Experiences with Distributed Searching

Nigel Ward
Senior Researcher
CRC for Enterprise Distributed Systems*
nigel@dstc.edu.au

Abstract:
The Resource Discovery Unit at the DSTC investigates techniques for improving access to information on heterogeneous networks like the Internet. Part of our research lead to the development of HotOIL - a search tool that distributes queries and collects search results from networked databases. This paper describes distributed searching as a resource discovery technique, how HotOIL implements distributed searching, and our experiences in deploying HotOIL to meet the needs of various communities.

* The work reported in this paper has been funded in part by the Co-operative Research Centre for Enterprise Distributed Systems Technology (DSTC) through the Federal Government's CRC Programme (Department of Industry, Science & Resources).

EXPERIENCES WITH DISTRIBUTED SEARCHNG

Introduction

The Resource Discovery Project[1] within the DSTC aims to investigate and develop tools, technologies, and information management processes that allow organisations to locate, access, retrieve, and manage information on highly distributed and heterogeneous networks such as the Internet. The project has investigated a number of technologies to meet these goals. This paper discusses our experience with one such technique: distributed searching.

Firstly we describe distributed searching and how it can help improve resource discovery. The rest of the paper then describes HotOIL - a distributed searching tool developed by the DSTC. In particular, we describe HotOIL from the user perspective, examine the assumptions and abstractions necessary to implement a distributed search tool, and finally, describe our experiences with deploying HotOIL to solve some real world problems.

What is Distributed Searching?

Much information currently being made available on the Internet is provided by backend databases. Existing web search engines cannot index this information. The result is the so-called "hidden web". Recent estimates claim the size of this hidden web to be an order of magnitude larger than the visible web[2].

Distributed searching is a technique for providing access to information in the hidden web. In essence, it involves sending a user's information request to a number of databases, unifying the results, and displaying them to the user.

What is HotOIL?

HotOIL is a distributed, heterogeneous search engine.

- **Distributed** because it can search other search engines on behalf of the user.

- **Heterogeneous** because it can search databases using multiple search protocols. Most other distributed search engines, such as DogPile only search using one information protocol.

Given a user's query HotOIL implements distributed searching in a number of steps. Firstly, HotOIL asks the user to choose which databases to query. This is done by asking the user to choose clusters of databases to query (see Figure 1).

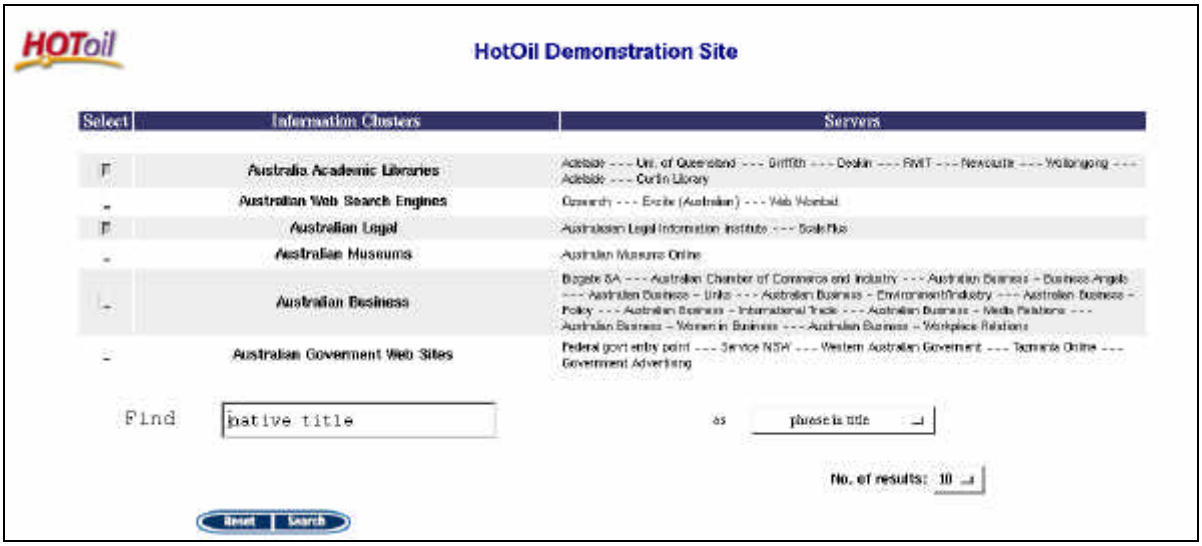

Figure 1: Search Interface

Once the databases have been selected, HotOIL translates the user query into queries for each database. These queries are then sent to database using a standard search protocol. HotOIL can currently interrogate databases supporting the HTTP and Z39.50 search protocols. Support for the LDAP and ODBC interface standards is being investigated.

