SALT: A semantic approach for generating document representations

Tudor Groza Alexander Schutz Siegfried Handschuh

Digital Enterprise Research Institute
National University of Ireland, Galway
IDA Business Park, Lower Dangan
Galway, Ireland
{tudor.groza, alex.schutz, siegfried.handschuh}@deri.org

ABSTRACT

The structure of a document has an important influence on the perception of its content. Considering scientific publications, we can affirm that by making use of the ordinary linear layout, a well organized publication, following a "red wire", will always be better understood and analyzed than one having a poor or chaotic structure, but not necessarily poor content. Reading a publication in a linear way, from the first page to the last page means a lot of unnecessary information processing to the reader. Looking at a publication from another perspective by accessing the key-points or argumentative structure directly can give better insights into the author's thoughts, and for certain tasks (i.e. getting a first impression of an article) a representation of the document reduced to its core could be more important than its linear structure. In this paper, we will show how one can build different representations of the same document, by exploiting the semantics captured in the text. The focus will be on scientific publications and as building foundation we use the SALT (Semantically Annotated LaTeX) annotation framework for creating Semantic PDF Documents.

Categories and Subject Descriptors

I.7.2 [**Document and Text Processing**]: Document Preparation; I.7.4 [**Document and Text Processing**]: Electronic Publishing

General Terms

Design, Experimentation

Keywords

LaTeX, PDF, semantic annotation, semantic document

1. INTRODUCTION

The dissemination stage of research can be seen as a communication process of scientists over time, where researchers present and support their findings and argue about claims in scientific publications. While this communication takes place over the course of several publications, each paper itself consists of a so-called rhetorical discourse structure which lays out supportive evidence for the raised claims. Annotating this rhetorical structure in a text does not only enable the linking of chunks of information, it also offers a different perspective than the – in terms of sentences – usually linear document representation. Semantic annotations are created by marking up chunks of text with rhetorical relations, and by referring to terminology from a publicly available ontology.

In order to achieve this goal, we make use of SALT (Semantically Annotated LaTeX) [2]. SALT is an authoring and annotation framework for creating *concurrent* semantic annotations for PDF documents, by exploiting the rich environment provided by LaTeX. The annotation process takes place while writing and the actual integration is realized at syntax level by exploiting regular LaTeX commands plus a series of newly introduced special annotation commands. The framework is mainly composed by two layers: (i) a semantic layer, capturing knowledge present in the publication, and (ii) a syntactic layer, providing the means for the authors to annotate their paper.

The syntactic layer extends the LaTeX environment, which is a markup language and typesetting system for documents. It is widely being used for publications in the scientific area and in research, and is capable of producing high quality typeset output. The original markup comprises of commands for the linear and logical structure of the document, i.e. partitioning the content into elements like sections or subsections, cross-referencing formerly defined elements or describing semistructured containers like tables. In order to cater for semantic annotations with respect to a pre-defined ontology, we extended the set of available LaTeX commands in such a way that it is possible to annotate chunks of text with rhetoric relations which are based on the Rhetorical Structure Theory (RST) [4], a theory that describes the organization of text and the relationships that hold between its parts.

RST represents the foundation of the semantic layer, which captures also concepts for the IBIS [3] methodology. Overall, it comprises a set of three ontologies, able to capture the structural information of the document as well as the semantics of its content. The three ontologies are (i) the *Document ontology*, (ii) the *Rhetorical ontology* and (iii) the *Annotation ontology*.

Using the LaTeX compiler, it is possible to translate the document content directly into the Portable Document Format (PDF), which supports the specification of metadata and acts as a container to hold the formerly created content and the annotations in a single file. Exploiting the semantic annotations during compilation time, different representations (or views) of the document can be created:

(i) the usual *linear view*, (ii) the *rhetorical block view*, and (iii) the *rhetorical tree view*

The final result of the annotation process is a semantically enriched PDF document encapsulating instances of the afore-mentioned ontologies together with associated visual annotations. We believe that the ontologies presented in our proposal can be used independently of the format used for the scientific publications. Therefore, we intend to use the current approach as a proof of concept and extend our investigations to other formats in the near future.

2. MODELING THE DOCUMENT STRUCTURE

The common approach for presenting or representing scientific publications, part of online proceedings or just as documents on the web, is by using the printing (linear) layout. This works fine, and it follows the natural way, for printed publications, but we argue that for digital documents the representation can be improved. The structure of a document has an important influence on the perception of its content. Thus, a well organized publication, following a "red wire", will always be better understood and analyzed than one having a poor or chaotic structure, but not necessarily poor content.

