Published in: The Association for Literary and Linguistic Computing and the Association for Computers and the Humanities ALLC/ACH (Ed.): New Directions in Humanities Computing. Proceedings of the 14th Joint International Conference, University of Tübingen, 24-28 July 2002. - Tübingen: Universität Tübingen, 2002. Pp. 60-62.

"Towards declarative descriptions of transformations: An approach based on topic maps"

Eva Lenz
University of Bielefeld
eva.lenz@uni-bielefeld.de

Andreas Witt
University of Bielefeld
andreas.witt@uni-bielefeld.de

Angelika Storrer
University of Dortmund
storrer@hytex.info

Introduction

The topic map standard ISO 13250 (Pepper 1999) provides an innovative and standardized way to express information about documents.. Although the relevance of topic maps for the humanities has been recognized, only few applications in this area have been developed up to now. In the following we will explain an approach for transforming document collections which uses a topic map to express meta information. We believe that this approach can be applied to various areas in the humanities where information about documents is important, and will present two research areas where our model can be employed: adaptive hypertext and relationships between document grammars. Brusilovsky (2001) reviews the field of adaptive hypermedia. Adaptive hypermedia systems build a model of the goals, preferences and knowledge of the individual user and use this model during user interaction in order to adapt the hypertext to the user's needs. Important application areas are adaptive educational hypermedia and systems for managing personalized views in information spaces. In this paper, we outline how topic maps can be used to support the development process and maintenance of adaptive hypertexts. Document grammars define the structure of XML or SGML annotations using a schema language. They can be used to check the validity of a document. The oldest schema language is used for expressing document type definitions (DTDs). Recently, a number of new schema languages have evolved in this field, most of which use an XML syntax (Cagle et al. 2001). It can be beneficial to formally express the relationships between two or more document grammars. In the case of DTDs, these relationships can be described using XML or SGML architectures as defined in an annex of ISO 10744 (HyTime). We will show how topic maps can be used to model relationships between different document grammars, abstracting from the schema language in which they are expressed.

Description of the model

Our model can be employed to describe and perform a transformation of a collection of documents into a new one, taking into account a specific context. In the following we speak of the initial document collection (or the corpus) and the generated document collection. The model is based on a three-level information architecture (see figure), consisting of

- 1. the document level: a corpus of annotated documents,
- 2. the conceptual level: a knowledge net, modelling relationships between the documents (or parts of annoted documents) in the form of a topic map, and
- 3. the application context level: information about the context in which the transformation is to take place. For example, in the case of adaptive hypertext, this level contains the user model.

A set of rules describes how the new document collection is to be generated from the corpus, taking into account the state of the knowledge net and the application context. We use XML and related technologies for the implementation of this model. This allows us to use and produce documents in the web context, and to apply available tools (parsers, browsers, generators, validators). We assume that all model components be expressed declaratively in an XML syntax: the initial document collection, the topic map (Pepper & Moore 2001), the information about the application context, and the rule language. By expressing the rules in a declarative language instead of hard-coding them into a programming language a level of abstraction is introduced. Each rule can be regarded as a mapping which takes a condition (on XML documents) as its input and produces an action (XML fragment) as its output. The expressions allowed in the rules, and thus the type of transformation the system is able to perform, are clearly defined by a document grammar. The generation of documents is realized in two steps, comparable to the implementation of the Schematron schema language (cf. Cagle et al. 2001, cp. 14). Each step is performed by an XSLT stylesheet. XSLT is a functional programming language optimized for parsing and generating XML documents. The first stylesheet takes the rules as its input and translates them into a second XSLT stylesheet, i.e. it translates the abstract descriptions of conditions and actions into an executable form. This second stylesheet takes the initial document collection, the topic map, and the application context information and generates the new document collection. This approach makes the rules more compact, easier to understand, and easier to maintain than program code.

