Digital Library Portal using Semantic Tools in WWWPal

John R. Punin and Mukkai S. Krishnamoorthy
Department of Computer Science,
Rensselaer Polytechnic Institute, Troy, NY 12180, USA
{puninj,moorthy}@cs.rpi.edu

Abstract

The WWWPal system and associated languages and tools (such as LOGML, XGMML, webbot and the graph browser) have been developed to perform syntactic analysis of web sites. In this paper, using WWWPal and semantic analysis tools (such as RGML, clustering and the graph browser) we construct digital library portals. We describe a method of obtaining the portal.

Introduction

Semantic web was introduced [BERNERS] to make the tangled information in the web more accessible to search engines and other applications. The semantic web is not a separate web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation [BERNERS01]. Specifying semantic information to the web content will make this task easier. On the other hand, it may be harder for the content developer to provide the semantic information that the user agent may want. Modifying the existing web pages with semantic information may result in additional errors. W3 Consortium arrived at a solution of specifying the meaning of a web resource using Resource Description Framework (RDF)[RDF]. RDF encodes the metadata in sets of triples, each triple being rather like the subject, verb and an object of an elementary sentence. A number of papers [SEMWEB01] have been published about RDF, vocabularies based on RDF and RDF applications.

In this paper we develop a general purpose framework, a collection of semantic tools using RDF, which can be applied for digital library web portals. Designing a portal for a digital library, a collection of information that can be browsed and searched by search engines and humans, is a simpler task with this framework.

WWWPAL Support for Digital Library Portals

WWWPal provides support for the Digital Library Research by collecting, filtering, and classifying the available metadata of a web site. In a digital library, there are two standard models of delivering information. This information is either static (i.e., supplied by a librarian with a keyword classification) or dynamic (i.e., the system tries to obtain a keyword using heuristics). WWWPal provides both the static support (by providing an RDF editor) and the dynamic support (by providing clustering and keyword classification). The web robot of WWWPal navigates a web site, saves the structure of the web site and collects the metadata information of the web site. The structure of the web site is saved in an XGMML (XML vocabulary for graphs) document [LOGML]. The metadata of the web pages and hyper links is appended to the nodes (web pages) and edges (hyperlinks) using the RDF/XML serialization. The XGMML document is transformed into an RGML (RDF vocabulary to describe graphs) document [RGML]. This RDF document is read by an RDF parser to produce a set of triples. Further, we can group several web documents by finding the clusters of the webgraph. These clusters are represented as subgraphs of the webgraph, and the metadata of each node of the subgraph is merged to form the metadata of the cluster. All of these subgraphs are saved in an RGML file. We have also developed a simple web portal so users can browse and search the information gathered in this repository.

Metadata Collection

The metadata collection is achieved using the web robot of WWWPal [WWWPAL]. The web robot navigates a web site using a breadth first search algorithm. Each visited web page is parsed to find the following metadata information:

- Title of the web page in the <TITLE> tag

- Metadata information in the <META> tag
- Metadata linked to the web page using the <LINK> tag
- Anchor text information in the <A> tag
- Headers of the page in the <H1> to <H6> tags

