Growth and Server Availability of the NCSTRL Digital Library

Allison L. Powell James C. French*
Department of Computer Science, University of Virginia
Charlottesville, VA
{alp4g|french}@cs.virginia.edu

ABSTRACT

This paper reports on measurements of the NCSTRL digital library taken over a two-year period. We report the growth of the system along two dimensions: number of participating institutions and number of documents indexed by the system. We also report an aspect of reliability for this distributed digital library system.

INTRODUCTION

The Networked Computer Science Technical Reference Library (NCSTRL) has been in existence for about five years. In this paper we explore some aspects of NCSTRL observed by regular polling of the system over a two-year period from 1997-1999.

Briefly, NCSTRL is a collection of Computer Science technical reports[1] organized as a loose federation of cooperating servers. Although the document repositories in NCSTRL are located at geographically distributed sites, there are two strategies in use for maintaining the metadata and providing indexing services: (1) a geographically distributed set of index servers; and (2) a centralized index server. The former configuration is maintained as a research vehicle and is the configuration that we examine in this paper. The latter is the production system designed to provide stability to end users.

Participating institutions (known as "publishing authorities") can be involved as "Standard" or "Lite" sites. A Standard site runs three services: user interface (UI), indexer, and repository. A Lite site maintains its technical reports at the home institution, but has its metadata held at a special site (the Central Server) that provides indexing services. The Central Server looks like a Standard site to the rest of the system. There is also a Backup Server to increase reliability. It is highly available but possibly a little out of date.

Within NCSTRL query processing proceeds as follows. First a user poses a query to a UI server. The query is broadcast to all the Standard sites including the Central server. Each local indexer processes the query and sends the results back to the issuing UI. If all sites do not respond, the Backup Server is contacted to supply results for the nonresponding sites.

In the sections that follow, we discuss the growth of NCSTRL over the two-year observation period and give some insight into its reliability. In this paper, we will judge reliability by estimating how often the Backup Server must be invoked. The reader is referred to Davis and Lagoze [1] for a more detailed description of NCSTRL and to Dushay *et* al. [2, 3] for more comprehensive performance measurements of the system.

GROWTH OF NCSTRL

We polled each NCSTRL site for a count of documents from 9-Aug-97 through 25-Sep-99. In response to the poll, the servers responded with either a count of documents or an error message. The Saturday polls from the two-year period of 6-Sep-97 through 28-Aug-99 are reported here. Over the course of the two-year period, there were instances of incomplete polls, e.g., a timed-out connection at some server might cause the poll to terminate prematurely. Weeks with incomplete polls were elided, leaving polls from 78 weeks. We have at least one complete poll for each month in the two-year period.

For purposes of counting the number of documents in NCSTRL, if a server replied with an error message, the most-recently reported document count for that server was used. If no document count had previously been reported, that server was assigned a count of zero documents until a document value was reported. As a result, the values reported here represent a lower bound on the number of documents in NCSTRL at a given time. There were also four sites that always replied with server errors and never reported a document count. These sites are counted in the total number of sites, but do not contribute to the total number of documents.

In Figure 1, we show the total number of documents in NC-

*This work supported in part by DARPA contract N66001-97-C-8542 and NASA GSRP NGT5-50062.

[1]Recently a wider variety of material has been accepted into the system.

Digital Libraries, San Antonio, TX.

STRL for the last poll of the month, for each month in the two-year period. We began with 17,406 documents (6,174 of these found at Lite sites) and finished with 29,367 (9,750 Lite), a total increase in holdings of 69%. This is an increase of 75% at Standard sites and 58% at Lite sites. The sharp increase in October, 1998 is due to the addition of a NCSTRL Standard site for INRIA (Institut National de Recherche en Informatique et en Automatique) which added over 3,500 documents.

We also tracked the growth of the number of Lite and Standard NCSTRL sites. We began the period with 98 sites (56 of these Lite) and finished with 123 sites (63 Lite). So the total number of participating sites increased by 26%; Standard sites grew by 43% while Lite sites grew only 13%. At the end of the observation period there were approximately the same number of Lite and Standard sites. Davis and Lagoze[1] also report the growth of NCSTRL sites, but for a different time period.

