Distributed Resource Discovery: Using Z39.50 to Build Cross-Domain Information Servers

Ray R. Larson
School of Information Management and Systems
University of California, Berkeley
(510)642-6046
ray@sherlock.berkeley.edu

ABSTRACT

This short paper describes the construction and application of Cross-Domain Information Servers using features of the standard Z39.50 information retrieval protocol[11]. We use the Z39.50 Explain Database to determine the databases and indexes of a given server, then use the SCAN facility to extract the contents of the indexes. This information is used to build "collection documents" that can be retrieved using probabilistic retrieval algorithms.

Keywords

Distributed Information Retrieval, Cross-Domain Resource Discovery, Distributed Search

1. INTRODUCTION

Information seekers must be able to identify information resources that are pertinent to their needs. Today they are also required to have the knowledge and skills to navigate those resources, once identified, and extract relevant information. The widespread distribution of recorded knowledge across the network landscape of the World Wide Web is only the beginning of the problem. The reality is that the repositories of recorded knowledge on the Web are only a small part of an environment with a bewildering variety of search engines, metadata, and protocols of very different kinds and of varying degrees of completeness and incompatibility. The challenge is to not only to decide how to mix, match, and combine one or more search engines with one or more knowledge repositories for any given inquiry, but also to have detailed understanding of the endless complexities of largely incompatible metadata, transfer protocols, and so on.

As Buckland and Plaunt[1] have pointed out, searching for recorded knowledge in a digital library environment involves three types of selection:

1. Selecting which library (repository) to look in;

2. Selecting which document(s) within a library to look at; and

3. Selecting fragments of data (text, numeric data, images) from within a document.

In the following discussion we focus on the first type of selection, that is, discovering which digital libraries are the best places for the user to begin a search. Our approach to the second selection problem has been discussed elsewhere[6,7].

Distributed information retrieval has been an area of active research interest for many years. Distributed IR presents three central research problems that echo the selection problems noted by Buckland and Plaunt. These are:

1. How to select appropriate databases or collections for search from a large number of distributed databases;

2. How to perform parallel or sequential distributed search over the selected databases, possibly using different query structures or search formulations, in a networked environment where not all resources are always available; and

3. How to merge results from the different search engines and collections, with differing record contents and structures (sometimes referred to as the collection fusion problem).

Each of these research problems presents a number of challenges that must be addressed to provide effective and efficient solutions to the overall problem of distributed information retrieval.

These problems been approached in a variety of ways by different researchers focusing on different aspects of retrieval effectiveness, and on construction of the index resources required to select and search distributed collections [2,3,4,5,8,9].

In this paper we present a method for building resource discovery indexes that is based on the Z39.50 protocol standard instead of requiring adoption of a new protocol. It does not require that remote servers perform any functions other than those already established in the Z39.50 standard.

2. Z39.50 FACILITIES

The Z39.50 Information retrieval protocol is made up of a number of "facilities", such as Initialization, Search, and Retrieval. Each facility is a logical group of services (or a single service) to perform various functions in the interaction between a client (origin) and a server (target). The facilities that we will be concerned with in this paper are the Explain Facility and the Browse Facility.

2.1 The Explain Facility

The Z39.50 Explain facility permits the client to obtain information about the server implementation, including information on databases supported, attribute sets used (an attribute set specifies the allowable search fields and semantics for a database), diagnostic or error information, record syntaxes and information on defined subsets of record elements that may be

JCDL '01, June 24-28, 2001, Roanoke, Virginia.
Copyright 2000 ACM 1-58113-000-0/00/0000…$5.00.

requested from the server (called elementsets). The server (optionally) maintains a database of Explain information about itself and may maintain Explain databases for other servers. The explain database appears to the client as any other database, and uses the Z39.50 Search and Retrieval facilities to query and retrieve information from it. There are specific attributes, search terms and record syntaxes defined in the standard for the Explain database to facilitate interoperability among different server implementations.

