December 1998

WDM: A Web Document Model and its Supporting Web Document Analyzer

William Song
Swedish Institute for Systems Development

WDM: A Web Document Model and its Supporting Web Document Analyzer

William W. Song
Swedish Institute for Systems Development

Abstract

The ever-growing information on the web has meanwhile brought people a great difficulty in quickly and easily finding the exact information they need. A problem causing the difficulty is that the web information and resources are less internally structured. In this paper, we propose a concise metadata model as an intermediate model that can transfer various styles of web models or metadata models used to describe web resources to a more general-used framework. We also develop a platform to realize the modeling functions in the framework.

Introduction

The use of the World Wide Web (web) has brought people a great convenience of obtaining information quickly and easily. However, the ever-growing information on the web has meanwhile brought people a great difficulty in quickly and easily finding the exact information they need. A problem causing the difficulty is that the web information and resources are less internally structured. In other words, there is little or even no additional information contained in a web information item used for identifying the web item. The term "metadata" is introduced to indicate this additional information to web information items for the purpose of description and identification.

Metadata can be considered to be the particular part of a web document, which contains information intentionally and especially for the web document identification. There are many metadata models at present to support the inclusion of metadata information in web documents. Some well-known metadata models, among others, are Dublin Core (DC) [1], Meta Content Framework (MCF) [2], and Extensible Markup Language (XML) [3]. These metadata models are usually focusing on a particular domain. For instance, DC is designed for description of publications in libraries. However, we consider it important to define a metadata model, which could be an integrated metadata model. In other words, a general-used metadata model required is able to interpret and represent other metadata models. We also think that a good metadata model would meet these requirements: simple in presentation and modeling, rich in semantics and expressiveness, easy to use, and flexible to adapt to various local models (such as relational model).

In this paper, we propose a concise metadata model as an intermediate model that can transfer various styles of web models or metadata models used to describe web resources to a more general-used framework. This metadata model, termed as WDM, will take in all the descriptive metadata information associated to and/or contained in web documents created by other metadata models. The metadata model also plays a role of document classification framework to support analysis of web documents. The framework is then realized in a platform, called web document analyzer, consisting of an analysis and classification mechanism and a knowledge and information repository for storing various metadata knowledge and intermediate analysis results.

The paper is organized as follows. In the next section, we present our metadata model for the web document analysis and classification. Then we describe a framework of the analysis method based on the model and discuss our future work in the section 3.

WDM: An Intermediate Metadata Model

This metadata model is developed in the project S-PICS [4] for management of web documents in various formats. An essential goal to introduce the web document metadata model is that we need to model various formats of web data types and functions in a unique way of representation and hence to consider an integral analysis and synthetic use of the web data. Although there are a number of metadata models used in designing web documents, we need a "super" metadata model over them to interactively apply the web data of different forms.

In addition, in our opinions, main requirements that should be met by an intermediate metadata model include 1) easy to understand and to use, 2) capable to represent other metadata models, and 3) well-defined for the web document analysis and classification.

The reason to propose these requirements is: we hope our intermediate metadata model to be general and expressive enough to translate various object types and relations existing in web documents and hence easier to represent them for analysis and classification in terms of our intermediate metadata model. The intermediate metadata model is based on our previous research work in metadata modeling [5].

We consider that WDM consists of two construct types: object type and attribute type. Any information pieces can be seen as instances of Object Type. For example, a web resource is an instance of Object Type. A database relation is another instance of Object Type. A web document is of course an instance of Object Type.

For Attribute Type, we define four sub-types of Attribute Type: Identifier Attribute, Description Attribute, Analysis Attribute, and Reference Attribute. Informally, each sub-type an be defined as follows:

- Identifier Attribute – A unique identifier is specified for an object. E.g., a uri can be an identifier for a web document.
- Description Attribute – A set of attributes define, e.g., who creates the object, when it was created, etc.
- Analysis Attribute – A set of attributes define the semantics, connotation of an object, for example, keywords, labels, etc.

- Reference Attribute – A set of attributes give links to other objects. Examples can be a uri to another web document, a path to a rating service referring to ratings applied by the object, and so on.

A diagram to describe the WDM model is shown in Figure 1.

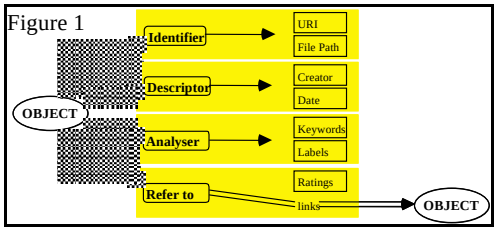

Figure 1

The merit of the WDM model is, first of all, to give a catego-rization to the content items in the current metadata models and the categories defined cover various items commonly used in classifying the web documents. A second advantage of the intermediate metadata model is that it can support management of the web documents, such as filtering, signature, property right setting, etc. Here, we just illustrate a representation of PICS rating service in terms of the WDM model.

