Report on PhD Thesis Work Progress

Marco Mesiti

November 2001

In this report we discuss how the work described in the PhD Thesis Proposal [15] has been carried out during last year, and how we plan to develop the work still to be done in the next year. The PhD Thesis Proposal has been revised for what concern task 6: developing query optimization strategies for XML. This task was part of a Telcordia Technologies project. However, due to general economical restrictions, the project has been canceled as well as the period of staying at the Telcordia Technologies Labs planned for last summer.

We recall that the main tasks of that research plan are:

1. Developing a formal definition for XML.
2. Developing a similarity-based approach for classifying XML documents.
3. Developing a similarity-based approach for querying XML documents.
4. Developing an access control mechanism for XML documents.

The report is organized as follows. Section 1 reports the current status of research. Section 2 presents a workplan for the last year of the thesis work. Section 3 proposes a possible structure for the thesis.

1 Current Status of Research

The purpose of this section is to discuss the work already developed with respect to the directions presented in the Thesis Proposal. The state of the research for specific topics is summarized in Figure 1. The state of the research quantifies the amount of research already done, with a number between 0 and 3. For each topic, a reference is given to the papers and reports presenting the results we have obtained so far. In particular, the current status of research can be summarized as follows.

1.1 Formalization of XML Documents

The formal definition of XML document and DTD has been presented in [4, 7]. Such definitions still lack of concepts of external links. However, the integration of such notion will be easy and we plan to complete it during the final writing of the thesis.

1.2 Classification of XML Documents

Starting from some results in the area of classification of semi-structured objects [5], we have defined an approach for classifying XML documents against a set of DTDs. The approach relies on a similarity measure that evaluates the distance between a document and a DTD. The similarity degree returned by the measure depends on parameters we have introduced in order to give a proper evaluation of the similarity for the applicative context in which it is used. Different relevance can be given to elements in the document not specified in the DTD (*plus* elements) w.r.t. elements required in the DTD that are missing in the document (*minus* elements). Moreover, different relevance can be given to elements at an high level of the hierarchical structure of the document w.r.t. those at a lower level. Finally, different relevance can be given to the *common* elements between the document and the DTD and for elements for which only a element tagged by a stem has been found in the DTD.

Basic Goal	Research State (0 ... 3)	Reference
A Formal Model of XML Documents	3	[4, 7]
A Similarity-Based Approach for Classifying XML Documents	3	[5, 6, 7]
Schema Evolution	2	[8]
A Similarity-Based Approach for Querying XML Documents	1	
An Access Control Model for XML Documents	3	[1, 2, 3, 4]

Figure 1: Current status of the research

The algorithm for computing the similarity measure has been implemented and the running time determined. A relevant result is that, in the most significant subset of DTDs, the similarity measure can be computed in polynomial time w.r.t. the number of nodes of the tree structures of the document and DTD. Moreover, approaches for reducing the running time for a general DTD have been studied and, currently, we are working for integrating them in the implementation.

Experiments have been performed on XML documents extracted from HTML pages describing software products and synthetic XML documents. Synthetic data are generated by means of a random generator we have developed. The generator produces XML documents containing elements arbitrarily chosen from the Dublin Core element set [10]. Each element, representing information about a resource (e.g. `title`, `creator`, `subject`), is optional, repeatable, and may appear in any order. Moreover, they can have qualifiers, i.e. subelements, refining their meaning. For example, the `creator` element can have the `personalName` subelement specifying the name of the creator.

We have rated this task as 3 (completed), even if some additional effort could be required for defining an interface (perhaps a web interface), by which it would be possible to set the different parameters on which the similarity measure is based and to "graphically" show the different behaviors of the similarity measure. Moreover, with few modifications of the similarity measure it is possible to extract the view of the document covered by the DTD and the view of the DTD.[1] However, we are planning to enroll a student to implement such facilities.

1.3 Schema Evolution

Starting from the requirements presented in the Thesis proposal [15] we have developed, with the help of a master thesis [16], an approach for evolving the structure of a DTD in order to adhere to the actual structure of its instances.

The proposed approach relies on the use of the similarity measure for gathering information on the plus and minus elements and on the frequent "*patterns*" identified in the elements of a document that weakly conform to the corresponding DTD declaration. A pattern is a subset of the tags of subelements of a non-valid element of the document w.r.t. a DTD. Patterns are used for identifying groups of elements frequently bound together and, thus, to extract the new structure of the DTD declaration of such element. Such information are stored in a data structure and used, when the evolution process is triggered, for evaluating which elements should evolve and how.

The evolution process cannot be triggered whenever a document not conforming to a DTD enters in the source. The evolution process has a high cost in terms of re-writing the applications that are working on the source. Therefore, it should be triggered whenever the DTD is not representative anymore of its instances and such "update" will improve the performance of applications that work on them. Different events have been considered. They depend on the access frequency to the DTD instances, on the number of non-conforming elements w.r.t the DTD, and on the number of documents currently considered as instances of the DTD.