Application A: Adaptive hypertexts

In the case of adaptive hypertexts, a collection of documents is tailored to suit the user by adapting the content and the linking structure. Brusilovsky (2001) describes a detailed taxonomy of the possible forms of adaptation, which he calls adaptation techniques. Typically, the collection is filtered (e.g., an expert of the domain sees other hypertext nodes and links as a novice). Other adaptation techniques comprise sorting, visual annotation, or "dimming" of links, as well as "stretchtext". When we apply our model to the field of adaptive hypertext, the initial document collection (on the document level) comprises all documents for every type of user, whereas the generated document collection is a hypertext adapted to one user. The relevant characteristics of that user are described in the user model (application content level). For example, we could model that users from a neighboring discipline "know" only certain concepts and technical terms of the domain in question. In the context of a project called HyTex (hypertext conversion based on textgrammatical annotation, cf. Lenz & Storrer 2002) which is part of a joint research group "Text-Technological Modelling of Information", we model technical terms of a subject domain together with their relationships in the form of a topic map (conceptual level). Different views on the subject domain (on its concepts, technical terms, on their relations to other terms and to documents) can be modelled using various topic map constructions. These views correspond to characteristics of the user model. The rules describe how an adapted hypertext is to be created from the initial document collection taking into account the user model and the topic map. The code that has to be generated will always be very similar for the same adaptation technique, e.g. for dimmed links. In this case, we would allow "dimmed text" as a possible action in the rule language, and implement the code once in the XSLT stylesheet. We benefit from this approach in the HyTex project, as it allows us to test different strategies for text-to-hypertext conversion, which can readily be expressed in the form of declarative rules. The model can thus serve as a tool in the area of hypertext research.

Application B: Expressing and processing relations between document grammars

XML and SGML architectures (architectural forms definition requirements, AFDR) allow one to express relations between two DTDs, a meta DTD and a client DTD. They have been used in areas relevant to the humanities (Simons 1999). Software that can process architectures can automatically transform a document annotated according to the client DTD into a document instance of the meta DTD. This process is called derivation. Examples of use include:

- filtering document content,
- filtering markup (e.g., transforming a richly annotated corpus into a corpus with annotations specialized for a certain purpose),
- modifying markup (e.g., translating tag names from german to english), and
- changing attributes into elements and vice versa.

On the poster, we will show how document grammars - not only DTDs - and their relationships can be expressed using a topic map, and how to apply our model to generate derivations. In this application scenario for our model, the initial document collection contains documents annotated according to the "client" document grammar. The generated ("derived") document collection will be annotated according to the "meta" document grammar. The following steps have to be performed:

- Translating the document grammars into a topic map (conceptual level). This can be done automatically, e.g. using XSLT, when the document grammar has an XML syntax.
- Modeling the relations between the document grammars in the topic map (conceptual level).
- Specifying which documents are to be derived according to which document grammar (application context level).
- Describing the translation process by means of the rule language.

Application B models the set of relations between document grammars as defined in the AFDR, but our model allows for the extension of this limited set. This application of our model is used in the context of the project "Secondary information structuring and comparative discourse analysis" (cf. Sasaki et al. 2002). This project is also part of the joint research group already mentioned.

Conclusion

We have described a model based on topic maps designed for a declarative description of a document collection and transformations on that collection in an application context. We have presented two important fields of application for this model, but we assume that its application can be extended to other areas relevant to the humanities.

Bibliography

Peter Brusilovsky. "Adaptive hypermedia." *User Modeling and User-Adapted Interaction.* 2001. 11: 87-110.
Kurt Cagle Jon Duckett Oliver Griffin Stephen Mohr Francis Norton Nik Ozu Ian Stokes-Rees Jeni Tennison Kevin Williams. *Professional XML Schemas.* Birmingham: Wrox Press, 2001.
Eva Anna Lenz Angelika Storrer. "Converting a corpus into a hypertext: An approach using XML topic maps and XSLT." *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002), Las Palmas.* : , 2002.
Steve Pepper. "Navigating haystacks and discovering needles. Introducing the new topic map standard." *Markup Languages: Theory & Practice.* 1999. 1: 47-74.
XML Topic Maps (XTM) 1.0. Ed. Steve Pepper Graham Moore. : , 2001.
Felix Sasaki Claudia Wegener Andreas Witt Dieter Metzing Jens Pönninghaus. "Co-reference annotation and resources: a multilingual corpus of typologically diverse languages." *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002), Las Palmas.* : , 2002.
Gary Simons. "Using Architectural Forms to Map TEI Data into an Object-Oriented Database." *Computers and the Humanities.* 1999. 33: 85-101.