Information Searching and Exploring on WWW:

Applying Clustering and Genetic Algorithm

Chia-Hui Chang
Department of Computer Science and Information Engineering
National Taiwan University, Taipei 106, Taiwan
chia@ails1.csie.ntu.edu.tw

The World Wide Web, fueled by the cost-free and easy self-publication characteristics, has exhibited exponential growth than any other system over the past three years. The vast information published on web obviously inflames the problem of "information quality". From the early stage of resource discovery to latter information retrieval and access interface, the process of information searching has often been a frustrating and time-consuming work.

Of the current searching tools on web, two approaches are generally employed: one is to organize information as subject catalog and let the users browse through the subject hierarchy, the other is via keyword searching and documents ranking provided by search engines. Subject browsing, like Yahoo, Galaxy, InfoSeek, and Magellan, etc., is a good starting point for finding what has been published on the Web for a broad topic, but is sometimes time-consuming and difficult to make a second guess. Keyword-based searching, for example, AltaVista, WebCrawler, and Lycos, etc. on the other hand, provides an immediate and a larger number of documents than subject browsing tool,. The problem of this model is that it is not always easy to formulate a request using searching features like Boolean controls or structure queries (which are advanced searching tips or options for better searching results). Thus natural language like's queries are often of the most popular. Unfortunately, according to Pinkerton's study in 1994 [4], the average number of keywords specified in each query is only 1.5, which usually leads to an unreasonable number of “matching” documents. Though different ranking algorithms have been used to rank the documents according to their relevance to the query, the effects is subjected to controversy since generally few keywords are specified in the inputs due to insufficiency of knowledge about the indexed keywords in the database. Therefore, it often turns out to be a tiring browsing process through the ranked documents.

In our approach, the searching results from other word-based search engines are processed using *clustering techniques* and are summarized to extract related words. The insight here is that the missed information in the user′s input may be embedded in the results of the first query. If we can summarize the “matching” documents as groups, the users could more easily identify the domains relating to what they really want. However, the main difficulties of this approach lie in cluster digesting and keyword extraction for a smooth query process in word-based search model. Therefore, we propose a combination technique of *majority principle* and *cue validity* (an idea first introduced by Bourne and Restle in 1959) to overcome these problems. The immediate rewards of our approach, as we will see, are the speed-up of the searching process and the saving of the browsing efforts compared to the available searching tools online. The query process is no longer a conditioned response but a feedback and learning interaction.

Further more, the learned information could enable the construction of an autonomous agent which searches for the new published documents regarding to the user’s interests in the background. A simple while effective method we adopted here is genetic algorithm, where the queries in the same interest domain are selected to generate new query expressions via genetic operators. This idea solves the problem since different combination of words about one topic derives different search outcomes and ranking results. Hence, by generating new query expression, we would be able to find out new information that has not yet been discovered.

References

[1] Chang, C.H. and Hsu, C.C. "Multi-Engine Search Tool with Clustering", in Proceedings of the Sixth International WWW Conference, Apr. 7-11, 1997, Santa Clara, CA, pp.257-264

[2] Chang, C.H. and Hsu, C.C. "InfoSearch: An Information Searching Agent applying Clustering and Cue Validity", submitted for publication. Prototype at *http://ails6.csie.ntu.edu.tw/~chia/Sep.html*

[3] Cutting, D.R. et al. "Scatter/Gather: A Cluster-Based Approach to Browsing Large Document Collections," Proceeding of International SIG Information Retrieval, ACM Press, New York, 1992. pp. 318-329.

[4] Pinkerton, B. "Finding What People Want: Experiences with the WebCrawler," the 2nd International WWW Conference, 1994, Chicago.