# Characterizing Information for Internet Resource Discovery

Michael F. Schwartz
University of Colorado - Boulder
Published in Internet Society News 1(2), Spring 1992

Resource discovery involves two basic problems: characterizing the resources of interest using name/attribute descriptions, and distributing this information so that it can be searched flexibly and efficiently. In this article we consider the characterization problem. We will consider attribute distribution and search in a future column.

The traditional approach to resource characterization is manual data registration, as used when loading an X.500 Directory System Agent. (A DSA might be loaded automatically from another database, but the original database will almost invariably have been created manually.) Manual registration is also used by Prospero, where users create "views" of existing files to organize them into related collections.

Manual characterization provides good control over what data is registered for each resource. This may be important for controlling what data is visible, or for providing highly conceptual descriptions. On the other hand, manual characterization is painstaking and error-prone in a large, dynamically changing environment like the Internet, and the information produced can quickly become dated and incomplete.

To automate the process, a popular approach is extracting keywords from the contents of documents. This technique is used by WAIS and bibliographic indexing systems like the UNIX "bib/refer" system. A simpler approach is to generate keywords from file and directory names. This approach is used by Archie and tools like the UNIX "find" command.

Automatic characterization can produce poor quality keywords, causing searches to match too few or too many resources. To improve keyword quality one can use techniques that exploit the context, semantics, or redundancy of the information being characterized. For example, WAIS eliminates common words, extracts root forms of word variants, and generates relevance weightings by frequency of occurrence.

Since WAIS operates on very general information (textual documents about any topic), it only exploits characteristics of human language text. Given a more focused resource discovery problem, more sophisticated characterization is possible. For example, netfind supports Internet "white pages" user searches by extracting keywords from a large, contextually focused source (a "seed database" of host and organization lines gathered from USENET news headers), and by honing the quality of this information in several ways. Biasing organization name selection from the seed database by frequency of occurrence eliminates many invalid search targets. Exploiting relationships between the seed data and sources of data consulted at the sites where searches are performed (the Domain Naming System and the Simple Mail Transfer Protocol) narrows the scope of searches to small, promising subsets.

As a second example, my research group developed a set of graph algorithms to locate people with particular interests, to support "colleague discovery". In essence, the algorithms exploit redundancy of graph neighbor information from the history of "From/To" lines monitored in electronic mail communications. This technique could be applied in other realms as well, for example to discover relationships between data in a file system.

As a third example, my research group is developing mechanisms that produce keywords and human browseable summaries of file data, in a file type-specific fashion. One summarizer extracts author, title, and abstract information from troff and TeX documents. Another summarizer samples bitmaps down to icon size for user browsing. A third summarizer extracts keywords in "manpages" associated with executable files. We will develop more summarizers over time.

Information about Archie is available from quiche.cs.mcgill.ca, in archie. Information about netfind, shared interest discovery, and file summarizers is available from ftp.cs.colorado.edu, in

pub/cs/techreports/schwartz/RD.Papers. Information about Prospero is available from cs.washington.edu, in pub/prospero.tar.Z. Information about WAIS is available from think.com, in wais. Information about the PSI X.500 pilot is available from uu.psi.com, in wp.