A PICS rating service provides a document with filtering mechanism. For example, a document adopting RSAC rating service may use "public" as its label and "3" as its rate. In this document, additional information (metadata) will be supplied for identifying the document itself by stating a rate and its rating service source, e.g. a link to WS-RS (WS Rating Service) by url: /www.sisu.se/~william/rating-bureau/. This document can be described in the WDM model as shown in Figure 2.

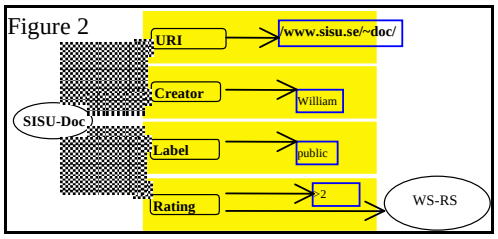

Figure 2

Web Document Analyzer

A major objective to develop the web document analyzer is to explore the semantics of web documents searched in terms of the metadata model WDM and hence to use the semantic information for the document identification, classification and management.

The modeling process in our web document analysis system takes in the web documents required and translates them into a collection of attribute slot sets in terms of the metadata model WDM. Each attribute set has four attribute slots: Identifier slot, Description slot, Analysis slot, and Reference slot. They contain information obtained from one document for next processing step.

These slots containing web document meta-information are then stored in a repository to be used as inputs for the web document analyzer.

The architecture of the web document analyzer based on the WDM metadata model consists of three main components, as shown in Figure 3.

1) Graphical presentation interface – The interface of our analyzer is built on a web browser (currently on Netscape Navigator). It adopts an object-attribute diagram with hierarchical structure.
2) Information and Knowledge Repository – We use the ACCESS database system as a supporting data repository to store the web information to be analyzed by our analyzer. These web documents are stored in the information part of the repository. Also stored in the information part is the intermediate results translated from the original web documents by the analyzer.

   In the knowledge part of the repository we store some pre-defined metadata models and web page authoring languages, like XML and rating services as a complementary analysis information. The web document metadata model, WDM, is also described as a framework for analysis and classification of web documents and its constructing items, like various attributes, are placed in different slots in the framework.

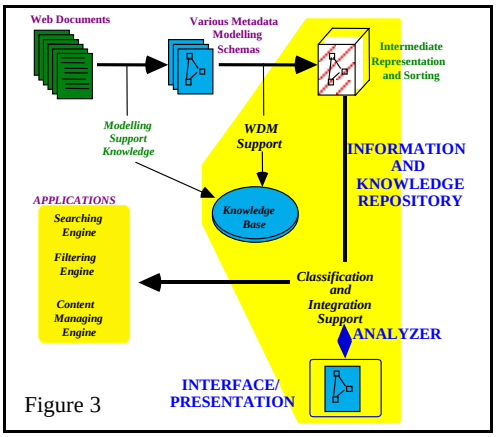

Figure 3

3) Analysis Engine – It is based on an integration engine for conceptual modeling schemas to which we introduced the WDM model as its core. Similarity comparison, analysis, and integration for the web information resources are performed here.

An initial implementation has been done for illustrating the functionality of the WDM model. The implemented prototype could sort out the web documents loaded and filled accordingly the modeling framework slots with all the metadata and descriptive information in terms of the modeling attributes. Problem arises due to the lack of information, which could be obtained from the web document heads. Either the authors of web documents do not supply them or supply them inappropriately. Our next is to improve the framework by providing the document authors with function of modifying their metadata information based on the suggestions from our meta-information analyzer.

We believe that the metadata model WDM has some addition advantages over the other metadata models like Dublin Core, PICS, etc. First, WDM supports he users to create their own specific metadata schemas for their particular purposes, whereas many metadata models are developed already for special aims and limit their application areas. Second, WDM provides a preliminary classification for the users so that they can make web data analysis based on this classification. Third, WDM focuses on its pictorial representation for application domains so it makes modeling and presentation easier and understandable.

References

[1] Metadata and Dublin Core, http://www.ub2.lu.se/tk/metadata/.
[2] Meta Content Framework, http://mcf.research.apple.com/mcf/.
[3] Extensible Markup Language, http://www.w3.org/TR/REC-xml.
[4] Song, W., ***Metadata and PICS Management.*** SISU PUBLIKATION. ***97(05).*** 1997.
[5] Song, W., Web Document Modelling and Clustering, CMMIS Workshop in the Int'l Conf. On Entity-Relationship Approaches, LA., USA., 1997.