HotOIL translates the results returned from each database into a common format - the Dublin Core Metadata Set[3]. The Dublin Core is emerging as the de facto standard for describing Internet resources.

Next, HotOIL merges the results from each database and attempts to remove duplicate results.

Finally, HotOIL summarises the results into a concise format using the Hyper-index Browser[4]. This summary provides an overview of the result set and allows the user to construct more precise queries by simply clicking on suggested queries (see Figure 2).

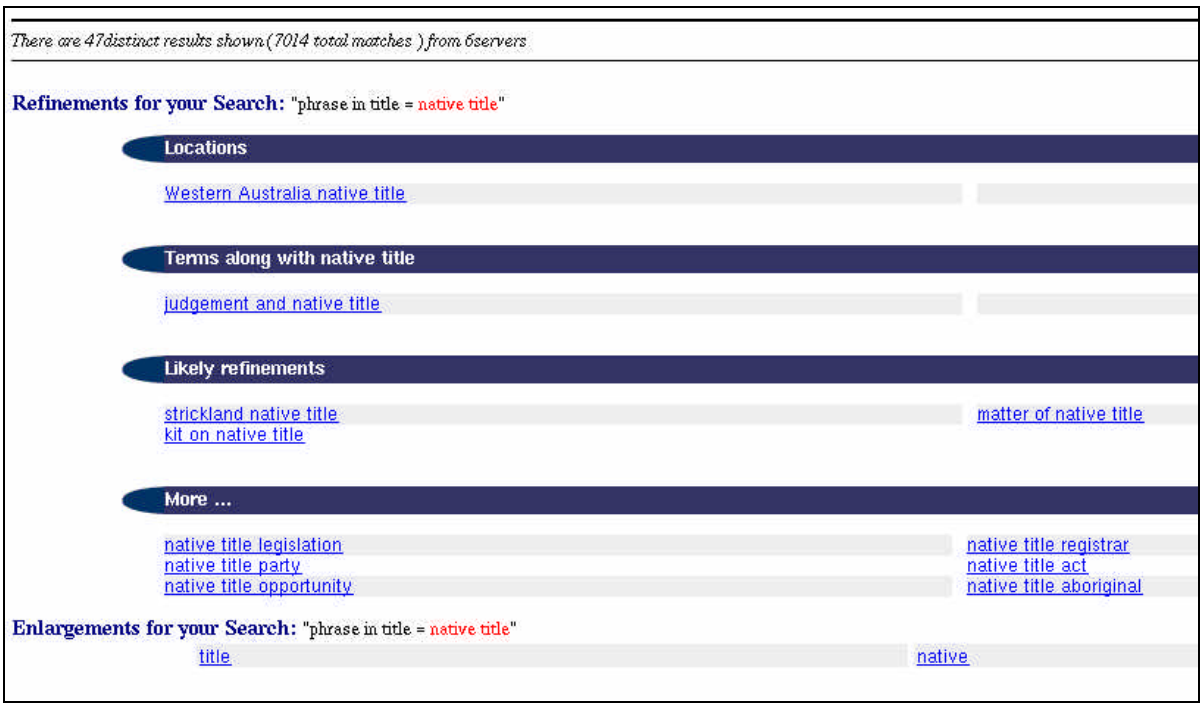

Figure 2: Result Summary

Following the result summary, the user is shown brief descriptions of each of the results, along with an indication of the databases that returned the results (see Figure 3).