The evolution process is based on three key principles.

- Use of data mining association rules [13, 14] for determining the most frequent patterns in the structure of subelements of each element. For each element of the DTD, relying on the patterns stored in the data structure, it is possible to determine:
  - elements that are always together (i.e. bound by an `AND` operator);

[1] The view of the DTD is the subpart of the DTD for which has been found a match in the document.

  - elements that are never together (i.e. bound by an `OR` operator);
  - elements, or groups of elements, that are repeated the same number of times (i.e. bound by a `*` or `+` operator);
  - elements, or groups of elements, that are optional (i.e. bound by an `?` operator).

  Note that the terms "always", "never", and "same number" should be considered in their statistical sense, i.e. in *most cases*. Moreover, in order to establish when the presence of an element implies the absence of another element, association rules like "*if element* `a` *is missing then element* `b` *is present*" should be considered.

- Incremental modification of the DTD. Approaches proposed in [11, 12] for inferring the "type" of a set of documents consider all the documents at once. Therefore, when a new documents is added to the set, in order to determine a new "type", the process starts from scratch. By contrast, in our approach we incrementally store the relevant information in the data structures and use them during the evolution process. The information in the data structures can be kept after the evolution or removed. In both cases we do not need to re-consider the documents that participated in the generation of the current DTD, because the DTD already takes them into account.

- Relevance of previous instances of the DTD. Different relevance can be given to the current structure of the DTD w.r.t. the documents classified against it since last DTD evolution. If the DTD was a *dummy* DTD generated from a training set of documents or, for the particular application area, the rule "more recent, more relevant" holds, then the DTD evolution process should forget the previous structure of the DTD and modify the DTD structure in order to be obtain a new one that closely represents the documents classified in the DTD since the last DTD evolution. By contrast, if the DTD structure is consolidated we want to minimize the DTD modifications in order to cover both the previous structure of the documents and the new structure deduced from the document classified since last DTD evolution.

We have rated this research task 2 because the data structures and the algorithms have already be defined, but we need to implement them and to perform some experiments in order to validate the proposed approach.

1.4 Querying XML Documents

This research task is still at an early stage. With respect to the work plan presented in the PhD Thesis proposal [15], the query language we are considering is Xpath and we do not consider the possibility to check the query w.r.t. the DTD of the source before evaluating the query against the DTD instances.

We have defined a mapping function that takes as input an Xpath expression and generates a document template, i.e. an XML document representing the structural and content constraints an XML document of the source should verify in order to be an answer to the query. We are currently adapting the classification approach for answering queries on documents based on their structures.

The need of shifting from exact queries with boolean answers to proximity queries with ranked approximate results has emerged as a requirement of XML query languages for searching the Web and some approaches in this direction have been developed [9]. These approaches, however, consider similarity only between terms appearing in the documents, that is, similarity of content, and disregard structure similarity. Relying on our approach to measure structure similarity, the query mechanism we aim at developing in our PhD Thesis will combine classical content-based queries and structure-based queries, both evaluated as proximity queries with a numeric rank that quantifies similarity.

1.5 An Access-Control Model for XML Documents

In this context we have obtained two main results:

- Definition of an access control model [1, 3, 4] that addresses the requirements outlined in the PhD Thesis Proposal [15]: presence of valid and well-formed documents, the hierarchical and interlinked structure of XML documents, the presence of subjects with different characteristics.

- Definition of facilities for the security officer for the definition of policy rules [2]. One of the most relevant facilities is the one that allows the security officer to define the access control policies for a document created outside the source by means of the classification approach developed in task 2 of this PhD thesis.

We think that for what concern this task there is enough contribution for the thesis. Therefore, the task is rated 3. However, the results obtained in the evolution DTD task encourages the idea to define mechanisms for automatically modifying the policies specified at DTD level whenever the DTD is modified. Actually, this facility will simplify the work of the security officer when the structure of the DTDs evolves and new access control rules should be implemented. However, we are not sure that there will be enough time to investigate such extension before the delivery of the thesis.

2 Work Plan for the Last Year of Thesis Work

As seen in Section 1, the main topics still to investigate, to complete the thesis work, are the following:

1. Implementation and validation of the schema evolution approach.
2. Investigation of the approach for querying the source of documents.

Among the previous tasks we believe that the second task will require more work than the previous one.