This metadata information may not be sufficient to describe with a web page. Our experiments suggested that we get most of the metadata information from the above five cases. All the metadata information of the web page is attached to the node of the graph that represents the web page. We use RDF to represent the metadata information and XGMML to save the whole webgraph structure. The following example is an XGMML document using RDF vocabulary to represent metadata. The collected metadata is title, date, format and keywords. We have used the Dublin Core vocabulary [DC] to represent these RDF properties of the web page. Figure 1 shows the structure of this webgraph using the WWWPal Graph Browser.

```
<?xml version="1.0"?>
<graph xmlns = "http://www.cs.rpi.edu/XGMML"
       directed="1" >
<node id="3" label="http://www.cs.rpi.edu/courses/" weight="6968">
<att>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.0/">
  <rdf:Description about="http://www.cs.rpi.edu/courses/"
    dc:title="Courses at Rensselaer Computer Science Department"
    dc:subject="www@cs.rpi.edu; M.S. requirements; CSCI-1190 Beginning C Programming for Engineers; Courses; People;
    Graduate Program; CSCI-4020 Computer  Algorithms; CSCI-2220-01  Programming in Java; Research; Course Selection Guide;
    CSCI-4961-01,  CSCI-6961-01 Advanced Robotics; Programming in Java; CSCI-2400 Models  of Computation"
    dc:date="2000-01-31"
    dc:type="Text"
    >
    <dc:format>
      <rdf:Bag
        rdf:_1="text/html"
        rdf:_2="6968 bytes"
      />
    </dc:format>
  </rdf:Description>
</rdf:RDF>
</att>
</node>
<node id="7" label="http://www.cs.rpi.edu/research/" weight="13732">
<att>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.0/">
  <rdf:Description about="http://www.cs.rpi.edu/research/"
    dc:title="Research at Rensselaer Computer Science
    Department"
    dc:subject="www@cs.rpi.edu; Computing  Twin Primes and Events; TEMPEST; Courses; People; Graduate Program;
    High-Performance  Object-Oriented Programming in Fortran 90;
    High Performance Problem-Solving Environment for Optimization and Control of  Chemical and Biological Processes;
    Computer Vision; Theory and Algorithms; technical report library; RPInfo; Undergraduate  Program;
    Research; Research; Design Conference Room; Rensselaer; Bryan Rudge; Engineering Databases; anonymous ftp; I.SEE;
    info@cs.rpi.edu; Scientific Computing; OpenMath; Proactive Network Problem Avoidance; Computing Facilities; Computer Science Departm
    dc:date="1999-11-19"
    dc:type="Text"
    >
    <dc:format>
      <rdf:Bag
        rdf:_1="text/html"
        rdf:_2="13732 bytes"
      />
    </dc:format>
  </rdf:Description>
</rdf:RDF>
</att>
</node>
<node id="8" label="http://www.cs.rpi.edu/undergrad/" weight="7672">
<att>
<rdf:RDF
```

```
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.0/">
  <rdf:Description about="http://www.cs.rpi.edu/undergrad/"
    dc:title="Undergraduate Program at Rensselaer Computer Science Department"
    dc:subject="www@cs.rpi.edu; Association for Computing Machinery; People; Graduate Program; Minor; Prospective Students FAQ;
    Admissions Office; Research; BS-MS Degree; Rensselaer Catalog; Dual Majors; Admissions;Computing Facilities; Undergraduate Program;
    Rensselaer; Course Descriptions; Computer Science Department"
    dc:date="2000-01-26"
    dc:type="Text"
    >
    <dc:format>
      <rdf:Bag
        rdf:_1="text/html"
        rdf:_2="7672 bytes"
      />
    </dc:format>
  </rdf:Description>
</rdf:RDF>
</att>
</node>
<node id="1" label="http://www.cs.rpi.edu/" weight="3402">
<att>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.0/">
  <rdf:Description about="http://www.cs.rpi.edu/"
    dc:title="Rensselaer Computer Science Department"
    dc:subject="www@cs.rpi.edu; faculty positions; Rensselaer; Graduate Program; info@cs.rpi.edu; Current Events;
    RPInfo; Research; Undergraduate  Program"
    dc:date="2000-01-26"
    dc:type="Text"
    >
    <dc:format>
      <rdf:Bag
        rdf:_1="text/html"
        rdf:_2="3402 bytes"
      />
    </dc:format>
  </rdf:Description>
</rdf:RDF>
</att>
</node>
<edge source="1" target="3" weight="0" label="SRC IMG gfx/courses2.jpg" />
<edge source="7" target="3" weight="0" label="SRC IMG ../gfx/courses2.jpg" />
<edge source="8" target="3" weight="0" label="SRC IMG ../gfx/courses2.jpg" />
<edge source="3" target="7" weight="0" label="SRC IMG ../../gfx/research2.jpg" /
<edge source="1" target="7" weight="0" label="SRC IMG gfx/research2.jpg" />
<edge source="8" target="7" weight="0" label="SRC IMG ../gfx/research2.jpg" />
<edge source="3" target="8" weight="0" label="SRC IMG ../../gfx/ugrad2.jpg" />
<edge source="7" target="8" weight="0" label="SRC IMG ../gfx/ugrad2.jpg" />
<edge source="1" target="8" weight="0" label="SRC IMG gfx/ugrad2.jpg" />
<edge source="3" target="1" weight="0" label="SRC IMG ../../gfx/corner2.jpg" />
<edge source="7" target="1" weight="0" label="SRC IMG ../gfx/corner2.jpg" />
<edge source="8" target="1" weight="0" label="SRC IMG ../gfx/corner2.jpg" />
</graph>
```

