SV: a visualization mechanism for ontologies of records based on SVG graphics

Ma. Auxilio Medina, Miriam Cruz, Rebeca Rodríguez, Argelia B. Urbina

Universidad Politécnica de Puebla
Tercer Carril del Ejido Serrano S/N
Juan C. Bonilla, Puebla, México
{mmedina, mcruz, rrodriguez, aurbina} @uppuebla.edu.mx,
WWW home page: http://informatica.uppuebla.edu.mx/
~mmedina, ~rrodriguez, ~aurbina

Abstract. This paper describes SV, a visualization mechanism used to explore digital collections represented as hierarchical structures called ontologies of records. These ontologies are XML files constructed using OAI-PMH records and a clustering algorithm. SV is composed by a web interface and SVG graphics. Through the interface, users can recognize the organization of the collection and access to metadata of documents.

1 Introduction

Digital libraries gather valuable information. Organizations such as the Open Archives Initiative (OAI[1]) have proposed different alternatives to share data. The Protocol for Metadata Harvesting (OAI-PMH protocol), for example, supports interoperability between federated digital libraries. Documents are described in metadata records. Dublin Core Metadata (DC[2]) is the default metadata format for this protocol.

The services and the collections of digital libraries are enriched in the Semantic Web. The use of XML, Resource Description Framework (RDF), OWL, conceptual maps and other metadata technologies are addressed to improve search tasks [1]. Semantic Digital Libraries (SDLs) refer to systems build upon digital libraries and social networking technologies (Web 2.0) [2]. Freely distributed software exists to construct SDLs such as Greenstone[3] or Jerome DL[4]. In this type of software, ontologies play a key role, they refer to explicit specifications of shared conceptualizations [3]. Ontologies enables the representation of knowledge that software and human agents can understand and use.

This paper proposes the use of ontologies called "ontologies of records" that are represented as XML documents as the basis of a visualization mechanism

[1] http://www.openarchives.org/
[2] http://dublincore.org
[3] http://www.greenstone.org/
[4] http://www.jeromedl.org/

called *semantic view (SV)*. The name also refers to the first two letters of "Support Vector Graphics". SV offers an interactive view to allow users to explore the content of a federated collection.

The paper is organized as follows. Section 2 describes the features of an ontology of records. Section 3 includes related work. Section 4 and 5 explains the design and implementation of SV, respectively. Experimental results are described in Section 6. Finally, Section 7 includes conclusions and suggests future directions of our work.

2 What is an ontology of records

An *ontology of records* is a hierarchical structure of clusters of OAI-PMH records that provides an unambiguous interpretation of its elements. Its construction is based on the *Frequent Itemset Hierarchical Clustering* algorithm [8]. This structure organizes a collection of documents, this has concept-term relationships useful for keyword based searches. An ontology of records is stored as a well formed XML file that is validated against an XML Schema. An ontology of records has the following features[9]:

1. Documents are clustered by similarity
2. Clusters in the k-level have labels of k-terms
3. All the records of a cluster share the terms of its label

3 Related work

This section describes some systems that have been used to visualize collections of documents. Proat et al. [4] use 3D trees to visualize documents organized according to the Library of Congress Classification (LCC). Documents are clustered in seven subsets. The interface has controls to rotate or zoom the nodes of trees. The leaf nodes contain metadata of documents.

Geroimenko et al. [5] have proposed the Generalizad Document Object Model tree Interface (G-DOM-Tree interface) to visualize metadata from XML DOM (Document Object Model) documents. The model displays a hierarchy of labels, this is very similar to the visualization that browsers offer of XML Schema. The interface is implemented as a Java applet or a Flash film.

Fluit et al. [6] describe *Spectacle*, this mechanism uses lightweight ontologies to represent classes of similar objects and their relationships. The navigation can be done by using hypertext or "cluster maps". A cluster map visualizes the objects and their classes.

At last, Sánchez et al. [7] use a star field grid to visualize documents from several collections. Documents are stored as OAI-PMH records. The axis of the grid represent attributes of the collections that can be chosen by users. Small polygons are associated with the type of document and different colors are used to distinguish the collections.

4 Desing of SV

The design of SV is addressed to reach the following objectives:

- Construct a visualization mechanism with semantic features that allow users to explore a collection of documents
- Represent the organization of a collection of documents
- Retrieve the metadata and the content of a determined document

In order to reach these objectives, we have used the levels of knowledge proposed by [2] in the design of SV. We want to uses CORTUPP as a test bed, this is a collection represented as an ontology of records[5].

1. **Level 1: Organization of the metadata**. Metadata is organized in the ontology of records. Content information is stored in `dc:title`, `dc:subject` and `dc:description` elements.
2. **Level 2: Organization of the information in the documents**. Technical reports have a common structure formed by six mandatory chapters: 1)research propose, 2)state of the art, 3)research design, 4)implementation, 5)results and 6)conclusions. This structure is defined in a Latex template. The BibTex file format is used to manage the bibliography. A technical report is described as a `@techreport` entry.
3. **Level 3: Organization of the information in databases**. The technical reports are stored as PDF files in a database that also includes data and counts of users. Documents are accessible through a web interface.
4. **Level 4: Organization of the topics treated in the documents**. The `dc:subject` element stores the topic of a document. Keywords of this element belong to the labels of the clusters in the ontology of records.
5. **Level 5: Organization of the concepts, terms and relations**. This level is also represented in the ontology of records.

