Integrating Data Mining and Data Management Technologies for Scholarly Inquiry: Project Summary

Paul Watry
Univ. of Liverpool
pwatry@liverpool.ac.uk

Ray R. Larson
Univ. of California, Berkeley
ray@sherlock.sims.berkeley.edu

Richard Marciano
University of North Carolina, Chapel Hill
richard_marciano@unc.edu

Abstract

The international "Digging Into Data Initiative" is exploring the uses and approaches for large-scale data analysis and processing for a variety of purposes primarily in the Humanities and Social Sciences. This paper discusses the "Integrating Data Mining and Data Management Technologies for Scholarly Inquiry" project being conducted at UC Berkeley, University of North Carolina, and the University of Liverpool. Our "big data" consists of the entire texts collection of the Internet Archive (approximately 3.6 million volumes) and the entire JSTOR database. We are performing surface-level natural language processing on this data to identify noun phrases and further refinements to identify personal, corporate, and geographic names. We are then using resources including library and archival authority records to identify variants and merge names. The goal is an integrated index of persons, places, and organizations and where they have been referenced in our collections.

Overview of goals and objectives

As noted in the abstract we had the twofold goal of developing a data analysis and processing system for large text collections, using a variety of prototype software, and then using the developed system to extract and index information about persons, organizations and places from very large text collections.

Our approach consists of developing a high level component framework extensible and flexible enough to accommodate radically different architectural models; one which supports large-scale preservation environments characteristic of content management systems; and support for semantic retrieval and natural language processes involving large-scale distributed datasets, served in a highly parallel environment [1,2]. Rather than prototyping a single solution, we have followed a modular approach with a variety of text and data mining, ontological, document rendering, and text retrieval tools that can be used in various combinations according to need. This framework is designed to address the twin aspects of access and preservation in new and sustainable ways.

The primary thrust of our development process has been to take a number of existing data-grid, digital library, and persistent archive management systems, develop them as an integrative framework for grid-based digital repositories, and apply this framework in ways which will support content management systems and workflow environments.

In support of this, we have integrated multiple information management technologies that have been developed across the data grid, digital library, and persistent archive communities, and apply these in an environment that will support highly scaled digital repositories.

Challenges and lessons learned

The primary challenge has been issues of development and component integration in our framework. Some components are being actively developed elsewhere, and changes in those added to the delays in preparing our system for operation. Since this system integration work was on the critical path for loading and analysis of the data, we were not able to start loading and indexing data until much later than expected.

In addition, while we had thought that there would be sufficient storage for the entire collections originally proposed, it turned out that there was, in fact, nowhere near the required 13-44 Terabytes estimated (under various assumptions). Given this scale of data and the estimated time (33-107 days of downloading) to make copies at our server cluster using the mechanism offered by the Internet Archive. Thus, we have not been able to obtain the planned local copy, but instead have been using the capabilities of our framework to incrementally download, parse and extract proper noun phrases from the Internet Archive sources, while retaining pointers to the original source for each unique name (noun phrase).

While this approach does accomplish the original goals, it takes much longer to do so. Thus, the primary lesson learned is to ensure that the required resources are in place and ready to go on projects with short time allotments and very limited budgets.

Indicators of success

Complete success would have been an online index to persons, organizations and places mentioned in any of the books in the Internet Archive, and in JSTOR. We may still achieve complete success, but not within the time limits of the current DID program.

Measuring impact

Since we have not been able to field a search tool for persons, places and organization in the Internet Archive and JSTOR, it is very difficult to estimate any impact. Certainly there has been a significant impact of the development of the framework used in the project

Knowledge dissemination mechanism and tools

We have presented information about the project, its goals, progress and technology at conferences, including the Joint ACM/IEEE Conference on Digital Libraries, and the Foundations of Information Science and Technology meeting in Shanghai. Specifically:

Ray R. Larson, Paul Watry. Richard Marciano – “Integrating Data Mining and Data Management Technologies for Scholarly Inquiry”, Presented at the Digging into Data Panel at JCDL, June 12, 2012.

Ray R. Larson, Paul Watry. Richard Marciano – "Digging Into Data: Data Mining for Information Access" , (Invited paper) Presented at the Frontiers of Information Science and Technology (FIST) Conference, Shanghai, China, December 12, 2012

John Harrison, Jerome Fuselier – "Python Support for iRODS (PyRODS + EmbedPython) & Content Search in iRODS (Cheshire3) Presented at the iRODS User Meeting – February 28 - March 1, 2013 RZG at IPP, Boltzmannstr. 2, 85748 Garching, Germany

Importance of working with libraries, archives and data repositories

We have had encouragement and help from colleagues at the Internet Archive in particular in obtaining the data that our project required. But we also faced some challenges in trying to avoid serious impacts on their infrastructure when attempting to obtain the data we wanted to use.

Capacity building and training (students and highly qualified personnel)

The students working on the project at Berkeley and the programmers working at UNC and Liverpool have had valuable training and skill enhancements working with a variety of new technologies.

References

[1] R. R. Larson and R. Sanderson, "Grid-Based Digital Libraries: Cheshire3 and Distributed Retrieval", *JCDL* 05, (June 7-11 2005.).

[2] T. Phelps, "Multivalent Documents: Anytime, Anywhere, Any Type, Every Way User-Improvable Digital Documents and Systems", Ph.D. Dissertation: University of California, Berkeley (1998).