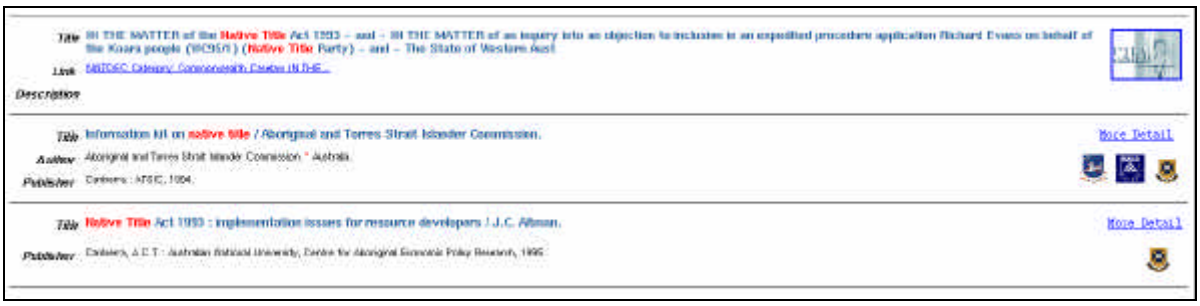

Figure 3: Result Screen

Some results contain more information than is shown on the result screen. The user can select to see full Dublin Core record of these results (see Figure 4).

Figure 4: Detailed Result

Implementation Abstractions

Because networked databases are built to meet a variety of information needs they have a variety of search interfaces. To avoid confusing the user with the idiosyncrasies of each networked database, HotOIL provides an *abstract view* of networked databases that gives the illusion that they all have a uniform interface. This section describes the abstractions used by HotOIL to create this illusion.

Information Protocol Abstraction

Networked databases use a variety of protocols to make their information available. Different protocols provide interaction models. For example, web search engines using HTTP have a simple interaction model: they allow a query to be submitted and return results immediately. Other protocols, such as Z39.50 provide more sophisticated interaction models: upon receipt of a query, Z39.50 returns an indication of the number of results matching that query. The user can then decide to submit a new query, or retrieve some or all of the results, in a variety of result formats.

HotOIL provides a common view of these varied interaction styles by assuming that every protocol provides at least two functions: the ability to receive a search and indicate how many results that search matches, and the ability to retrieve a number of results (e.g. give me the next 20 results). Although this means that HotOIL does not use the full range of capabilities of some protocols, it allows the greatest number of databases to be accessed.

Query Abstraction

Networked databases support a variety of query languages, ranging from simple keyword style searches supported by web search engines, through to the complex SQL queries supported by relational databases.

HotOIL gives the illusion that all of these databases provide a single style of search interface: fielded boolean queries like those used in online library catalogues. That is, queries that ask for keywords to appear in certain fields and that use logical connectors to separate parts of the query. For example,

Title: “Preamble” AND Author: “John Howard”

asks for results that have the word “Preamble” in the Title field, and the phrase “John Howard” in the Author field.

The illusion that all databases support this type of query is performed during the query translation phase of a HotOIL distributed search. During initial set-up HotOIL is told how to translate a fielded boolean query into the query structure understood by each database. This translation will typically be configured differently for each database, and represents a substantial amount of work on behalf of the HotOIL administrator. This configuration work, however, gives the user an illusion of uniformity of the underlying networked databases.

Query Attributes

Different networked databases provide different fields that can be search on. For example, a library database may support searching on the fields "title", "author", and "subject", whereas a web search engine may only support searching on "title" and "keywords" fields.

HotOIL provides the illusion that all of the databases it queries support the same set of search fields by firstly allowing the HotOIL administrator to select a set of search fields that are shown to the user. The administrator then configures a query translation for each database to map these user fields into fields understood by that database. This is illustrated in Figure 5 below.

