Automatically Generated Topic Maps of World Wide Web Resources

Carol Jean Godby, Eric Miller, Ray Reighart
OCLC Online Computer Library Center
6565 Frantz Road
Dublin, Ohio, USA, 43017
{godby, miller, reighart} @ oclc.org

Abstract

At the world's major center for research and development in library automation, we are interested in technology that organizes, standardizes, and facilitates access to electronic text. Currently, the World Wide Web affords unprecedented access to globally distributed, primarily textual, information, but users quickly learn that sites of interest can be difficult to find without expert guidance or a measure of luck. Though search engines have been available for five or six years, other methods for discovering information are just as popular, such as personal recommendations or collections of links from trusted sources. But the reliance on serendipitous discovery strategies is odd because the information needs of an individual user are often confined to a single subject or closely related subjects, as evidenced by the growing number of personal Web pages that collect links on such diverse subjects as computational linguistics, freight trains, Japanese art, calligraphy, computer viruses, the worldwide disappearacne of frogs, or noninvasive breast cancer. This behavior suggests both a problem and an opportunity. The problem is that subject access to Web information is still primitive, at best. But many users are taking the initiative to organize portions of the Web themselves, making it is possible to leverage their work with the help of appropriately designed software.

To address this problem, OCLC and the World Wide Web Consortium are sponsoring the development of the Resource Description Framework (RDF), an international standard for an infrastructure that enables the encoding, exchange and reuse of descriptive data about Web resources. At the core of RDF is a syntax-independent model for describing resources, properties associated with these resources, and relationships among resources. RDF enables groups that manage data collections to define the vocabularies for describing resources (e.g. *author*, *title*, *subject* for Web documents; or *name*, *address*, *hair-color* for people, etc.) and to build collections of these properties. Figure 1 illustrates a simple RDF description of a Web page that has properties including a *title* and an *author.* The *author* property, in turn, has the properties *name*, *email* and *affiliation*. Since RDF is a generic data model, a syntax is required for communicating this model. RDF utilizes XML eXtensible Markup Language as a syntax for the transmission of this model among Web-based applications. See Miller (1998) for additional information regarding this data model and its corresponding syntactic representation.

```
<?xml version = '1.0'?>
<rdf:RDF xmlns:rdf = "http://www.w3.org/TR/rec-
rdf-syntax#">
 <rdf:Description about = "http://uri-of-Document-1">
     <Author><name> John Smith </name>
     <affiliation> Home, Inc.  </affiliation>
     <email> smith@home.com </email></Author>
     <title> The Smith family tree </title>
  </rdf:Description>
</rdf:RDF>
```

Figure 1. XML code for a simple RDF graph.

If collections of Web resources are described as in Figure 1, it is possible to create generic tools for the navigation of this graph based on particular properties. To aid in the creation of such navigation tools, Netscape (Note 1) has developed an engine that produces a memory-resident acyclic graph from collections of RDF descriptions. The content of the graph and the details of the display are left to the application developer, but the RDF engine provides an easy way to create meta-indexes of Web pages that are now laboriously created by hand. For example, one possible input to an application using the RDF engine is a bookmark file. If application software can automatically extract authors, titles and URLs from the Web pages referenced in a bookmark file, browseable author-title displays can be created with no human input.

In our demo, an early research prototype, we show how RDF graphs can be enhanced to include subject access through terminology automatically extracted from Web pages. This work is a natural extension of the research reported by Ibekwe-San Juan (1998), which showed that syntactic term variants such as *root hair, deformed root hair,* and *root hair deformation* form a graph structure that produces an effective schematic representation of the growth and development of topics in a corpus of academic research articles in plant biology. The broad goals of our research program in terminology identification are to identify high-quality subject terminology in ways that are fully automated, computationally efficient, conceptually simple, and generic; to impose reasonable organizations on this terminology; and to embed this terminology into useful applications that enhance access to electronic text.

Our terminology identification system is conceptually similar to that described by Daille (1994) and implemented in Java as a set of filters that perform three steps. First, the data is preprocessed with association statistics to identify ngrams most likely to contain subject vocabulary. The ngrams serve as input to the second step, a partial parser that extracts noun phrases according to patterns specified by the software application. Finally, the set of noun phrases is filtered with heuristics like those described in Nakagawa and Kori (1998) and Wacholder (1998) that identify subject-rich compound nominals as well as atomic nouns and can be used to create simple hierarchical or graphical structures based on syntactic similarity. Our system, described more fully in Godby and Reighart (1999), has many configurable parameters, but it can result in drastic reductions of the input text. The final terminology graph may represent as little as .0003% of the original file, a good result considering our goal of creating relatively sparse terminology maps.

Figure 2 shows a portion of an RDF graph populated with automatically extracted subject terminology. In this example, *electronic library* has two narrower terms, *electronic library model* and *electronic library model implementation*. The RDF graph links the terms, identifies their relationships, and clusters the pointers to the resources that contain them.

```
<Topic rdf:about = "electronic_library">
  <name>electronic library</name>
  <narrow rdf:resource = "electronic_library_model">
  <isDefinedIn rdf:resource = "http://r1"/>
  <isDefinedIn rdf:resource = "http://r2"/>
  <isDefinedIn rdf:resource = "http://r3"/>
</Topic>
...
<Topic rdf:about = "electronic_library_model">
  <name>electronic library model</name>
  <narrow rdf:resource =
"electronic_library_model_implementation"/>
  <broad rdf:resource = "electronic_library"/>
  <isDefinedIn rdf:resource = "http://r3"/>
</Topic>
```

Figure 3. An RDF graph with subject vocabulary.

Note 1

Accessible at: http://www.mozilla.org/rdf/doc

References

Daille, B. (1994). Study and implementation of combined techniques for automatic extraction of terminology. *Proceedings from the Balancing Act*, the ACL workshop held on July 1, in Las Cruces, New Mexico, 29-36.

Godby, C. and Reighart, R. (1999). The WordSmith Indexing System. *The Annual Review of OCLC Research*. http://cweb.oclc.org/wip/review98/wordsmith.htm

Ibekwe-San Juan, F. (1998). Building a prototype system for trends survey through the use of syntactic term variants. *Computerm '98: Proceedings from the First Workshop on Computational Terminology*. Montreal, Quebec, Canada: COLING-ACL. 22-28.

Miller, E. (1998). An introduction to the Resource Description Framework. *D-Lib Magazine*. May 1998. http://www.dlib.org/dlib/may98/miller/05miller.html

Nakagawa, H. and Kori, T. (1998). Nested collocation and compound noun for term extraction. *Computerm '98: Proceedings from the First Workshop on Computational Terminology*. Montreal, Quebec, Canada: COLING-ACL. 64-70.

Wacholder, N. (1998). Simplex NPs clustered by head: a method for identifying significant topics within a document. *The Computational Treatment of Nominals: Proceedings of the Workshop*. Montreal, Quebec, Canada: COLING-ACL. 70-79.