Fig. 2 and let John Smith be a “self-employed” citizen (i.e. belonging to class “7”) retrieving the norm: hence, the sample article in the Figure will be selected as pertinent, but only the second paragraph will be actually presented as applicable. Furthermore, the applicability constraint can be combined with the other three ones in order to fully support a multi-dimensional selection. For instance, John Smith could be interested in all the norms ...

- which contain paragraphs (structural constraint) dealing with health care (textual constraint), ...
- which were valid and in effect between 2002 and 2004 (temporal constraint), ...
- which are applicable to his personal case (applicability constraint).

Such a query can be issued to our system using the standard XQuery FLWR syntax (XQuery 2005) displayed in Fig. 3, where textConstr, tempConstr, and applConstr are suitable functions allowing the specification of the textual, temporal and applicability constraints, respectively (the structural constraint is implicit in the XPath expressions used in the XQuery statement). Notice that the temporal constraints can involve all the four available time dimensions (publication, validity, efficacy and transaction), allowing high flexibility in satisfying the information needs of users in the eGovernment scenario. In particular, by means of validity and efficacy time constraints, a user is able to extract consolidated current versions from the multi-version repository, or to access past versions of particular normative texts, all consistently reconstructed by the system on the basis of the user’s requirements and personalized on the basis of his/her identity.

Implementation and performance evaluation

In order to test both the efficacy of the approach and the efficiency of the proposed technique, we built a prototype system implementing the data model. The system exploits

```xml
FOR $a IN norm
WHERE textConstr ($a//@paragraph/text(), 'health AND care')
AND tempConstr ('vTime OVERLAPS PERIOD('2002-01-01','2004-12-31')')
AND tempConstr ('eTime OVERLAPS PERIOD('2002-01-01','2004-12-31')')
AND applConstr ('class.7')
RETURN $a
```

Figure 3: An XQuery-equivalent query executable on our second system.
an “XML-native” architecture, as it is composed of a Multi-
version XML Query Processor designed on purpose, which
is able to manage the XML data repository and to support
all the temporal, structural, textual, and applicability query
facilities in a single component.

The prototype is implemented in Java JDK 1.5 and ex-
ecutes ad-hoc data structures (relying on embedded “light”
DBMS libraries) and algorithms which allow users to store
and reconstruct on-the-fly the XML norm versions satisfying
the four types of constraints. Such a component stores the
XML norms not as entire documents but decomposed into a
collection of ad-hoc temporal tuples, each representing one
of their multi-version parts (i.e. paragraphs, articles, and
so on); these data structures are then exploited to efficiently
perform structural join algorithms (Al-Khalifa et al. 2002)
we specifically devised and tuned for the temporal/semantic
multi-version context. A detailed presentation of the pro-
duced data structures and structural join algorithms, together
with a related work discussion on these topics, is given in
(Mandreoli, Martoglia, & Ronchetti 2006). Textual con-
straints are handled by means of an inverted index. The im-
provement with respect to our first temporal prototype are
manifest:

- by querying ad-hoc and temporally-enhanced structures
  (which have a finer granularity than the entire documents
  managed by standard XML engines), our system is able
to access and retrieve only the strictly necessary data;
- only the parts which are required and which satisfy the
temporal constraints are used for the reconstruction of the
retrieved documents;
- there is no need to retrieve whole XML documents and
build space-consuming structures such as DOM trees to
process a query.

Furthermore, the new architecture also provides support
to personalized access by fast handling of the applicability
constraints. Owing to the properties of the adopted pre- and
post-order encoding of the civic classes, the system is able
to very efficiently deal with applicability constraints during
query processing by means of simple comparisons involving
such encodings and semantic annotations.

As a consequence, we expected a high overall query pro-
cessing efficiency together with low memory requirements.

In order to evaluate the performance of our system, we
built a specific query benchmark and conducted a number
of exploratory experiments to test its behavior under differ-
ent workloads. The experiments have been effected on a
Pentium 4 2.5Ghz Windows XP Professional workstation,
equipped with 512MB RAM and a RAID0 cluster of two
80GB EIDE disks with NT file system (NTFS). We per-
formed the tests on three XML document sets of increasing
size: collection C1 (5,000 XML normative text documents),
C2 (10,000 documents) and C3 (20,000 documents). In this
section we will present in detail the results obtained on the
collection C1, then we will briefly describe the scalability
performance shown on the other two collections. The to-
total size of the collections is 120MB, 240MB, and 480MB,
respectively. In all collections the documents were syn-
thetically generated by means of an ad-hoc XML genera-
tor we developed, which is able to produce different docu-
ments compliant to our multi-version model. For each col-
lection, the average, minimum and maximum document size
is 24KB, 2KB and 125KB, respectively.