By exploiting the semantics captured in the text, one can provide different representations of the same publication. Based on an annotation framework introduced in the following section, we are able to represent a document using a structure built from the semantics present in the document's content. Therefore, besides the linear layout, we propose two additional representations: (i) a *rhetorical block representation*, and (ii) a *rhetorical tree representation*. Both representations have their roots in the Rhetorical Structure of Text (RST) Theory [4] and provide an argumentation support for the document's content. (i) models the document based on a best practice structure for scientific publications, whereas (ii) shows the relationships existing between different information chunks present in the text.

Starting from the logical structure of the document, the rhetorical block representation creates a perspective over text, along a best-practice structure for scientific publications (see Figure 1.a). It (re)organizes the structure into nine main sections, out of which some can be absent, depending on the author's goals. The sections are: *Abstract*, *Background*, *Motivation*, *Scenario*, *Contribution*, *Evaluation*, *Discussion*, *Conclusion* and *Entities*

During the writing process, the author can model directly the publication according to this structure. In practice this is hard to achieve, and the semantics captured by these concepts is usually interleaved between the sections composing the document, and sometimes several concepts are present within a single section. Therefore, the goal of this representation is to unite the author's thoughts spread over the borders of the document's sections, giving the reader the possibility to analyze the publication through a fixed rhetorical structure.

If the previous representation gave a high level perspective over the text, the rhetorical tree perspective (see Figure 1.b) provides a finer granularity analysis and presentation. As background, in the Rhetorical Structure of the Text Theory (RST), a piece of text is broken into many text spans, which we call rhetorical elements. The rhetorical tree structure represents a form of organizing the text by means of rhetorical elements (leaves) and the relations existing between them (branches) [5]. The central claim of RST [4] is

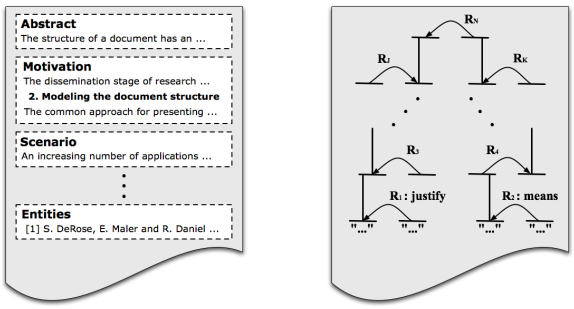

Figure 1: Rhetorical (a) Block and (b) Tree representation

that the structure of every coherent discourse can be described by a single rhetorical structure tree. From our point of view, this representation offers the reader the possibility of navigating through the publication and having as guide the semantics encapsulated in the relations indicated directly by the author.

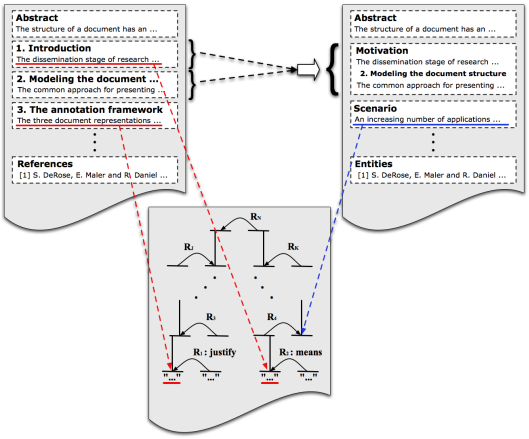

Figure 2: A parallel view over the three proposed document representations

Figure 2 provides a parallel view between the three representations of the same text. It can be observed how the logical (linear) structure of the text transforms its shape when the semantics of the document's content is taken into account, at different levels of granularity. A challenge that arises from this parallelism, from the interaction point of view, is to find an appropriate mechanism for easily (and on-the-fly) switching between the perspectives, without the user losing her current focus.

In order to provide a working example, we implemented an application showing how the semantic documents enable an easy, low-effort information distribution and integration for the purpose of an innovative online workshop proceedings. The metadata is used not only to ease the process of creation of the online proceedings, but also provide added value to the reader of these proceedings, by generating the different representations generated from the semantics encapsulated in the documents' content.

We propose an approach which is flexible in a way that formatting choices are left to the editor: templates for the HTML generation process are offered and can be exchanged on the fly. For each annotated paper, we create an individual HTML page display-

ing a particular representation of the publication, whereas for the entire proceedings we generate the index, giving a short overview of all papers. The publications' perspective is chosen at proceedings compile time, with the remark that currently the system allows only an exclusive choice between the linear and rhetorical block representations. In the meantime we are experimenting with the generation of the rhetorical tree representation and of interleaved representations with the possibility of directly switching between them. We also investigate the possiblity of creating lenses around a particular rhetorical element.

