Deep Web Mining through Web Services

Monica Maceli
Drexel University, USA

Min Song
New Jersey Institute of Technology & Temple University, USA

INTRODUCTION

With the increase in Web-based databases and dynamically-generated Web pages, the concept of the "deep Web" has arisen. The deep Web refers to Web content that, while it may be freely and publicly accessible, is stored, queried, and retrieved through a database and one or more search interfaces, rendering the Web content largely hidden from conventional search and spidering techniques. These methods are adapted to a more static model of the "surface Web", or series of static, linked Web pages. The amount of deep Web data is truly staggering; a July 2000 study claimed 550 billion documents (Bergman, 2000), while a September 2004 study estimated 450,000 deep Web databases (Chang, He, Li, Patel, & Zhang, 2004).

In pursuit of a truly searchable Web, it comes as no surprise that the deep Web is an important and increasingly studied area of research in the field of Web mining. The challenges include issues such as new crawling and Web mining techniques, query translation across multiple target databases, and the integration and discovery of often quite disparate interfaces and database structures (He, Chang, & Han, 2004; He, Zhang, & Chang, 2004; Liddle, Yau, & Embley, 2002; Zhang, He, & Chang, 2004).

Similarly, as the Web platform continues to evolve to support applications more complex than the simple transfer of HTML documents over HTTP, there is a strong need for the interoperability of applications and data across a variety of platforms. From the client perspective, there is the need to encapsulate these interactions out of view of the end user (Balke & Wagner, 2004). Web services provide a robust, scalable and increasingly commonplace solution to these needs.

As identified in earlier research efforts, due to the inherent nature of the deep Web, dynamic and ad hoc information retrieval becomes a requirement for mining such sources (Chang, He, & Zhang, 2004; Chang, He, Li, Patel, & Zhang, 2004). The platform and program-agnostic nature of Web services, combined with the power and simplicity of HTTP transport, makes Web services an ideal technique for application to the field of deep Web mining. We have identified, and will explore, specific areas in which Web services can offer solutions in the realm of deep Web mining, particularly when serving the need for dynamic, ad-hoc information gathering.

BACKGROUND

Web Services

In the distributed computing environment of the internet, Web services provide for application-to-application interaction through a set of standards and protocols that are agnostic to vendor, platform and language (W3C, 2004). First developed in the late 1990s (with the first version of SOAP being submitted to the W3C in 2000), Web services are an XML-based framework for passing messages between software applications (Haas, 2003). Web services operate on a request/response paradigm, with messages being transmitted back and forth between applications using the common standards and protocols of HTTP, eXtensible Markup Language (XML), Simple Object Access Protocol (SOAP), and Web Services Description Language (WSDL) (W3C, 2004). Web services are currently used in many contexts, with a common function being to facilitate inter-application communication between the large number of vendors, customers, and partners that interact with today's complex organizations (Nandigam, Gudivada, & Kalavala, 2005). A simple Web service is illustrated by the below diagram; a Web service provider (consisting of a Web server connecting to a database server) exposes an XML-based API (Application Programming Interface) to a catalog application. The application manipulates the data (in this example, results of a query on a collection of books) to serve both the needs of an end user, and those of other applications.

Semantic Web Vision

Web services are considered an integral part of the semantic Web vision, which consists of Web content described through markup in order to become fully machine-interpretable and processable. The existing Web is endowed with only minimal semantics; the semantic Web movement endeavors to enhance and increase this semantic data. Additionally, the semantic Web will provide great strides in overcoming the issues of language complexity and ambiguity that currently inhibit effective machine processing. The projected result will be a more useful Web where information can be intelligently shared, combined, and identified.

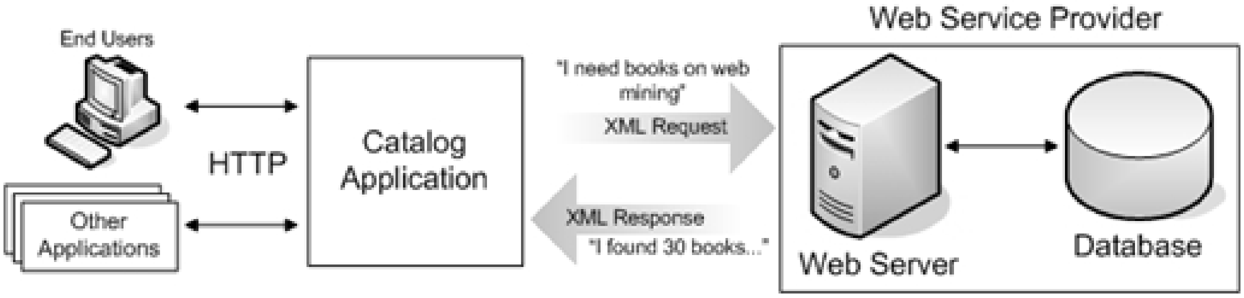
Semantic Web services are Web services combined with ontologies (high-level metadata) that describe Web service content, capabilities and properties, effectively merging the technical strengths of Web services with the descriptive capacity of ontologies (Narayanan & McIlraith, 2002; Terziyan & Kononenko, 2003). Ontologies are vital to the concept of the semantic Web, describing and defining the relationships and concepts which allow for interoperability. Not surprisingly, there has been a great deal of recent research exploring topics such as how to construct, model, use, personalize, maintain, classify, and transform the semantic Web ontologies (Acuña & Marcos, 2006; Alani, 2006; Jiang & Tan, 2006; Lei, 2005; Pathak & Koul, 2005). Ultimately, by describing both what the Web service provides and how to interact with it in a machine-readable fashion, semantic Web services will allow automation in such areas as Web service discovery, integration and interoperation.

Deep Web Challenges

As identified in earlier works (Chang, He, & Zhang, 2004; Chang, He, Li, Patel, & Zhang, 2004), dynamic and ad-hoc integration will be a requirement for large-scale efforts to mine the deep Web. Current deep Web mining research is exploring Web query interface integration, which allows for the querying of multiple hidden Web databases through a unified interface (Bergholz & Chidlovskii, 2004; He, Chang, & Han, 2004; Liu & Chen-Chuan-Chang, 2004). The below figure illustrates the current challenges of the distributed Web environment. The Web consists of both static Web pages and database-specific interfaces; while static pages are easily mined, the Web databases are hidden behind one or more dynamic querying interfaces, presenting an obstacle to mining efforts.

These studies exploit such techniques as interface extraction, schema matching, and interface unification to integrate the variations in database schema and display into a meaningful and useful service for the user (He, Zhang, & Chang, 2004). Interface extraction seeks to automatically extract the relevant attributes from the HTML of the query interface Web page, schema matching identifies semantic similarity between these attributes, and interface unification refers to the construction of a single unified interface based upon the identified matches (He, Zhang, & Chang, 2004). Such unified interfaces then serve as a mediator between the user and the multiple deep Web databases that are being queried; the request and aggregation of this data

Figure 1. High-level Web service example


End Users
Other Applications
HTTP
Catalog Application
"I need books on web mining"
XML Request
XML Response
"I found 30 books..."
Web Service Provider
Web Server
Database