5 Implementation of SV

SV is formed by a web interface and SVG graphics[6]. SVG is a format developed and maintained by the W3C SVG Working Group. This is an XML application used to describe animated or static two dimensional vectorial graphics. The main feature of these graphics is scalability.

SV uses Xerces, this is a Java parser used to extract data from an ontology of records. The classes of SV are built using Java language. In the interface, each document, that is, an OAI-PMH record, is represented with a yellow star in a blue gradient background. The background is divided in five parts that correspond to the first levels of the ontology. These levels are divided by lines that form angles of 90 degrees. The distribution of the lines try to reflect an estimation of the amount of documents that can be found in each level. The documents closer to

[5] CORTUPP is available at http://server3.uppuebla.edu.mx/cortupp/

[6] http://www.w3.org/svg/

the upper left corner belong to the first level of the ontology, these documents share one term. The second level shows the documents that share two terms, and then on. The stars have different size according to their level, they are bigger at the first level and smaller at the last one.

The interface of SV is a SVG graphic of 502 per 502 pixels. XML Parser is the Java application used to construct the XML document that contains the interface. XLink is used to create hyperlinks between documents and their metadata. Given a click on a star, users can allow the metadata on the right panel. Figure 1 shows the SV interface where only six documents at the second and third level were included, however SV is designed to support until 500 documents. The colors can be modified without requiring compilation because they are stored in a text file. The mechanism is accessible at http://informatica.uppuebla.edu.mx/ visualizacionPI/index.html.

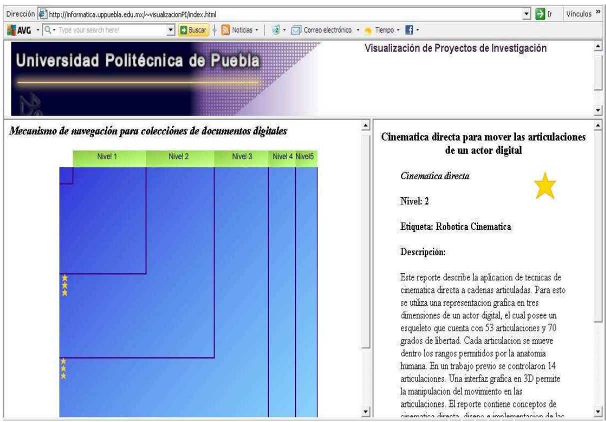

Fig. 1. Using VS to visualize CORTUPP

6 Experimental results

Different configuration of ontologies of records were constructed in order to check SV, that is, unity tests and integration tests were performed successfully. After the installation of the SVG Plugin Version 1.7, the visualization of SV was successful using Internet Explorer 8, Google Chrome 7.0.517.41 and Opera 10.6, however, there were some inconveniences using Firefox 1.5, Firefox 3.6 and Firefox Beta due to these versions do not support the animation features of SVG graphics.

7 Conclusions

We have described SV, a visualization mechanism of federated collections based on ontologies. SV has semantic features represented in the interface such as the location of documents in the ontology and the similarity between documents. Additional semantic information is stored in the metadata attached to each document and in the ontology of records. Through SV interface, users can access to metadata or download a document.

CORTUPP was used as a test bed for SV, however, any collection of OAI-PMH records represented as an ontology of records can be visualized. Although the size of an ontology of records can impact the visualization of SV, its design is flexible enough to support distinct collections. As future work, we plan to expand SV to show the clusters and their labels. Then, we would like to incorporate tagging and recommendation mechanisms.

References

1. Geroimenko V., C.C.: Visualizing the Semantic Web. XML-based Internet and Information Visualization. Segunda edición edn. Volume 1. Springer, Wokingham, England (2003) Libro. los primeros cuatro datos son obligatorios.
2. Kruk, S.R., McDaniel, B.: Semantic Digital Libraries. Springer-Verlag, Berlin, Heidelberg (2009)
3. Gruber, T.: A translation approach to portable ontology specification. Knowledge Adquisition **5**(2) (1993) 199–220
4. C., P.: Sistema uva: interfaces para visualización de grandes colecciones digitales. Tesis de maestría, Universidad de las Américas Puebla, Santa Catarina Mártir S/N, San Andrés Cholula, Puebla, México. (2002) Tesis de maestría. Los primeros cuatro campos son obligatorios.
5. Geroimenko V., G.L.: Interactive interfaces for mapping e-commerce ontologies in Visualizing the Semantic Web. XML-based Internet and Information Visualization. Segunda edición edn. Volume 1. Springer, Wokingham, England (2003) Libro. los primeros cuatro datos son obligatorios.
6. Fluit C., Sabou N., H.v.F.: Ontology-based information visualization. In: Visualizing the semantic web: XML based internet and information visualization. Volume 1. Segunda edición edn. Springer, Wokingham, England (2002)
7. Sánchez J. A., Quintana M. G., R.A.: Star-fish: Starfields+fisheye visualization and its application to federated digital libraries. Proceedings of the 3rd Latin American Conference on Human-Computer Interaction (CLIHC 2007, Nov.) (2007)
8. Fung, B., Wang, K., Ester, M.: Hierarchical document clustering using frequent itemsets. In: Proceedings of the Third SIAM International Conference on Data Mining, (SDM'03, San Francisco, California, May),, San Francisco, CA, USA, SIAM (2003) 59–70
9. Medina, M.A., Sánchez, J.A.: Ontoair: A method to construct lightweight ontologies from document collections. Mexican International Conference on Computer Science **0** (2008) 115–125