3. RELATED WORK

In our analysis of related work, we will focus on two main aspects: (i) related ontological frameworks, and (ii) approaches targeting document representation issues. As theoretical background, our ontologies have their roots in the Rhetorical Structure of the Text (RST) Theory [4], where the authors provide the underlying semantics of the concepts modelled by the theory together with their definitions. Taboada and Mann [7] provide a deep analysis of the application domains in which RST was used until a certain point in time. It is interesting to observe that the mentioned range of domains varies from computational linguistics or cross-linguistic studies or dialogue to multimedia presentations.

A similar ontological framework is presented by Tempich et. al in [8]. The DILIGENT *Argumentation Ontology* was designed in line with the terminology proposed by the IBIS methodology [3] and captures the argumentative support for building discussions in DILIGENT processes. In DILIGENT, the argumentative support is equivalent to only one of the three parts of the Rhetorical Ontology and so is less expressive. Uren et. al [9] describe a framework for sensemaking tools in the context of the Scholarly Ontologies Project. Their starting point is represented by the requirements for a discourse ontology, which has its roots in the CCR (Cognitive Coherence Relations) Theory and models the rhetorical links in terms of similarity, causality or challenges. Although the ontological foundation is very similar, the application focus is on providing support for finding claims in scientific publications and visualize the constructed claim networks using a central knowledge server.

The other target of our analysis is represented by systems dealing with document representation, and a wide category of such systems are Wikis, or to be more precise, Semantic Wikis. Systems like IkeWiki [6] or WikSar [1] have the ability of presenting part of the information space through a particular perspective, depending on the type of the information. Some of the methods used to achieve this functionality are: interpreting the content of the information space, or classifying it by using a set of well established types (this case assumes the presence of an underlying knowledge base). In both cases, depending on the analysis result, a template is applied in order to give the document a different representation. Our approach is similar to the first method. We create the representation based on the annotated information chunks, considering the text's semantics, while using the erected rhetorical structure as a representation template.

4. CONCLUSIONS

In this paper we have described a solution for generating different representations of the same document, based on the metadata created by using our authoring and annotation framework. SALT leaves the semantic data where it can be handled best, merged with the document. Moreover, it provides a means to create Semantic Documents in a comparatively simple and intuitive way to use for LaTeX authors. The framework brings added value to the applications using PDF documents and to the users, as shown in our online proceedings scenario, where we used it to automate and improve the presentation and navigation of the scientific publications. Utilised in this way, SALT could also be integrated into the workflow of generating the semantic metadata for conferences such as ESWC or ISWC, a process that can be long and tedious if performed manually.

Acknowledgments

This work is funded by the European Commission 6th Framework Programme in context of the EU IST NEPOMUK IP - The Social Semantic Desktop Project, FP6-027705.

5. REFERENCES

[1] David Aumueller. Semantic authoring and retrieval within a wiki. In *Proceedings of the European Semantic Web Conference (ESWC), Demos and Posters, Heraklion, Greece, 29. May 1. June*, 2005.

[2] T. Groza, S. Handschuh, K. Möller, and S. Decker. SALT – Semantically Annotated LaTeX for Scientific Publications. In *Proceedings of the Fourth European Semantic Web Conference, (ESWC 2007), Innsbruck, Austria, May*, 2007.

[3] W. Kunz and H.W.J. Rittel. Issues as elements of information system. Working paper 131, Institute of Urban and Regional Development, University of California, 1970.

[4] W. C. Mann and S. A. Thompson. Rhetorical structure theory: A theory of text organization. Technical Report RS-87-190, Information Science Institute, 1987.

[5] Daniel Marcu. Building up rhetorical structure trees. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence, Portland, Oregon, August*, volume 2, pages 1069–1074, 1996.

[6] Sebastian Schaffert, Diana Bischof, Tobias Buerger, Andreas Gruber, Wolf Hilzensauer, and Sandra Schaffert. Learning with semantic wikis. In *Proceedings of the First Workshop on Semantic Wikis - From Wiki to Semantics, co-located with the European Semantic Web Conference (ESWC), Budva, Montenegro, June*, 2006.

[7] M. Taboada and W. C. Mann. Applications of rhetorical structure theory. *Discourse Studies*, 8, No. 4:567–588, 2006.

[8] C. Tempich, H. S. Pinto, Y. Sure, and S. Staab. An Argumentation Ontology for Distributed, Loosely-controlled and evolvInG Engineering processes of oNTologies (DILIGENT). In *Proceedings of the Second European Semantic Web Conference, (ESWC 2005), Heraklion, Crete, Greece, May*, 2005.

[9] V. Uren, S. B. Shum, G. Li, and M. Bachler. Sensemaking tools for understanding research literatures: Design, implementation and user evaluation. *Int. Jnl. Human Computer Studies*, 64, No.5:420–445, 2006.