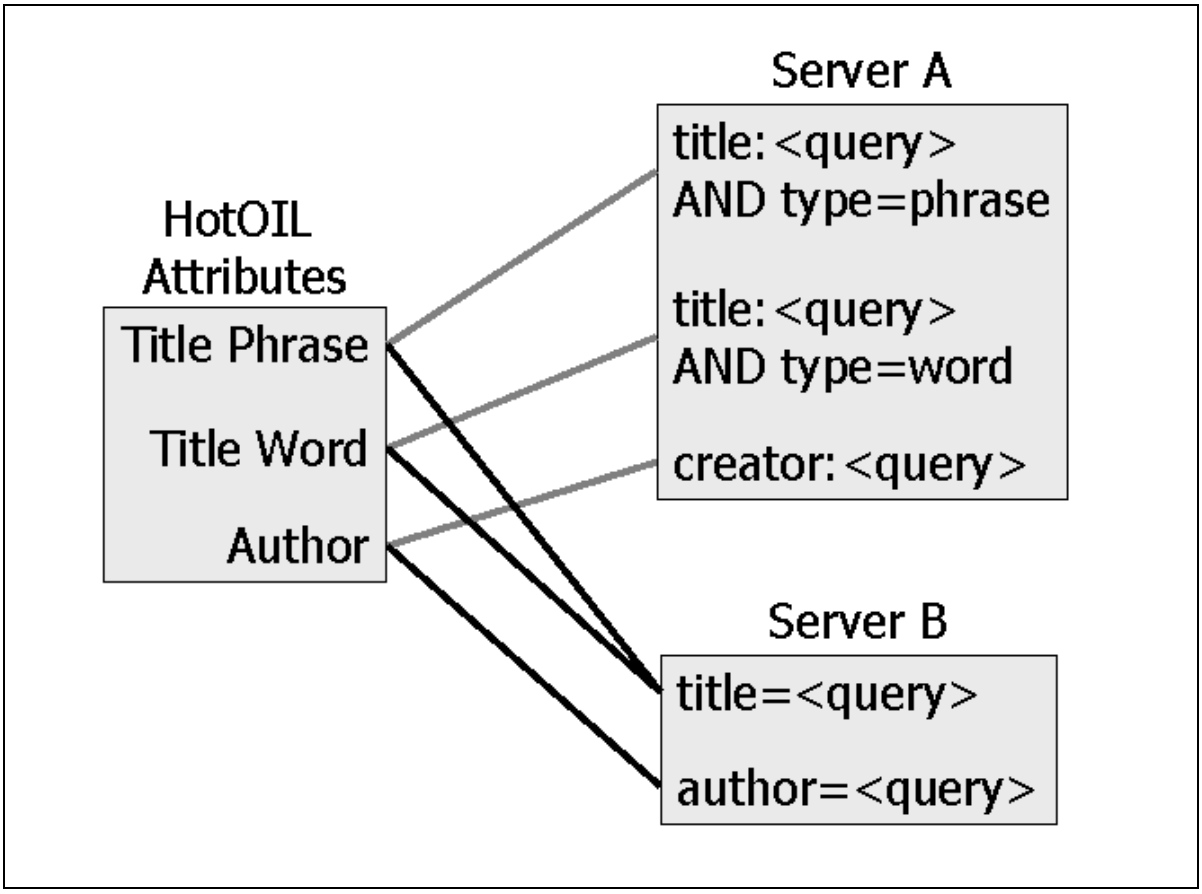

Figure 5: Query Attribute Translation

Result Abstraction

Networked databases return a wide variety of result formats. HotOIL gives the user an illusion that all networked databases support the same result format by translating each result into a common format: qualified Dublin Core[3]. Dublin Core is a metadata set initially intended to facilitate discovery of electronic resources through simple description of those resources. It consists of fifteen descriptive elements that have a have simple and commonly understood semantics. These elements have proven to provide the basis for good cross-domain description. Descriptive records from many communities have been (at least partially) translated into Dublin Core descriptions.

As before, the HotOIL administrator configures the system to translate results from individual networked databases into qualified Dublin Core records. Results from web search engines and ODBC databases contain a wide variety of fields. For this reason, the translation of these results is done on a per-server basis. Other networked databases return results conforming to descriptive standards (e.g. results from library databases often conform to the USMARC standard). In this case, the translation for all such results can be configured once, independently of the server they are retrieved from. The result translation is illustrated in Figure 6.

