Toward Information Retrieval Web Services for Digital Libraries

Yueyu Fu, Javed Mostafa, and Weimao Ke
Laboratory of Applied Informatics Research
Indiana University, Bloomington, IN, 47405-3907
(812)856-4182, 01
{yufu, jm, wke}@indiana.edu

Information retrieval (IR) functions serve a critical role in many information systems. There are many techniques and operations have been developed in IR that do not require radical changes or re-implementation. The implemented IR algorithms can be distributed or their functions made available through the framework of web services. A key idea behind web services is that frequently used functions can be implemented once and offered to other application or software environments through programmatic interfaces. Web services in the IR domain have not been widely tested. Library of User-Oriented Concepts for Access Services (LUCAS) offers IR functions such as term extraction, term clustering, and document classification as web services. It has successfully served other information systems such as the Biological Knowledge through Ontologies (BioKnOT) and the Extensible Networked Association-based Bioinformatics Learning Environment (ENABLE) Knowledge Base. In this paper, we demonstrate the utility of LUCAS to the ENALBE system.

Introduction

Library of User-Oriented Concepts for Access Services (LUCAS) offers IR functions such as term extraction, term clustering, and document classification as web services. It has successfully served other information systems such as the Extensible Networked Association-based Bioinformatics Learning Environment (ENABLE) Knowledge Base. ENABLE project is funded by NSF to extend digital library technologies in the emerging domain of bioinformatics and develop novel interfaces to support association-based learning of bioinformatics resources. The ENABLE Knowledge Base provides functions for bioinformatics resource collection and management. We will present the design of the LUCAS system, implemented using the web service framework, and how the LUCAS system can be used to develop effective and efficient complex information systems. LUCAS interacts with the ENABLE Knowledge Base by means of web services, providing an automated indexing service. Sections that follow present further details.

LUCAS Architecture

LUCAS mainly consists of three components: the web service server, which accepts and processes web service calls; the client, which passes the user selected parameters to the server and retrieves the results based on SOAP protocol, and the IR function modules. The basic operation supported by LUCAS is term extraction, term clustering, and document classification.

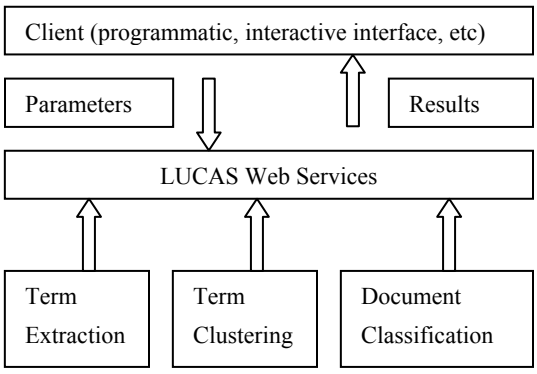

Figure 1: LUCAS architecture

Web Services

Three web service operations were developed in LUCAS. First, the term extraction operation retrieves "important" terms from a given web document. All the terms in the web document are ranked based on their TF*IDF weights [3]. Users can specify the URL of a web document and the number of terms requested. A list of top ranked terms is returned from this service. Second, the term clustering operation extracts top weighted terms from a document collection, computes term-term association, and cluster terms with centroids. A list of clusters is returned from this service. Third, the document classification operation classifies a web document into one of the pre-defined categories. Users can specify the URL of the web document and the domain of the document. A classification label for this document is returned from this service. Further information on LUCAS and applications of some of its services can be found at: http://tara.slis.indiana.edu:8080/lucas2/lucas2.html.

Application

LUCAS web service has successfully served other information systems such as ENABLE Knowledge Base. ENABLE project is funded by NSF to extend digital library technologies in the emerging domain of bioinformatics and develop novel interfaces to support association-based learning of bioinformatics resources. The ENABLE Knowledge Base provides functions for bioinformatics resource collection and management. LUCAS interacts with the ENABLE Knowledge Base by means of web services, providing an automated indexing service.

ENABLE Knowledge Base

ENABLE Knowledge Base consists of two major components: bioinformatics education resource collection and bioinformatics education resource management. The resource collection component automatically collects and refines information of on-line bioinformatics education resources. Based on the collected education resources, the resource management component provides learning services to users through a query interface based on some visualization techniques [2].