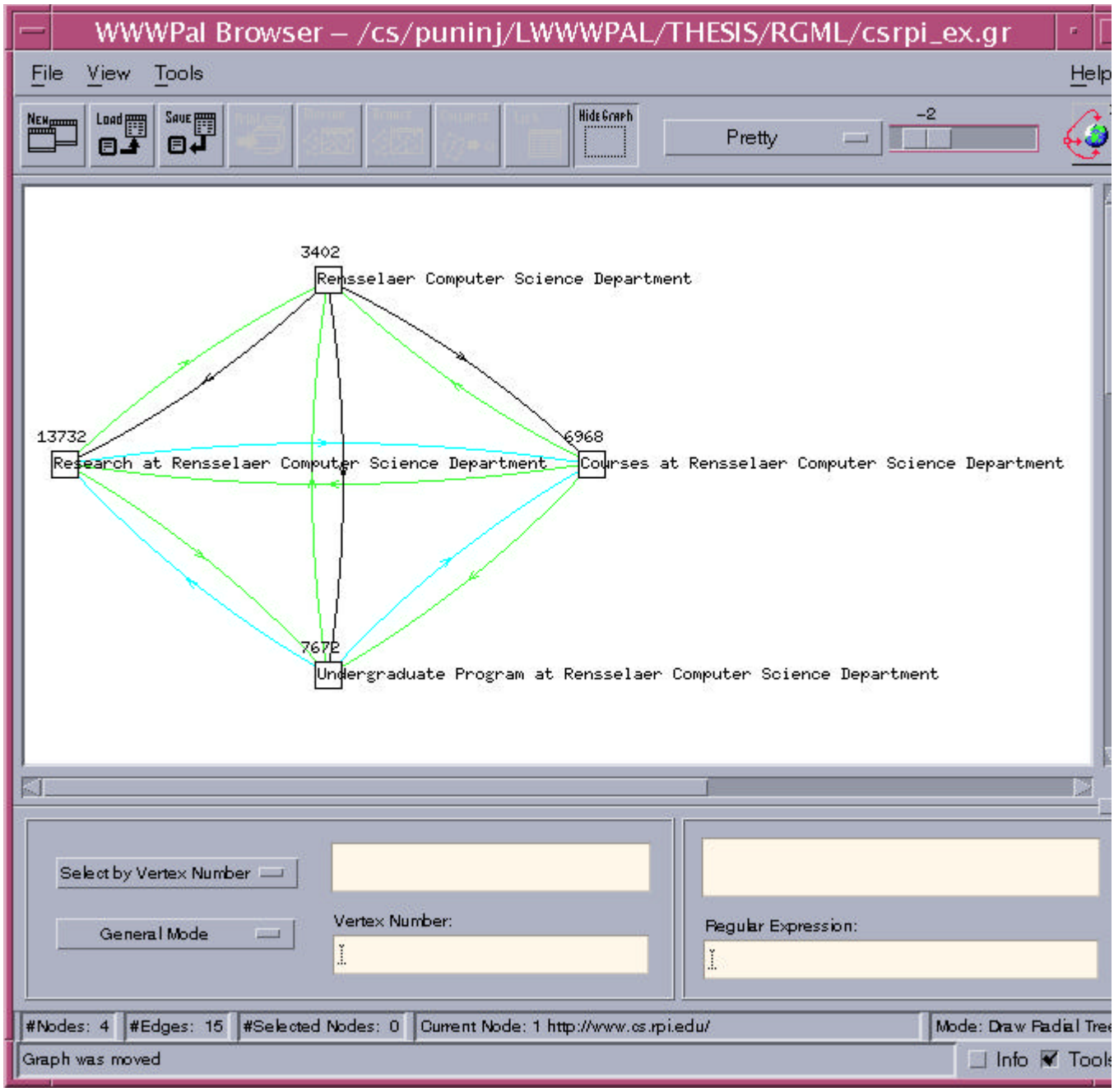

Figure 1: Webgraph of the main web pages of the RPI Department of Computer Science.

Metadata Filtering

RDF has been conceived to represent semantic information. We use RGML to save the metadata of a given web site . One of the WWWPal modules transforms the XGMML document into an RGML document. The RGML vocabulary follows the RDF syntax and makes it simple to combine different RDF vocabularies such as Dublin Core [DC] and Vcard [VCARD]. The generated RGML file can be read by any RDF parser. The parser generates a set of triples which are the RDF statements: subject, predicate and object. Before parsing the RGML file, it is important to cluster the webgraph. The generated clusters provide metadata information such as keywords. Clusters are subgraphs of the webgraph and hence RGML can represent them by using the *graphs* property. WWWPal has implemented several clustering methods and they are fully explained in [WWWPAL].

Digital Library Portals

Popular web portals such as Yahoo and Netscape classify and present important information such as news, weather and entertainment in just one web page so the user does not have to spend too much time in finding valuable information. An educational web portal has been described in [JASIG]. We know that most of the educational web sites do not offer portals and hence it is difficult to find information. WWWPal provides a web interface to browse and search a repository of RGML documents. Each RGML document contains the information of an educational web site such as Computer Science Department sites. The user can visit this web interface and quickly find the specific information that they are looking