Paradigms for Resource Discovery

Michael F. Schwartz, University of Colorado - Boulder
schwartz@cs.colorado.edu

For the past five years, the Networked Resource Discovery Project at the University of Colorado has been investigating two broad problems: automated resource characterization, and flexible information distribution/search. Our approach to these problems is based on the *Two Phase Discovery* paradigm, in which both characterization and distribution/search are dynamic, exploratory processes. Rather than depending on manually registered data about resources, we use mechanisms that characterize resources by extracting data from the resources themselves, or from other existing sources of information. Rather than using a hierarchy or other relatively static organizational structure for distributing and searching information, we use information where it naturally resides.

Netfind provides an example of Two Phase Discovery, in the context of Internet white pages. Netfind builds a "seed" database that maps geographical and organizational keywords to Internet domain information, by continually monitoring a number of data sources, including USENET messages, UUCP maps, NIC WHOIS data, logs from FTP and other services, and information supplied by users. When a search is requested, Netfind consults the seed database to obtain domain search hints. If the database matches fall within more than three naming domains, the user is asked to select at most three domains to search. Netfind then contacts the Domain Naming System, to locate authoritative name server hosts for each domain. The idea is that these are often central administrative machines that have accounts and/or mail forwarding information for many users at a site. Each of these machines is then queried using the Simple Mail Transfer Protocol, in an attempt to find mail forwarding information about the specified user. If such information is found, the located machines are searched using the "finger" protocol. Netfind employs a number of search strategies to accommodate cases where only a subset of the remote information services is available. Moreover, we have developed a number of mechanisms to improve the quality of the seed database, based on semantics, redundancy, and context. These adaptive search and characterization mechanisms allow Netfind to locate more people than any other Internet directory service.

As a second example of Two Phase Discovery, we are developing mechanisms that extract keywords and human browseable summaries of file data, in a file type-specific fashion. This effort is meant to complement the capabilities present in WAIS, which indexes only text. Our goal is to provide mechanisms that can index more general types of information, and produce smaller indices. One summarizer we have developed extracts author, title, and abstract information from troff and TeX documents. A second summarizer samples bitmaps down to icon size for user browsing. A third summarizer extracts keywords in "manpages" associated with executable files. We are also building summarizers for directories, archives, program source code, database dictionaries, audio files, library code, mail and news messages, revision controlled data, PostScript documents, and files with nested structure (such as "tar.Z" files).

More recently, we have begun investigating a new paradigm for resource discovery, in the context of the Internet Research Task Force research group on Resource Discovery and Directory Service. To motivate this paradigm, we observe that existing Internet resource discovery tools fall roughly into two categories: organizational systems, such as Prospero, WorldWideWeb, Alex, and Gopher; and search systems, such as WAIS, Archie, Netfind, and Knowbots™. While there are gateways between many of these systems, there is no way for a user to request a "flat" search, and then explore a structured space surrounding the located resources. For example, it would be quite powerful if one could use Archie to locate a particular piece of software, and then be placed in a Prospero view that includes that software plus related software and documents. Similarly, WAIS might let a user search for some data, and then place the user on a WWW link that includes programs that can display the data in various ways.

Beyond supporting a combined search/browse paradigm, a second problem that must be addressed is how to support efficient global searches. While systems like Archie currently support global searches, eventually there will be too much information to place entire indicies on a single machine. For example, we estimate that within five years, Archie's current indexing mechanism will require a 16 gigabyte index, to hold information about just one archive per Internet domain. This does not include other types and sources of networked information, or indicies that contain more information than just file names. Clearly, Archie and other "flat" directory services will need to focus on information for smaller communities, such as individual countries or particular scientific disciplines.

Both of these problems (the need for a combined search/browse paradigm, and the need to support global searches) have led us to a new paradigm, called *perspective discovery*. The idea is motivated by editorial perspectives, as found in newspapers and moderated electronic news groups. A perspective is an entry point into a structured information space, based on a unifying theme (e.g., a perspective about biological software, or a perspective about current cultural events). Each perspective will register pieces of a structured information space into a flat searchable space. There will be many entry points, with automated means of advertising, interrelating, and searching them. The perspective discovery paradigm is different from the directory-of-directories approach, since perspective discovery does not imply a cascading hierarchy (where eventually there are directories of directories of directories, etc.). Scalability will be achieved through specialization, rather than through global nesting.

Papers about the Networked Resource Discovery Project are available by anonymous FTP from ftp.cs.colorado.edu, in the directory pub/cs/techreports/schwartz/RD.Papers. Users can try out Netfind by telnet/rlogin to bruno.cs.colorado.edu, and logging in as "netfind" (with no password). The mailing list archive for the Internet Research Task Force research group on resource discovery and directory service is available by anonymous FTP from ftp.cs.colorado.edu, in pub/cs/misc/irtf-rd/rd-archive.