3 Structure of the Thesis

The thesis title will be "*Classification and Querying of XML Documents*". The thesis to be developed could be structured in the following chapters:

1. **Introduction** motivating the thesis, and presenting its structure.
2. **Related Work** surveying the pertinent literature, giving the starting point of the work;
3. **The XML Document Model**, introducing the formal definition of XML documents and DTDs.
4. **A Similarity-Based Approach for Classifying XML Documents**, proposing the classification approach and the similarity measures.
5. **Schema Evolution**, proposing the schema evolution approach for updating the DTD schema in order to adhere to its real instances.
6. **A Similarity-Based Approach for Querying XML Documents**, presenting the querying approach, and the similarity measures specifically tailored for the querying process.
7. **An Access Control Model for XML Documents**, presenting the access control model for XML document sources and the facilities developed for the security officer.
8. **Conclusions** summarizing the contribution of the thesis and outlining further developments.

With respect to the Thesis structure presented in [15], two chapters have been devoted to the approach for classifying XML documents. The first one will be devoted to present the similarity measure and the algorithm for matching XML documents against a set of DTD. By contrast, the latter will be devoted to present the schema evolution process. The alternative solution to maintain a single chapter will produce a very long chapter dealing with different aspects.

Moreover, if there will be time for investigating the extension of access control mechanisms for adapting the access control policies specified for a DTD to the restructured one, Chapter 7 will be split into two chapters. The first one will deal with the access control model developed for XML and the facilities developed for helping the security operator in the specification of access control policies for documents created outside a source by means of the classification approach. The latter will deal with the approaches for restructuring the access control policies defined on a DTD when such DTD evolves to another one that is a better representation of its instances.

Since the topics addressed by the thesis are quite broad, the chapter on Related Work may become very long and not well integrated. A possible alternative solution is to split the discussion on related works in various sections in specific chapters of the thesis (e.g. the discussion of existing approach for protecting the access to XML documents can be presented as a section of Chapter 7 rather than in Chapter 2). Thus, we could present in Chapter 2 only general overview of the work done in the context of document classification, evolution and querying and defer the discussion of the extensions to the specific chapters.

References

[1] E. Bertino, M. Braun, S. Castano, E. Ferrari, and M. Mesiti. Author-X: a Java-Based System for XML Data Protection. In *14th IFIP 11.3 Working Conference in Database Security*, Schoorl, The Netherlands, 2000.

[2] E. Bertino, S. Castano, E. Ferrari, and M. Mesiti. Protection and Administration of XML Data Sources. Submitted for journal pubblication.

[3] E. Bertino, S. Castano, E. Ferrari, and M. Mesiti. Controlled Access and Dissemination of XML Documents. In *Proc. 2nd ACM Workshop on Web Information and Data Management (WIDM'99)*, pages 22–27, Kansas City, Missouri, 1999.

[4] E. Bertino, S. Castano, E. Ferrari, and M. Mesiti. Specifying and Enforcing Access Control Policies for XML Document Sources. *World Wide Web Journal*, 3(3), 2000.

[5] E. Bertino, G. Guerrini, I. Merlo, and M. Mesiti. An Approach to Classify Semi-Structured Objects. In *Proc. Thirteenth European Conference on Object-Oriented Programming*, LNCS 1628, pages 416–440, 1999.

[6] E. Bertino, G. Guerrini, and M. Mesiti. Matching XML Documents against a Set of DTDs. Submitted for publication.

[7] E. Bertino, G. Guerrini, and M. Mesiti. Measuring the Structural Similary among XML Documents and DTDs. Submitted for journal publication.

[8] E. Bertino, G. Guerrini, M. Mesiti, and L. Tosetto. Evolving a Set of DTDs according to a Dynamic Set of XML Documents. In preparation.

[9] T. Chinenyanga and N. Kushmerick. An Expressive and Efficient Language for XML Information Retrieval. *American Society for Information Science and Technology*, 2001. To appear.

[10] Dublin Core Initiative. `http://bublingcore.org`.

[11] M. N. Garofalakis, A. Gionis, R. Rastogi, S. Seshadri, and K. Shim. XTRACT: A System for Extracting Document Type Descriptors from XML Documents. In *ACM SIGMOD International Conference on Management of Data, 2000, Dallas, Texas, USA*, pages 165–176. 2000.

[12] J. Hammer, H. Garcia-Molina, J. Cho, R. Aranha, and A. Crespo. Extracting Semistructured Information from the Web, 1997. ftp://db.stanford.edu/pub/paper/extract.ps.

[13] J. Hipp, U. Guntzer, and G. Nakhaeizadeh. Algorithms for Association Rule Mining- a General Survey and Comparison. *SIGKDD Explorations*, 2(1):58–64, 2000.

[14] G. Lee, K. Lee, and A. Chen. Efficient Graph-Based Algoritms for Discovering and Maintaining Association Rules in Large Databases. *Knowledge and Information System*, 3(3):338–355, 2001.

[15] M. Mesiti. Classification and Querying of XML Documents. PhD Thesis Proposal, PhD in Computer Science, University of Genova, 2000.

[16] L. Tosetto. Evoluzione di un insieme di DTD in base ad un insieme dinamico di documenti XML. Master thesis, University of Genova, 2001. In italian.