for. We consider this web interface as a simple example of a Digital Library Portal where the RGML documents provide semantic information. Future development will add an inference engine and several rules to construct a powerful knowledge base so that the Digital Library supports semantic search and presentation of knowledge.

References

[BERNERS] T. Berners-Lee, *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*, San Francisco, CA, HarperCollins 1999

[BERNERS01] T. Berners-Lee, J. Hendler, and O. Lassila, *The Semantic Web*, Scientific American, May 2001.

[DC] S. Weibel, J.Kunze, C. Lagoze, and M. Wolf. *Dublin Core Metadata for Resource Discovery*, Internet RFC 2413. 1998.

[JASIG] JA-SIG Portal Framework Project, www.mis2.udel.edu/ ja-sig/portal.html; Bernard W. Gleason, Boston College University-Wide Information Portal: Concepts and Recommend ed Course of Action, Jan. 26, 2000.

[LOGML] J. Punin, M. Krishnamoorthy, M. J. Zaki, *Web Usage Mining: Languages and Algorithms*, to appear in Studies in Classification, Data Analysis, and Knowledge Organization, Springer-Verlag, 2001.

[RGML] J. Punin and M. Krishnamoorthy, *Describing Structure and Semantics of Graphs Using a RDF Vocabulary*, to appear at Extreme Markup Languages 2001.

[RDF] O. Lassila, and R. Swick, *RDF Model and Syntax Specification W3C Recommendation* , February 1999.

[SEMWEB01] *The Second International Workshop on the Semantic Web*, April 2001.

[VCARD] R. Iannella, *Representing vCard Objects in RDF/XML,* W3C Note February 2001.

[WWWPAL] J. Punin, M. Krishnamoorthy. *WWWPal System - A System for Analysis and Synthesis of Web Pages*. In Proceedings of the WebNet 98 Conference, Orlando, November, 1998.