

NSF/JISC REPOSITORIES WORKSHOP

Repositories for Large-scale Digital Libraries

William Y. Arms
Cornell University
March 27, 2007

Large-scale digital libraries

Emerging digital libraries are large, even by the standards of high-performance computing. Here are some examples: the Internet Archive's historical collection of the web, the Library of Congress's National Digital Information Infrastructure and Preservation Program, the USC Shoah Foundation's video collection of interviews with survivors of genocide, and the web search services such as Google and Yahoo. The mass digitization of books might become the most important of all. The size of these collections is measured in hundreds of terabytes or petabytes; several of them have billions of items.

With disappointingly few exceptions, the digital libraries community has paid little attention to the questions of how to manage these very large collections and how to support researchers in using them. Academic libraries and digital library researchers have largely abandoned large-scale digital libraries to commercial companies and the not-for-profit Internet Archive. This is doubly unfortunately: it isolates universities from a vibrant area of research and innovation, and it forces libraries into alliances with commercial companies that they may regret in the long term.

In fall 2006, we received an NSF Small Grant for Experimental Research to study one aspect of large-scale digital libraries: the use of the NSF Cyberinfrastructure to index very large volumes of text. Specifically, we are exploring the challenge of building a full text index to all the textual web pages in the historical collections of the Internet Archive. The total size of this corpus after uncompressing is several petabytes. Our work builds on an existing project, the Cornell Web Lab, which is organizing large portions of the collections for research by social scientists and computer scientists.

Research in organizing large-scale digital libraries

Once digital libraries grow beyond about a hundred terabytes or have billions of records, the most flexible and cost-effective way to manage them is with large clusters of small computers. Because organizations such as Google, Yahoo, and the Internet Archive run large-scale production services, they have been forced to develop expertise in programming and operating such clusters. This expertise enables them to carry out research and introduce new services at low marginal cost. Unfortunately, the academic community has been left behind in the development of cluster computing for digital libraries and is in danger of being left further behind.

However, there is hope. Google has published some valuable papers about the principles behind its distributed file system and its use of the map/reduce programming paradigm. The Internet Archive has been energetic in advocating the development of open source software. Several commercial companies, notably Yahoo, have contributed greatly to open source software development.

Three open source software packages are particularly valuable: the Heritrix web crawler, the Nutch family of indexing software, and Hadoop, which implements the map/reduce paradigm. At Cornell, we are running all three packages with considerable success. In particular, we have found that the map/reduce model provides an easy way to build and organize very large collections, while hiding most of the complexity of parallel programming from the developers.

Many of the tasks of organizing very large collections can be carried out on equipment that is within the budgets of universities. For the Web Lab, we have a 16-processor SMP machine, which is used as a relational data base server, and we share a 128-node Linux cluster. Some tasks, such as building a full text index to a complete web crawl, require computational power far beyond the resources of a single university, but only for short periods of time. This is a natural use of national supercomputers and the NSF's Cyberinfrastructure.

Research that uses large-scale digital libraries

In a recent seminar, Greg Crane of Tufts University made the simple but profound statement, "When collections get large, only the computer reads every word." A scholar can read only one document at a time, but a supercomputer can read millions. In a traditional library the scholar personally browses the collection, searches through the catalog, and takes books off the shelves. With very large digital collections, the equivalent functions are performed by computer programs, acting as agents for people. Researchers do not interact with the collections directly. They use computer programs to extract small parts of the collection and rarely view individual items except after preliminary screening by programs.

When we began to work with data from the Internet Archive's web collection, one of our goals was to enable research by people, such as social scientists, who are not professional programmers. As this work has progressed, the critical resource in carrying out research has been skilled programmers. Many social scientists can write simple programs in Java or scripting languages. Few have the expertise to write high performance parallel code.

We must avoid a situation where every use of the collections requires complex programming. This is a fascinating research area. This semester, several masters students have run map/reduce tasks on the Linux cluster. A typical task is to build the adjacency matrix of a web graph starting with about a billion raw links. Once the map/reduce concept is understood their progress has been pleasingly rapid. Other colleagues are exploring different approaches to this same problem.

How can the NSF/JISC help?

Managing very-large digital libraries and carrying out research on the collections are substantial challenges, but they are within the reach of academic libraries and digital library research groups if certain criteria can be met.

- 1) Multi-year funding must be available to support the development of and research on large-scale collections. It takes three to five years to build a cadre of expertise. Running long-term projects on a series of short-term grants is highly inefficient.
- 2) In the US, the Cyberinfrastructure needs to support this research explicitly. Current supercomputers are configured for computations, or for the management of scientific data, not for semi-structured information such as text.
- 3) The development of open source software and its support on the Cyberinfrastructure is essential. It is possible that a software regime based on map/reduce might become as important for digital libraries as the Grid stack is for large-scale computation.
- 4) Research is needed into the tools that allow non-computing specialists to analyze very large collections, with minimal assistance from professional programmers.

Acknowledgements

The work described above would not be possible without the assistance of the Internet Archive. This work is funded in part by National Science Foundation grants CNS-0403340, SES-0537606, and IIS 0634677.