As a gross characterization of NCSTRL, we can say that two thirds of the documents are stored in about one half the sites, i.e., the Standard sites.

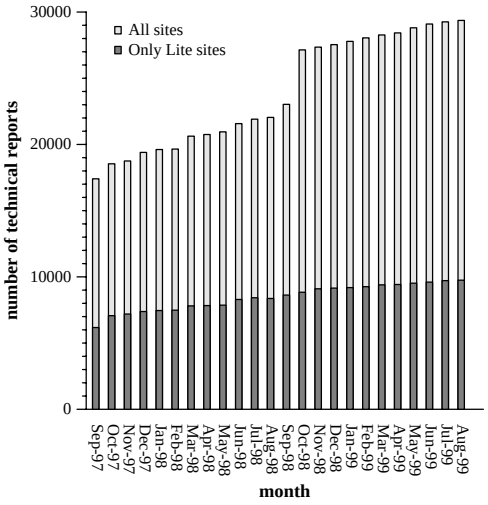

Figure 1: Number of technical reports.

RELIABILITY

We eliminated a number of sites for the purposes of examining server reliability. We eliminated sites that always responded with an error (in case the problem was due to our polling approach at that site) and the sites that added no new reports during the two-year period (assuming that they were inactive and not monitoring the server). It is also the case that a single server might serve up documents for multiple publishing authorities, e.g., the Lite server handles all 63 Lite sites. We report on the availability of 38 servers. Reliability percentages are calculated over the 78 weekly polls. A variety of errors were encountered, in some cases the cause was unclear. We only concerned ourselves with three of the most commonly occurring ones, "could not connect", "connection timed out" and "dienst server unavailable".

Table 1 shows the availability of servers on a server-by-server basis. For example, if a server responded to 77 out of 78 polls, it would have an uptime of 98.7%. Two sites had an uptime of 100%. The average observed uptime was 86.7%. Dushay, *et al.* [3] observed average uptimes of 87% and 89% when studying a smaller subset of servers. The number of sites that failed on a given percentage of polls (on a poll-by-poll basis) is shown in Table 2. In 100% of the polling attempts, we experience at least one server failure. This implies that during query processing we would have contacted the backup server for every query.

Uptime (%)	Num. Servers
95-100	14
90-94.9	9
85-89.9	5
80-84.9	2
75-79.9	3
70-74.9	1
less than 75	4

Table 1: The number of servers that responded to k percent of the polling attempts.

# Servers Down	6	5	4	3	2	1	0
% of Polls	3.8	7.7	28.2	34.6	23.1	2.6	0

Table 2: The percentage of polling attempts in which exactly j servers were down.

DISCUSSION

NCSTRL is a growing digital library. In the two-year monitoring period the number of participating sites increased by 26% and the holdings grew by 69%. Reliability of the distributed system is low. Table 1 shows that many servers are highly available, specifically 23 of 38 (61%) are up 90% of the time. However, from Table 2 we see that the system had at least one server failure 100% of the time. This implies that during query processing, every query will have to be routed to the Backup Server, increasing the overall system response time.

Our measurements indicate that engineering reliable, distributed digital libraries will be a challenge. A federated system is vulnerable to its weakest component. Strong institutional commitment will be necessary for success.

REFERENCES

1. Davis, J. R. and C. Lagoze, "NCSTRL: Design and Deployment of a Globally Distributed Digital Library," *JASIS*, 51(3):273-280, 2000.

2. Dushay, N., J. C. French and C. Lagoze, "Using Query Mediators for Distributed Searching in Federated Digital Libraries," *Proc. Fourth ACM Conf. on Digital Libraries*, Aug. 1999, pp. 171-178.

3. Dushay, N., J. C. French and C. Lagoze, "Predicting Indexer Performance in a Distributed Digital Library," *European Conf. on Digital Libraries*, Sept. 1999, pp. 142-166.