IGI PUBLISHING **ITJ3831**

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.igi-pub.com

This paper appears in the publication, International Journal of Web Services Research Volume 4, Issue 3
edited by **Liang-Jie Zhang © 2007, IGI Global**

Service Class Driven Dynamic Data Source Discovery with DynaBot[1]

Daniel Rocco, University of West Georgia, USA

James Caverlee, Georgia Institute of Technology, USA

Ling Liu, Georgia Institute of Technology, USA

Terence Critchlow, Lawrence Livermore National Laboratory, USA

ABSTRACT

Dynamic Web data sources on the Deep Web provide intuitive access to real-time information and large data repositories anywhere that Web access is available. Although recent studies suggest that the dynamic Web is larger and growing faster than static Web, dynamic content is often ignored by existing search engine indexers owing to technical challenges inherent in searching dynamic sources. To address these challenges, we present DynaBot, a service-centric crawler for discovering and clustering Deep Web sources. DynaBot has three unique characteristics. First, DynaBot utilizes a service class model implemented through the construction of service class descriptions (SCDs). Second, DynaBot employs a modular architecture for focused crawling of the Deep Web. Third, DynaBot incorporates algorithms for efficiently probing, discovering, and clustering Deep Web sources through SCD-based service analysis. Experimental results demonstrate DynaBot's effectiveness and suggest techniques for efficiently managing service discovery given the immense scale of the Deep Web.

Keywords: deep Web; dynamic Web data; service discovery, Web crawling

INTRODUCTION

The World Wide Web is the product of two unique approaches to document publication. The traditional or "static" Web consists of documents materialized in the secondary storage of server systems that are hyperlinked to other Web documents. These documents are generally accessible to unauthenticated users and automated agents like search engine crawlers. The dynamic or "Deep Web," in contrast, refers to the dynamic collection of Web documents that are created as a direct response to some user query. Deep Web services provide access to real-time information, like entertainment event listings,

or present a Web interface to large databases or other data repositories. Recent studies suggest that the size and growth rate of the dynamic Web greatly exceed that of the static Web (Lawrence & Giles, 1998, 1999). Estimates suggest that the practical size of the Deep Web may be greater than 550 billion individual documents (Bergman, 2003). More than half of the content of the Deep Web resides in topic-specific databases, many of which are made available through Web services. A full 95% of the Deep Web is publicly accessible information that is not subject to fees or subscriptions.

Dynamic content is often ignored by existing search engine indexers owing to the technical challenges that arise when attempting to search the Deep Web. The most significant challenge is the philosophical difference between the static and Deep Web with respect to how data are stored: in the static Web, data are stored in documents while in the dynamic Web, data are stored in databases or produced as the result of a computation. This difference is fundamental and implies that traditional document indexing techniques, which have been applied with extraordinary success on the static Web, are inappropriate for the Deep Web. Related to the data storage issue is the problem of data retrieval, since static Web documents are retrieved via simple HTTP calls while dynamic Web documents often reside behind form interfaces that are impenetrable to traditional crawlers. Finally, Deep Web sources tend to be more domain focused than their static Web counterparts. While there is much to be gained from discovering and clustering Deep Web sources, any significant exploration of the Deep Web will require techniques that exploit service-oriented functionality through intelligent analysis of search forms and result samples.

With these challenges in mind, we present DynaBot, a service-centric crawler for discovering and clustering Deep Web sources. DynaBot has three unique characteristics. First, DynaBot utilizes a service class model of the Web implemented through the construction of service class descriptions (SCDs). Second, DynaBot employs a modular, self-tuning system architecture for focused crawling of the Deep Web. Third, DynaBot incorporates methods and algorithms for efficient probing of the Deep Web and for discovering and clustering Deep Web sources and services through SCD-based service matching analysis.

We demonstrate the capability of DynaBot through the BLAST (**B**asic **L**ocal **Al**ignment **S**earch **T**ool) service discovery scenario. Our initial experimental results are very encouraging—demonstrating up to 73% success rates of service discovery and showing how the incorporation of service clues into the search process may improve service matching throughput. These results suggest an opportunity for efficient service discovery in the face of the large and growing number of Web services. The DynaBot prototype has been successfully deployed by Lawrence Livermore National Lab for use in aiding bioinformatic service discovery and integration, and its further development and testing is continuing as part of the LDRD project; the goal of this project is to provide scientists with access to hundreds of data sources through a single, intuitive interface, thereby simplifying their interaction with data and enabling them to answer more complex questions than currently possible.[2]

The remainder of the article is organized as follows. We first describe our service class model and the construction of service class descriptions. We then outline the architectural design of DynaBot with a focus on system-level design and development issues, including our Deep Web probing methodology and the SCD-based service matching algorithms. We then present our initial experimental results that demonstrate the effectiveness and scalability of DynaBot for discovering domain specific Deep Web sources and services. We then conclude the article with a summary of related work and a discussion of open research issues.

THE SERVICE CLASS MODEL

Research on DynaBot for automatically discovering and classifying Web services is motivated by the need to fill the gap between

QoS Evaluation of End-to-End Services in Virtualized Computing Environments: A Stochastic Model Approach

Guofeng Yan, Yuxing Peng, Shuhong Chen and Pengfei You (2015). *International Journal of Web Services Research (pp. 27-44).*
www.igi-global.com/article/qos-evaluation-of-end-to-end-services-in-virtualized-computing-environments/125457?camid=4v1a