Reusing Dynamic Document Fragment through Virtual Documents: Key Issues In Document Engineering

FALQUET, Gilles, et al.

Abstract

Electronic documents may obviously no more be considered as a simple representation of their paper counterpart. They become dynamic components whose content may be modified accordingly to users interaction but also in reaction to modifications in the user's environment (such as an update operation in a database). They not only include data but also behavior (Scripts, Applets, etc.) that may aim at different purposes (provide appropriate rendering method, supply interaction mechanism to the users, etc.). They also include the hypertext dimension and play an important role in the management of information as well as associated knowledge. The extensive use of such hyperdocuments through the Internet, to disseminate and interact with pieces of information, is currently gaining in importance. Designing information systems that rely on this emerging paradigm clearly requires new methods to produce, share and access distributed information. The paper focus on the reusability of information in distributed environments. It aims at presenting our document based approach to both model active document fragments and specify [...]
Reusing Dynamic Document Fragment through Virtual Documents: Key Issues In Document Engineering

Yassine A. Rekik*, Christine Vanoirbeek*, Gilles Falquet**, Luka Nerima**

* MEDIA Research Group
LITH-DI-EPFL – IN Eculbens – CH 1015 Lausanne - Switzerland
*Yassine.Rekik@epfl.ch, Christine.Vanoirbeek@epfl.ch
Tel +41 21 693 6782

* Centre Universitaire d'Informatique
Université de Genève - 24, rue Général-Dufour – CH 1211 Genève 4 – Switzerland
**Gilles.Falquet@cui.unige.ch, Luka.Nerima@cui.unige.ch
Tel +41 22 705 77 72

Abstract
Electronic documents may obviously no more be considered as a simple representation of their paper counterpart. They become dynamic components whose content may be modified accordingly to users interaction but also in reaction to modifications in the user’s environment (such as an update operation in a database). They not only include data but also behavior (Scripts, Applets, etc.) that may aim at different purposes (provide appropriate rendering method, supply interaction mechanism to the users, etc.). They also include the hypertext dimension and play an important role in the management of information as well as associated knowledge.

The extensive use of such hyperdocuments through the Internet, to disseminate and interact with pieces of information, is currently gaining in importance. Designing information systems that rely on this emerging paradigm clearly requires new methods to produce, share and access distributed information. The paper focus on the reusability of information in distributed environments. It aims at presenting our document based approach to both model active document fragments and specify appropriate methods to build customized virtual documents on the basis of so-called fragments.

Introduction
The concept of structured document relies on the use of attribute grammars to formally describe documents classes [Furuta87]. Such an approach allows the production of documents that conform to an abstract model and, thus, enhance the specification of appropriate processing operations on documents belonging to a given class. Initially used for the production and exchange of documents in publishing purposes, the concept of structured document appeared to be of much wider interest.

Document reuse
How to reuse documents, or pieces of document, is currently addressed in many ways and multiple purposes. We propose a classification based on approaches originating from different research areas. We distinguish between three main problems to be faced in this respect:

- How to add structure to existing documents? This is a matter of analysis;
- How to transform existing documents to reuse them within other applications? It ranges from...
format transformations to more complex structure manipulations;
- How to design documents in such a way that facilitates their reuse by existing or further applications? This is a “document engineering” problem, very similar to design problems addressed in the software engineering domain.

**Structuring or re-structuring documents**

It is commonly accepted that availability of structured documents greatly facilitates the reuse of content. In this respect, a number of research works are dedicated to the analysis of raw or semi-structured documents in order to structure or re-structure them in such a way that facilitates their reuse through existing applications. In this respect, a lot of effort is dedicated to the analysis of document images, either obtained by scanning operation or generated by applications, either obtained by scanning operation or generated by applications, or either obtained by scanning operation or generated by applications, such as Postscript files [Bapst98]. Another approach is adopted in MarkItUp [Fankhauser94], a system designed to recognize the structure of untagged electronic documents; it is based on a learning by example approach to gradually build recognition grammars. Extracting logical structure of library references has been addressed by a mixed strategy combining image analysis and structural information provided by the representation standard used (Unimarc). The system is based on a constraint propagation method [Belaid97]. Finally, we may also cite work performed to interactively restructuring HTML documents, an approach based on the use of a transformation language [Bonhomme96].

**Transforming documents**

Despite the obvious advantages conveyed by the manipulation of documents in a structured way, reusing them within users’ environments raises a number of fundamental problems to transform or to adapt their intrinsic structure:

- How to provide the users with appropriate editing operations (such as cut and paste operations) that maintain the document consistency?
- How to merge documents conforming to different structures?
- How to guarantee the consistency of existing documents that relate to an evolving generic structure?
- How to transform existing structured documents to reuse them in the framework of applications relying on another document model?

The issue of structures transformations is complex. Depending on the context, several approaches have been proposed or are currently under investigation to address this problem. An interactive context (such as document editing) highly promotes the use of automatic transformations that have to be performed in an efficient way to accommodate a required response time [Akapotsui97]. Transforming existing structured documents in order to fit a different target structure may be performed on the basis of explicit specification provided through descriptive rules to guide the transformation process [XSL99]. A combined approach has also been proposed to take benefit from the two mentioned approaches [Bonhomme97][Bonhomme98].

