FERRET: An Intelligent Assistant for Internet Searching

Juhua Zhou and Jacky Baltes[1]

University of Auckland, Auckland, New Zealand,
j.baltes@auckland.ac.nz,
WWW: http://www.citr.auckland.ac.nz/ jacky

This paper describes the design and implementation of Ferret, an information-seeking assistant that helps a user find information on the World Wide Web. It analyzes and automatically clusters the returned pages from a search engine.

The Internet today contains millions of WWW pages containing a huge amount of information. In theory, this information is only a few keystrokes away. Automatic search engines are the most popular method for finding information. However, these search engines often return thousands of results even for specific queries. A lot of the returned pages are not relevant.

One reason for the large number of returned pages is that keywords often refer to different concepts. For example, cookies can refer to a method for maintaining state information on a WWW browser or oven-baked goodies.

The goal of FERRET is to automatically determine different concepts for the returned WWW pages and to cluster the pages into different groups. FERRET completes these tasks in several steps. Firstly, it lets users submit their search query and the desired page type and fetches the result pages from the search engine. Secondly, it filters the returned results according to the selected type. Then it extracts keyphrases from the pages and represents them with generated attributes. Finally it clusters the pages and presents the results in dynamic web pages where users can browse the results.

The paper primarily investigates three issues: (1) extracting keyphrases from web pages, (2) extracting features and creating a representation of web pages, and (3) clustering web pages. The paper contrasts the related machine learning techniques and selects KEA to extract keyphrases from pages and AutoClass to cluster pages. The representation of pages has been the main focus of the thesis. To build the representation for pages, the paper examines the possible resources of extracting features from text content and HTML components of web pages and selects extracted keyphrases and links.

The evaluation shows that FERRET is able to cluster web pages with high accuracy on both correctly clustered page numbers and class number and significantly outperforms random clustering.

Based on initial exploration, the authors are optimistic that FERRET sets up a mechanism by which the software agent can analyze the web pages and cluster them. FERRET can guide search engine users to access relevant pages and to avoid reading irrelevant pages and thus lead users to efficient searching. The clustering can provide users with a concept map of knowledge areas.