Cheshire3: Retrieving from Tera-Scale Grid-Based Digital Libraries

Ray R. Larson
School of Information
University of California, Berkeley
Berkeley, California, USA, 94720-4600
ray@sims.berkeley.edu

Robert Sanderson
Department of Computer Science
University of Liverpool
Liverpool, L69 3BX, U.K.
azaroth@liverpool.ac.uk

Categories and Subject Descriptors:
H.3.3 [Information Systems]: Information Search and Retrieval—retrieval models, search process; H.3.7 [Information Systems]: Digital Libraries—systems issues

General Terms: Algorithms, Performance, Design

Keywords: Grid-Based Digital Libraries

1. EXTENDED ABSTRACT

Recent research in designing and developing digital library services has been focused on approaches to indexing and searching in a steadily increasing range of genres and materials. An important aspect of this research is concerned with providing effective and scalable IR services for digital libraries as these diverse collections grow to sizes measured in terabytes and petabytes. The Cheshire project has had a central research focus on large-scale digital library collections for more than a decade, with a current focus on supporting distributed digital libraries in a Grid environment. At the same time we have been prototyping systems for very long term digital preservation, and examining how grid-scale information retrieval systems can interoperate with petabytes of diverse data stored over many years.

In order for Information Retrieval (IR) in the evolving “Grid” parallel distributed computing environment[1] to work effectively, there must be a single flexible and extensible series of “Grid Services” with identifiable objects and a known API to handle the IR functions needed for Digital Libraries or other retrieval tasks. The Cheshire3 system builds on the work of the Cheshire project[4] over the past decade to define and implement an easy to use set of IR objects with precisely defined roles that can effectively provide a Grid Service for IR.

This demonstration will show how the Cheshire3 system is being applied in distributed information retrieval tasks in large-scale grid-based digital libraries. We will show how the system has been integrated with “Datagrid” services provided by distributed storage systems like the Storage Resource Broker (SRB)[3], and how this enables very large scale storage and retrieval systems with support for data preservation services using the “Multivalent Document” framework[2].

In this demo we will present the results of testing Grid-based parallel approaches in indexing and retrieval for a variety of information resources, ranging from small test collections like the TREC and INEX collections, to medium-scale metadata collections like Medline and a test version of University of California Online Union Catalog, MELVYL (with 15 million and 9.5 million records respectively) ranging up to large-scale collections like the US National Records and Archives Administration (NARA) Preservation Prototype.

We will describe our approaches to indexing and retrieving from these collections and the architecture of the system that supports them, as well as providing a live demonstrations of Grid-based retrieval over these collections.

We will also show how the system is managed, in both standalone and distributed processing, by a set of user defined workflow descriptions (which are specified simply through XML configuration files). We will demonstrate how Cheshire3 databases are created and how each database is actually a logical collection of records, that can be easily split across many nodes in a Grid environment or combined at a single location. We will show the performance statistics for indexing and retrieval using standard test collections, and examine the performance issues in a distributed environment.

2. ACKNOWLEDGMENTS

Development of the Cheshire3 system was supported in part by the Joint Information Systems Committee(U.K.)

3. REFERENCES