FishNet: Finding and Maintaining Information on the Net

Paul De Bra[1] and Pim Lemmens
Department of Mathematics and Computing Science
Eindhoven University of Technology
PO BOX 513, 5600 MB Eindhoven
The Netherlands
{debra, wsinpim}@win.tue.nl

Abstract: This short paper presents a tool for keeping a hotlist or homepage up to date. It combines two existing tools:

- MOMspider [Fielding 1994] is a tool to verify whether links are still valid and whether documents they point to have been modified or moved.
- Fish-Search [De Bra & Post 1994b] is a search tool for finding new interesting documents in the neighborhood of a given set of (addresses of) documents.

FishNet keeps track of the evolution of a domain of interest by periodically running MOMspider and FishSearch and presenting the user with newly found documents. The user can put documents in the hotlist or in a reject list. This positive and negative feedback are constantly used to improve the precision of the search.

1. Overview and Motivation

Beginning World Wide Web users start collecting addresses of interesting documents they find, by storing them in the browser's bookmark list. Later they may also move this information to their home page to share their findings with the world. Keeping the list consistent and adding addresses of new interesting documents to ensure that the list remains a valuable resource can quickly become a full-time job.

Existing large search engines such as Alta Vista, Excite or Lycos do not offer a solution to this problem, because no small set of keywords is sufficiently discriminating to perform a search without returning a high rate of non-relevant documents. Browsing through the answers of these engines, in search of some new interesting documents, often takes much more time than it's worth.

The **FishNet** toolkit offers a platform for automating hotlist maintenance. It offers the following features:

- Verification of link consistency and of updates to documents through the standard MOMspider package [Fielding 1994] (developed by Roy Fielding, not by us).
- Multi-threaded Fish-Search navigation engine [De Bra & Post 1994a, De Bra & Post 1994b] for finding new documents. This engine can be extended by means of external filters for determining relevance of documents. FishNet contains a set of such filters.
- History of documents previously marked as relevant or non-relevant, to improve the selection of new documents.
- (HTML) Report generator through which the bookmark list or home page can be updated.

By means of FishNet the user can ensure that the list or home page always contains valid links, that the descriptions of these documents remain accurate, and that new documents on the topics of interest are found and added to the list. Using FishNet can reduce the full-time information discovery job to just a few minutes a day.

For use with FishNet the Fish-Search tool has been improved significantly since its original development back in 1994. The most important new features of Fish-Search are:

- Fish-Search used to be integrated into a Web-browser. The new version is a stand-alone program that can be activated as a CGI-script.

[1] Paul De Bra is also affiliated to the University of Antwerp, and with the "Centrum voor Wiskunde en Informatica" in Amsterdam.

- Use of multi-threading (through the standard W3C library) to load documents from different servers in parallel.
- Fish-Search now obeys the Robot-Exclusion protocol [Koster 1994].
- External filters can be used in addition to the built-in keyword- regular-expression and approximate maching algorithms. These filters must reside in a special directory to avoid abuse.

2. Using FishNet

FishNet is normally run at night, from the Unix *cron* utility. It first activates MOMspider to find which documents need closer examination. It then performs the following actions:

- For documents that have been relocated, FishNet updates the hotlist to note the new address.
- Documents that have been modified are starting points for a search-run, in order to look for new interesting documents. FishNet comes with a set of filters for finding related documents.
- For documents that have been deleted, or possibly moved without leaving a relocation, FishNet will start a search from the root of the server(s) these documents used to be on. If the documents were simply moved chances are they will be found again.
- New potentially interesting (URLs of) documents are combined into a report for the user. From the report the user can move the documents to the hotlist or to a reject list.

If FishNet is run through a proxy cache [De Bra & Post 1994a] and the user's browser goes through the same cache, the documents that need to be examined by the user can be retrieved very efficiently.

Some systems try to locate information based on a user profile which is deduced from the user's browsing behaviour [Brown & Benford 1996]. Since a user may be interested in more than one subject, it is more difficult to determine which information satisfies the user profile than when only one specific topic is used. Some packages like those described in [Maarek & Shaul 1995] and [Gaines & Shaw 1995] try to distribute documents over a set of topics automatically. FishNet does not deal with multiple areas of interest. Instead, for different subjects separate lists or Web pages should be created, and each of the lists is treated separately by FishNet. In order to do so, FishNet identifies each "job" by the user identification and the URL of the list.

We believe FishNet is a valuable tool for teaching students about hotlist maintenance. For mainstream end-users some commercial maintenance and search tools are entering the market, with more user-friendly interfaces.

3. References

[Brown & Benford 1996] Chris Brown, Steve Benford, Tracking WWW Users: Experience from the Design of HyperVis, WebNet'96, World Conference of the Web Society, pp. 57–63, San Francisco, 1996.
(URL: http://aace.virginia.edu/aace/conf/webnet/html/174.htm)

[De Bra & Post 1994a] P. De Bra, R. Post, Information Retrieval in the World-Wide Web: Making Client-based searching feasible, First International World Wide Web Conference, Geneva, 1994.
(URL: http://www.win.tue.nl/win/cs/is/reinpost/www94/www94.html)

[De Bra & Post 1994b] P. De Bra, R. Post, Searching for Arbitrary Information in the WWW: the Fish-Search for Mosaic, (Poster and demo at) Second International World Wide Web Conference, Chicago, 1994.
(URL: http://www.win.tue.nl/win/cs/is/debra/wwwf94/article.html)

[Fielding 1994] Roy T. Fielding, Maintaining Distributed Hypertext Infostructures: Welcome to MOMspider's Web, First International World Wide Web Conference, Geneva, 1994. (URL: http://www.ics.uci.edu/pub/websoft/MOMspider/WWW94/paper.html)

[Gaines & Shaw 1995] Brian R. Gaines, Mildred L.G. Shaw, WebMap, Concept Mapping on the Web, Fourth International World Wide Web Conference, Boston, 1995. (URL: http://ksi.cpsc.ucalgary.ca/articles/WWW/WWW4WM/)

[Koster 1994] Martijn Koster, A Standard for Robot Exclusion, (Unofficial standard obeyed by most robots on the Web).
(URL: http://info.webcrawler.com/mak/projects/robots/norobots.html)

[Maarek & Shaul 1995] Y.S. Maarek, I.S. Ben Shaul, Automatically Organizing Bookmarks per Contents, Fifth International World Wide Web Conference, Paris, 1995, Computer Networks and ISDN Systems, Vol. 28, p. 1321–1335.
(URL: http://www.ics.forth.gr/ telemed/www5/www185/overview.htm)