**Modeling modular structured documents**

The formal description of a document class aims at constraining the logical organisation of document instances and, thus guarantee their adequacy to a given model. Attributed grammars have been originally used to define classes of document for publishing purposes. Initially, such formal descriptions consisted in "monolithic" descriptions, each of them, representing the (logical) editorial structure of the document class. The need to build such descriptions in a modular way, taking benefit from existing document fragment descriptions, has been rapidly identified and addressed in different ways. Some document management systems explicitly provide appropriate mechanism to define reusable units of structured pieces of documents.

However, dealing with reusable fragments of documents raises problems, very similar to those encountered in the software engineering domain, that are not currently solved. Ongoing research works have mostly adopted an object-oriented design for the appropriate representation of documents models [SOX99][Abiteboul99]. This approach is relevant in many respects. It allows the precise definition of pieces of information as well as the processing operations to be performed on identified structured data. It aims at facilitating the exchange and reuse of document fragments between heterogeneous applications.

**Reusable Dynamic Fragments and Virtual Documents**

Reusing information implies that there exists a mechanism to combine (or compose) the available pieces of information into new documents derived according to the users’ needs. Two major issues have to be addressed to reach this objective: how to model pieces of information and how to provide methods for assembling such pieces in a consistent way?

**Dynamic Document Fragment and Dynamic Fragments Server**

Many work has been done in order to offer repositories of documents in order to reuse them. However, the majority of works emphasized on documents collection, documents identification and documents access. The integration of those fragments when authoring new documents, especially in the case of structured documents is not yet answered in a satisfying way. In addition, the majority of works
focused on documents content reuse. The reuse of document content, structure and behavior remains an open problem.

The first concept that seems for us a major concept for document reuse is what we call "Dynamic Structured Document Fragment". We consider the document fragment as an independent, self-described piece of information that can be reused and adapts to new documents. The definition of a dynamic structured fragment includes several aspects:

- the definition of the granularity of such reusable fragments
- the definition of a fragment content model
- the definition of fragment metadata, to characterize and identify it within an authoring process.
- the definition of a mechanism to associate methods to describe the fragment interface and behavior.

In our approach "reusing" consists in integrating both the structured content of the fragment and the associated methods, in order enhance processing operations on the fragment. The fragment can be manipulated separately with appropriate authoring tools, or within a more general context by applications that manipulate the new created document.

In order to access and reuse document fragments, our approach is based on a "Fragment Server". The design of such a server must take into account three major points. First, how to manage and store such dynamic structured fragment. Two promising approaches may be adopted for this task. The first is the use of multimedia object databases which are appropriate for handling dynamic fragment. The second in the structured document databases which are emerging currently and seem to be appropriate for managing document based data. The second task will be to define the access functions and mechanisms that a fragments server have to offer. Also in this task, we will investigate two directions. The querying based approach represent the first direction and is based on recently proposed document querying languages. The second direction is based on navigational models. This will be based on the recent work done in the domain of hypertext navigation and information retrieval. Finally, the third point to be addressed when designing a fragment server in the exchange protocol and format. In fact, a model to exchange dynamic fragment with authoring tools and/or document processing applications is necessary to ensure the import/export function between authors and the fragment server. When proposing this exchange model, a special attention will be accorded to the use of standards and interoperable technologies such as XML, Java, Corba, etc.

**Database Views and Virtual Document**

The notion of view has been extensively studied in the field of databases. Views were first introduced within the relational data model as derived relations. Theoretical and applied works have dealt with problems such as updating databases through views, computing and updating materialized (stored) views, etc. More recently, the notion of view has been (partially) adapted to other models such as object-oriented data models, semi-structured models and document models.

In recent years there has been a strong interest for virtual documents and virtual hypertexts. For instance, the development of the World Wide Web has prompted research on the generation of virtual hypertexts (made of Web pages) to represent the content of databases. Works on this theme are being presented in the WebDB workshops and SIGMOD conferences. In the proposed research we intend to adapt recent results on hypertext views on databases to the generation of virtual documents on a repository of document fragments. Recently, there has been a strong interest for Personalized virtual documents, and a community of researchers is gathering on this theme.

The virtual document model will precisely define how to derive each virtual document from the stored document fragments. The specification of a virtual document will contain at least:

- a selection specification: determines which fragments to use; it is a set of queries on the fragment repository (based either on fragment content or meta-content);
- a global document structure, which can be either static (fixed in the specification) or dynamic (computed on the document fragments, e.g. links obtained by computing a semantic proximity between two fragments); global structures may take various forms, such as sequences, trees, graphs, etc.
- a presentation specification which indicates how to represent the global structure (for instance, a tree can be represented graphically, with embedded boxes, linearly with parentheses, etc.)
- fragment operations: specify which operation (method) to apply on each fragment (e.g. a fragment representing a formula may be expressed in a given system of units, a picture may be resized or rotated, etc.)
- inheritance relationships: a specification may be derived from one or more other specifications (this is particularly useful to adapt existing specifications to new needs)

We will develop a language to express such specifications. This language will be an extension of a document specification language. Its formal semantics will be expressed in terms of the virtual document model.
Bibliography


