

Efficient Searching in Distributed Digital Libraries

James C. French Allison L. Powell Walter R. Creighton, III *
 Department of Computer Science
 University of Virginia
 Charlottesville, VA 22903-2442
 Tel: 1-804-982-2213
 E-mail: {french|alp4g|wrc7m}@cs.virginia.edu

ABSTRACT

When a digital library is decomposed into many geographically distributed repositories, search efficiency becomes an issue. Increasing network congestion makes this a compelling issue. We discuss an effective method for reducing the number of servers needed to respond to a query and give examples of search space reduction in the NCSTRL distributed digital library.

KEYWORDS: collection selection, database selection, text resource discovery, distributed searching

INTRODUCTION

There are three fundamental activities associated with distributed searching in digital libraries: (1) choosing the specific database collections to search; (2) searching the chosen databases; and (3) merging the results into a cohesive response. Each of these is an important area of research, but here we focus specifically on the first activity. This has been called the *collection selection* problem by Callan *et al.*[1], while Gravano *et al.*[3] refer to it as the *text database resource discovery*. It is also referred to as *database selection*[2].

The database selection problem is generally formulated as: given a query q , rank the available databases in order of presumed utility to q . The simple techniques discussed here can be used to *exclude* databases from consideration when it can be determined from the query that they could not possibly yield a response. A simple example will illustrate the difference. Assume that we have a query that includes a term specifying an author, among potentially many other search predicates. We can then rule out all database sites which are known not to have any documents written by the requested author. This will not affect the effectiveness of the search; we will achieve exactly the same result by processing the query at fewer sites. This will, however, increase the efficiency of the search by obviating the need to process the query at

superfluous sites.

It is clear that the most efficient way to search a site is never to process a query there if it can be avoided. Here we discuss this principle as applied to author and title searching in the NCSTRL collection.

NCSTRL

NCSTRL is the Networked Computer Science Technical Reports Library, a collection of the technical reports of over 100 computer science departments and research laboratories worldwide. At the present time NCSTRL includes about 22,000 technical reports. It is a distributed architecture based on an open protocol, DIENST[4]. NCSTRL participants elect to participate as either "standard" or "lite" sites. Standard sites support user interface, index and repository services; lite sites provide a repository only and are indexed at the NCSTRL Central Server which is itself a modified standard site. We used 39 standard sites and 43 lite sites for the results reported here.

Example

Table 1 demonstrates the potential of site selection in the NCSTRL collection when relatively uncommon and relatively common names are used as part of a search. The table shows the number of sites containing at least one instance of each named author. The first column of the table considers only the 39 standard sites; the second column considers all 82 sites. The search space reduction is from 95% – 100% for the relatively uncommon names and from 62% – 82% for the common names. The search space reduction is clearly significant for these examples. Below we extend this analysis to include all the authors in the NCSTRL collection.

| Author Last Name | Standard only | Standard plus Lite |
|------------------|---------------|--------------------|
| Creighton | 0 | 0 |
| French | 1 | 3 |
| Powell | 2 | 3 |
| Jones | 5 | 12 |
| Lee | 15 | 25 |
| Smith | 15 | 24 |

Table 1: The number of NCSTRL sites containing at least one instance of each author's last name.

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

METHODOLOGY AND ANALYSIS

We examined the efficacy of a query routing mechanism based on author and title words. To do that we analyzed the NCSTRL data for authors, title words, and stemmed title words. For each unique author or word we determined the number of sites containing that author or having that word in a title. Finally, we histogrammed the data and created cumulative distributions.

Authors

Authors were extracted from the bibliographic records of each site and reduced to last name only. There was some error in this process because some of the bibliographic records strayed from the RFC 1807 standard and could not be easily disambiguated by automated means. There was also no standard encoding of diacritical marks. For example, we found at least five different ways that umlauts were encoded. Nevertheless, the error due to these problems is very small.

Title Words

Each title was extracted from the bibliographic records at each site. All nonalphabetic characters were deleted. The practical effect of this in most cases was to convert constructions such as $O(\log n)$ to the nonsense string *Ologn*. More sophisticated parsing could be used to preserve these constructions, but the effect on the work reported here is minimal.

