WebFind: Automatic Retrieval of Scientific Papers
Over the World Wide Web

Alvaro E. Monge  Charles P. Elkan

Computer Science and Engineering Department, 0114
University of California, San Diego
La Jolla, California 92093, USA
{amonge, elkan} @cs.ucsd.edu

The World Wide Web (WWW) has become the most convenient place in which to disseminate information to internet users. Unfortunately, due to the continuous growth, a lot of the information in the WWW is not of any interest to most users. This leaves many users stranded in a sea of information, with no clue of where to find the information relevant to them. This phenomenon is what has propelled the research in resource discovery. The work in this area has given way to a large number of software applications which in some way or another attempt to provide the user with the capability of locating information that is important to them. These software applications go by different names, among which are spiders, agents, robots, wanderers, worms, and others. This paper describes WebFind, a tool which tries to solve a user's quest for online copies of a published paper.

WebFind is a new software tool for locating authors and the full text of their papers on the internet. Users can easily find author and published paper information scattered throughout the ever growing, dynamic, distributed WWW. WebFind first extracts information from Melvyl databases. Melvyl is a traditional information retrieval system for magazine and journal articles. The information collected is used to locate an author's home in the World Wide Web, which in turn is used to direct the search for the author's published paper. WebFind is useful in scientific areas where authors (who are on the internet) usually make some form of the published paper available online through the WWW.

Melvyl, an important University of California resource, is comprehensive and systematic but it has limited content. The WWW, on the other hand, can have unlimited content, but it is chaotic. WebFind is a unique combination of these complementary information sources.

WebFind is unlike many internet tools which attempt to solve the resource discovery problem. Typical applications first create off-line indices of the WWW and then use a search engine on the indices. In WebFind we use external information sources to aid the discovery of information on the WWW. There are three main types of sources used: library resource, white page directories, and full text paper archives; Melvyl, Netfind, and authors' home pages, respectively, are the instances of each resource that are used in WebFind. This approach allows us to significantly narrow down the WWW space which is searched.

One of the problems with doing resource discovery in an environment as large as the WWW is to select the starting point(s) for the search which will quickly lead to the main sources of information being sought. We believe that an author's home on the internet is the best starting point to use in WebFind. To determine an author's home, first we use information which is provided by the library resources such as Melvyl. Specifically, we use the affiliation field for a principal author of a paper. The affiliation field is a text string representing the author's place of work. This information is used to do a search in the white page directory Netfind. The search results provide us with a set of candidate internet domains where we might locate the author. In these internet domains, WebFind discovers WWW servers which are used for the discovery of the author's home page and of the paper.

Throughout the discovery phase, WebFind must make decisions regarding the direction in which to continue the discovery for additional information. We need to select the most promising link among a set of candidate information links. In WebFind we are investigating the use of heuristic methods to measure each link's relevance and then continue the discovery phase only along those links with the highest relevance measurement.

Another research problem that we are addressing is the field matching problem. An instance of this problem in WebFind appears when we want to match the affiliation field of a Melvyl database with the field describing an internet domain in the Netfind database. There are other instances of this problem occurring throughout WebFind. This is an important and difficult problem to solve as it occurs in different forms and no single algorithm may work for all the instances that we have encountered so far. Instead, our approach to this problem will be to apply heuristic techniques to solve it.