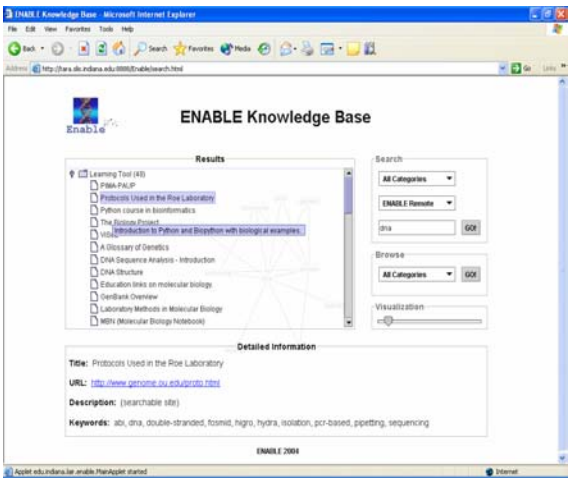

Figure 2: ENABLE user interface

ENABLE-LUCAS Interaction

In ENABLE Knowledge Base, the online bioinformatics education resources are collected by crawling web sites. Information thus collected is indexed for retrieval purposes. To generate the index words, the resource collection component sends each of the URLs of the HTML documents to LUCAS using SOAP. LUCAS term extraction service parses the HTML document, extracts keywords for each document, and sends the keywords back to the ENABLE Knowledge Base as a set of index words (see Figure 3). A demo (see Figure 4) shows how index words are extracted from a web page.

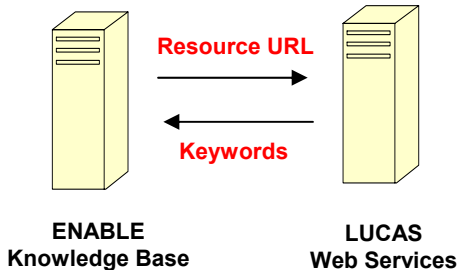

Figure 3: ENABLE – LUCAS interaction

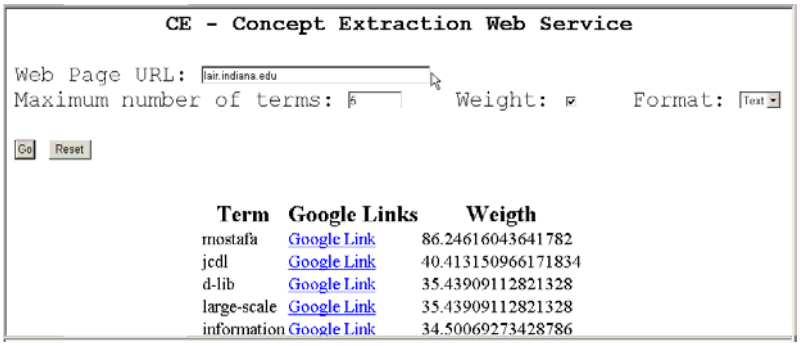

Figure 4: Index words extraction for a web page

Conclusion

Our system demonstrated that mature IR algorithms can be successfully turned into web services and can be accessed in a variety of ways. This flexibility enables easier integration of IR systems and/or algorithms without duplicating efforts. Future development of LUCAS will include development of advanced web services for more sophisticated IR functionality.

ACKNOWLEDGMENTS

This work was partially supported through a grant from the National Science Foundation Award#:0333623.

REFERENCES

Chen, H. (1999). Semantic Research for Digital Libraries. D-Lib Magazine, 5(10).

Liu, Y. and Mostafa, J. (2005). Fast clustering algorithm for scatter-gather browsing. Submitted to SIGIR'05.

Salton, G. (1983). Introduction to modern information retrieval. McGraw-Hill, New York.

Tilley, S., Gerdes, J., Hamilton, T., Huang, S., Müller, H., and Wong, K. (2002). Adoption Challenges in Migrating to Web Services. Proceedings of the Fourth International Workshop on Web Site Evolution.

Truner, M., Budgen, D., and Brereton, P. (2003). Turning Software into a Service. IEEE Computer, 36(10).

Web Term Document Frequency and Rank. Retrieved January 20, 2005 from http://elib.cs.berkeley.edu/docfreq/.