2.2 Z39.50 Browse Facility

As the name of this facility implies, it was originally intended to support browsing of the server contents, specifically the items extracted for indexing the databases. The single service in the Browse facility is the Scan service. It is used to scan an ordered list of terms (subject headings, titles, keyword, text terms, etc.) drawn from the database. Most implementations of the Scan service directly access the contents of the indexes on the server and return requested portions of those indexes as an ordered list of terms along with the document frequency for each term.

3. IMPLEMENTATION

Our implementation relies on these two Z39.50 Facilities to derive information from Z39.50 servers (including library catalogs, full-text search systems, and digital library systems) in order to build a GlOSS-like[5] index for distributed resources. The procedure followed is:

1. Search the Explain Database to derive the server information about each database maintained by the server and the attributes available for searching that server.
2. For each database, we determine whether Dublin Core attributes are available, and if not, we select from the available attributes those most commonly associated with Dublin Core information.
3. For each of the attributes discovered, we send a sequence of Scan requests to the server and collect the resulting lists of index terms. As the lists are collected they are verified for uniqueness (since a server may allow multiple search attributes to be processed by the same index) so that duplication is avoided.
4. For each database an XML *collection document* is constructed to act as a surrogate for the database using the information obtained from the server Explain database and the Scans of the various indexes.
5. A database of collection documents is created and indexed using all of the terms and frequency information derived above.

Probabilistic ranking methods[6] are used to retrieve and rank these collection documents for presentation to the user for selection, or for automatic distributed search of the most highly ranked databases using method similar to [2] and [9]. Although the Z39.50 protocol has been used previously for resource discovery databases[8], in that work random samples of the records in the collection were used to build the indexes. The method described here gives the ability to use the server's own processing of its records in extracting the terms to be matched in the resource discovery index.

4. CONCLUSION

This brief paper has presented a standards-based method for building a resource discovery database from distributed Z39.50 servers. We plan to combine this system with others that support SDLIP, and possible other protocols for further investigation of distributed search for Digital Libraries.

5. ACKNOWLEDGMENTS

This work was sponsored by NSF and JISC under the International Digital Libraries Program award #IIS-9975164.

6. REFERENCES

[1] Buckland, M. K. and Plaunt, C. Selecting Libraries, Selecting Documents, Selecting Data. In ISDL'97: (Tsukuba City, Japan, 1997). http://www.dl.ulis.ac.jp/ISDL97/proceedings/.

[2] Callan, J. P., Lu, Z. and Croft, W. B. Searching Distributed Collections with Inference Networks. In SIGIR '95: (Seattle, WA, 1995), ACM Press, 21-28.

[3] Danzig, P.B., Ahn, J., Noll, J. and Obraczka, K. Distibuted Indexing: A Scalable mechanism for Distributed Information Retrieval. In SIGIR '91 (Chicago, IL, 1991), ACM Press, 220-229.

[4] French, J. C., et al. Evaluating Database Selection Techniques: A Testbed and Experiment. In SIGIR '98 (Melbourne, Australia, 1998), ACM Press, 121-129

[5] Gravano, L. and Garcia-Molina, H. Generalizing GIOSS to Vector-Space Databases and Broker Hierarchies. In VLDB '95 (Zurich, Switzerland, Sept 1995), ACM Press, 78-89

[6] Larson, R. R. and Carson, C. Information Access for A Digital Library: Cheshire II and the Berkeley Environmental Digital Library. In Proceedings ASIS '99 (Washington, DC, 1999), Information Today, 515-535.

[7] Larson, R. R. TREC Interactive with Cheshire II. Information Processing and Management. (In Press, 2001).

[8] Lin, Y., et al. Zbroker: A Query Routing Broker for Z39.50 Databases. In: CIKM '99: (Kansas City, MO, 1999), ACM Press, 202-209.

[9] Xu, J. and Callan, J. (1998) Effective Retrieval with Distributed Collections. In SIGIR '98, (Melbourne, Australia, 1998), ACM Press, 112-120.

[10] Z39.50 Maintenance Agency. Information Retrieval (Z39.50): Application Service Definition and Protocol Specification (ANSI/NISO Z39.50-1995), Washington: Library of Congress, 1995