Stemmed Title Words

Titles words were extracted as above and then conflated using a stemming algorithm due to Paice[5]. We were interested to see if stemmed title words would retain the selectivity to be of use in search space reduction.

Cases to Consider

We used 39 NCSTRL standard sites and 43 NCSTRL lite sites from the Central Server as the basis for this analysis. Our selection of sites was based on the availability of sufficient bibliographic data at the sites. We considered the sites in three configurations for the purpose of the analysis.

Case 1. This configuration consisted of 40 standard sites, the 39 standard sites from above plus all 43 lites sites considered as a single standard site. This case reflects the distributed searching scenario in NCSTRL where a query is broadcast to all standard sites and the Central Server.

Case 2. This configuration consisted of 82 standard sites, the 39 standard sites from above plus the 43 lite sites, each considered as an individual standard site. This case represents a distributed system about twice as large as Case 1.

Case 3. This configuration consisted only of the 39 standard sites from above. Thus, Case 3 differs from Case 1 by the removal of the composite standard site composed of the lite sites. Because the composite site is made up of 42 sites, we wanted to make sure that it did not have unusual properties that unduly dominated Case 1.

RESULTS

We extracted the authors and title words for each of the cases described above. In each case we computed the expected number of sites to search given a single author or title word. Table 2 shows the results of these calculations.

| | Case | | |
|---------------------|------|------|------|
| | 1 | 2 | 3 |
| Authors' names | 1.33 | 1.42 | 1.31 |
| Title words | 2.85 | 3.81 | 2.97 |
| Stemmed title words | 3.27 | 4.61 | 3.55 |

Table 2: For each case we show the expected number of sites to be searched given an arbitrary author or title word from the collection. We do not consider authors or words not occurring in the collection.

It is clear from the data in the table that substantial search space reduction is possible. On average we can expect to enjoy a 97% reduction of the search space based on a single author in a query while title words can be used to reduce the search space 92–95%. This latter fact is most surprising. We had not anticipated such a widely skewed distribution of title words, but our data showed that individual title words occur at only one site in approximately 57% of the cases. The stemmed title words have good selectivity (91–94%) as well.

Note also that Case 1 and Case 3 do not differ appreciably. This implies that the skewed distribution is not due to some artifact of the composition of the 43 lite sites into a single standard site.

CONCLUSIONS

In the NCSTRL collection the distribution of author and title word data is sufficiently skewed so that it can be exploited favorably to radically reduce the number of sites that have to be searched. Moreover, we can reasonably expect to retain author selectivity as the collection grows. It is not clear how well the title word selectivity will hold up as the system grows, although the selectivity of the stemmed title words seems to suggest that title word selectivity might scale reasonably well.

A back-of-the-envelope calculation shows that at the present time a data structure for author selection would require about 50K to 60K bytes, sufficiently small to be replicated in each standard site's user interface to provide query routing services. The title data can be widely replicated as well. This will allow queries to be multicast to many fewer sites without degrading query effectiveness.

REFERENCES

1. J. P. Callan, Z. Lu, and W. B. Croft. Searching Distributed Collections with Inference Networks. In *Proc. of SIGIR'95*, pages 21–29, Seattle, WA, 1995.
2. J. C. French, A. L. Powell, C. L. Viles, T. Emmitt, and K. J. Prey. Evaluating Database Selection Techniques: A Testbed and Experiment. In *Proc. of SIGIR'98*, Melbourne, Australia, August 1998. (To appear).
3. L. Gravano, H. Garcia-Molina, and A. Tomasic. The Effectiveness of GLOSS for the Text Database Discovery Problem. In *SIGMOD'94*, pages 126–137, Minneapolis, MN, May 1994.
4. C. Lagoze and J. R. Davis. Dienst: An Architecture for Distributed Document Libraries. *CACM*, 38(4):47, 1995.
5. C. D. Paice. Another Stemmer. *SIGIR Forum*, 24(3):56–61, 1990.