Z39.50 and Distributed Searching

Ralph LeVan
Research Scientist
OCLC Online Computer Library Center
ralph_levan@oclc.org

Why Distributed Searching?

In searching the Internet, your choices as a user, or user agent, are to go to one site, multiple sites or all sites. Given the number of users on the Internet today, having them all search all sites seems impractical. Alternatively, depending on only one site seems dangerous. A change in availability, due to any number of reasons, leaves the user with nowhere to go. In addition, a single site leaves you at the mercy of a source that might decide to start charging for it's service. Clearly, having multiple sites is desirable.

But, searching multiple sites is a problem if HTTP is the protocol used. Each site can have it's own interface. While this might not seem to be a serious problem for humans interacting with these servers, it is a significant problem for robots. A standard for searching those sites is desirable. Z39.50 is such a standard.

Why Z39.50?

Z39.50 is an international standard developed specifically to support search and retrieval of text-like objects. It was originally developed in the library community with its well developed requirements for searching and retrieving bibliographic data and has been applied to much richer data, including full-text and image databases. Its strengths include:

Interoperability

Z39.50 supports the unambiguous transmission of queries between clients and servers. It currently supports Boolean queries extended to allow proximity operators. Work is underway to support relevance feedback queries unambiguously. Either of these grammars is a big step beyond having the user type something and some magic happens at the search engine, which is what we have today.

It is important to stress that Z39.50 is a peer level application protocol. It is NOT an end-user protocol. Just because Z39.50 uses Boolean queries does not mean that the user needs to. It's up to the client software to convert its user's queries into interoperable Z39.50 queries.

Specificity

Z39.50 allows for specificity in searching that is unavailable in most of the search engines on the net today. When I search for "Mark Twain", am I looking for books about Mark Twain or by Mark Twain? This distinction can be expressed in Z39.50. Try that at Yahoo.

In addition, Z39.50 supports a rich grammar for retrieving portions of a record; not simply the entire record or the first N bytes of the records, but specific fields from a record. This allows for the retrieval of, for instance, only the author and title of a record, reducing significantly the amount of data being transmitted.