Experiments were conducted by submitting queries of five
different types (Q1-Q5). Table 1 presents the features of
the test queries and the query execution time for each of
them. All the queries require structural support (St con-
straint); types Q1 and Q2 also involve textual search by key-
words (Tx constraint), with different selectivities; type Q3
contains temporal conditions (Tm constraint) on three time
dimensions: transaction, valid and publication time; types
Q4 and Q5 mix the previous ones since they involve both
keywords and temporal conditions. For each query type, we
also present a personalized access variant involving an ad-
ditional applicability constraint (Ap constraint), denoted as
Qx-A in the first column of Table 1.

Let us first focus on the “standard” queries. Our ap-
proach shows a good efficiency in every context, provid-
ing a short response time (including query analysis, retrieval
of the qualifying norm parts and reconstruction of the re-
sults) of approximately one or two seconds for most of the
queries. Notice that the selectivity of the query predicates
does not impair performances, even when large amounts of
documents containing some (typically small) relevant por-
tions have to be retrieved, as it happens for queries Q2 and
Q3. Our new system is able to deliver a fast and re-
liable performance in all cases, since it practically avoids

<table>
<thead>
<tr>
<th>Query</th>
<th>Constraints</th>
<th>Selectivity</th>
<th>Performance (msec)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tm</td>
<td>St</td>
<td>Tx</td>
</tr>
<tr>
<td>Q1-A</td>
<td>-</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Q2-A</td>
<td>-</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Q3-A</td>
<td>√</td>
<td>√</td>
<td>-</td>
</tr>
<tr>
<td>Q4-A</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Q5-A</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Q1</td>
<td>-</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Q2</td>
<td>-</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Q3</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Q4</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Q5</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
</tbody>
</table>

Table 1: Features of the test queries and query execution time (time in milliseconds, collection C1)
the retrieval of useless document parts. Furthermore, consider that, for the same reasons, the main memory requirements of the Multi-version XML Query Processor are very small, less than 5% with respect to “DOM-based” approaches such as the one adopted in (Grandi et al. 2003; Grandi, Mandreoli, & Tiberio 2005). Notice that this property is also very promising towards future extensions to cope with concurrent multi-user query processing, since memory requirements are not crucial for performance in the “XML-native” approach.

The time needed to answer the personalized access versions of the Q1–Q5 queries is approximately 0.5-1% more than for the original versions. Moreover, since the applicability annotations of each part of an XML document are stored as simple integers, the size of the tuples with applicability annotations is practically unchanged (only a 3-4% storage space overhead is required with respect to documents without semantic versioning), even with quite complex annotations involving several applicability extensions and restrictions.

Finally, we only make here a comment about the performance of our current prototype in querying the other two collections C2 and C3 and, therefore, concerning the scalability of the system. We ran the same queries of the previous tests on the larger collections and saw that the computing time always grows sub-linearly with the number of documents. For instance, query Q1 executed on the 10,000 documents of collection C2 (which is as double as C1) took 1,366 msec (i.e. the system was only 30% slower); similarly, on the 20,000 documents of collection C3, the average response time was 1,741 msec (i.e. the system was less than 30% slower than with C2). Also with the other queries the measured trend was the same, thus showing the good scalability of the system in every type of query context.

Conclusions and future work

In this paper, we presented the current achievements of an ongoing research activity we are carrying out in the context of a national research project in order to support efficient and personalized access to multi-version XML document repositories for eGovernment applications. We defined a data model supporting both temporal versioning and personalized access, built a prototype system implementing the data model and evaluated its performance through some exploratory experiments. Up-to-date results are very encouraging as to query response time, storage requirements and system scalability.

Our current research work is devoted to the extensions of this framework and to the development of the complete technological infrastructure enabling our approach to be self-contained and usable in a large Web-based eGovernment scenario. In particular, we plan to complete the infrastructure with the specification and implementation of the remaining auxiliary services described in the Introduction. Moreover, the adoption of a tree-like civic ontology –that is based on a taxonomy induced by the IS-A relationship– is sufficient to satisfy basic application requirements as far as applicability constraints and personalization services are concerned. However, more advanced application requirements include a more sophisticated ontology definition. As a matter of fact, we are currently working in extending the framework in order to fully support generic graph-like ontologies, containing, for instance, multiple interconnected IS-A taxonomies, PART-OF and equivalence relations between the classes. To this aim, the XML documents annotation scheme and their storage organization need to be enhanced. Further, we are also considering the use of new ad-hoc indexes on the semantic attributes to enhance the efficiency of the query processing algorithms even more for highly-selective applicability queries.

On the other hand, further work will also include the assessment of our developed systems in a concrete working environment, with real users and in the presence of a large repository of real legal documents. In particular, a civic ontology based on a corpus of real norms (concerning infancy schools) is currently under development.

Finally, as the reference ontology evolves in time due to modifications of the founding acts, we are trying to provide the ontology with a temporal versioning mechanism similar to the one we employed for XML documents. Furthermore, we also plan to extend our approach to a broader application range, by considering ontology-based personalization of generic Web resources.

Acknowledgment

This work has been supported by the MIUR-PRIN Project: “The European citizen in e-Governance: philosophical-juridical, legal, information and economic profiles” (PRIN Project 2003).

References