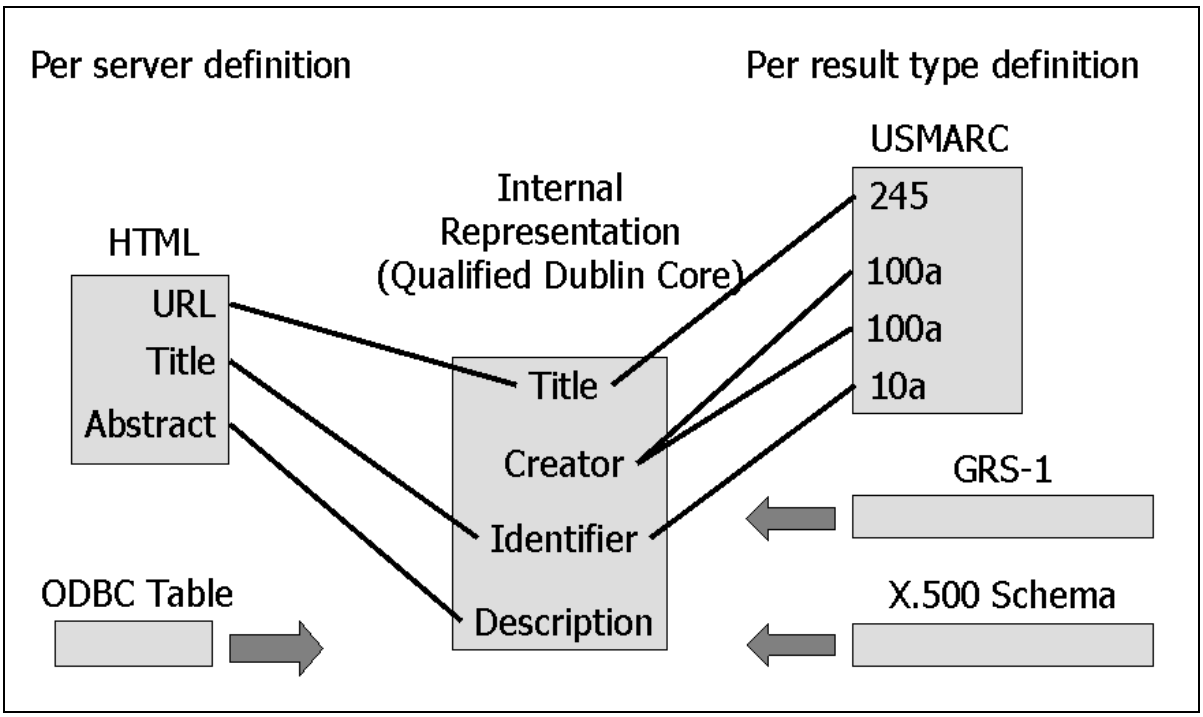

Figure 6: Result Translation

Applications and Experience

HotOIL has been used to provide integrated interface to a number of online databases:

- The DSTC HotOIL demonstration site[5] uses HotOIL to provide a unified interface to Australian libraries, museums, web search engines, legal, business, and government information.
- The ZAVIER project[6] used HotOIL to demonstrate the feasibility of using Z39.50 to search the databases of major Victorian cultural organisations.
- The ZedWeb project[7] used HotOIL to provide a public service that integrated Australian Z39.50 servers situated in libraries.

Each of these applications had varying amounts of user acceptance, depending on the amount that the user communities accepted the HotOIL abstractions. HotOIL translates user queries into queries on each database, and translates returned results into a common format. Due to the diversity of search fields and result formats supported by the databases being queried, this translation is not always exact. For example, a user query for a "Subject Word" could be translated exactly for one library database, but may only be approximated as a "Subject Phrase" query on another library database that does not support "Subject Word" searches. This type of search approximation is often accepted by general web users, probably due to the greater number of databases that HotOIL provides access to. Search approximation is, however, less readily accepted in a more formal discovery communities such as the library community, where a search for an author is expected to only return results for that author.

References

1. *"DSTC Resource Discovery Project."* http://www.dstc.edu.au/RDU
2. *"AT1 Homepage."* http://www.at1.com/

3. "The Dublin Core Metadata Initiative." http://purl.org/dc/
4. *"The HyperIndex Search Engine."* http://www.dstc.edu.au/Research/Projects/hib/
5. *"DSTC HotOIL Demonstration."* http://flare.dstc.edu.au:8888/hotoil.asp
6. Beaumont, Anne. *"ZAVIER - Wider than Libraries, Deeper than the Web."* VALA 2000 Conference, February 2000.
7. *"ZedWeb Project."* http://www.dstc.edu.au/Research/Projects/zedweb/