A Tool to Personalize the Ranking of the Documents Returned by an Internet Search Engine

Wadee S. Al halabi*, Miroslav Kubat, Moiez Tapia
Department of Electrical and Computer Engineering, University of Miami
* Department of Computer Science, King Abdulaziz University
wsalhalabi@kau.edu.sa, *mkubat@miami.edu, mtapia@miami.edu*

Abstract
Internet search engines identify web pages that contain user-specified keywords, and then rank these pages according to their (heuristically assessed) relevance to the user's query. In this paper, we investigated the possibility of evaluating this relevance by the similarity of the returned web page to web pages previously visited by the same user: these previously visited pages thus serve as positive training examples from which a machine-learning program induces an internal model of the user's interests and preferences. We describe two different ways to represent this model. Our experiments indicate that this approach can indeed improve the ranking.

1. Introduction
Upon receiving a user's query in the form of a list of keywords, a typical Internet search engine identifies those web pages (documents) that contain the query keywords, and then estimates the relevance of these documents to the user's query. The engine then offers the user a list of hyperlinks pointing to these documents, ordered according to the relevance to the query. Historically, a document's relevance to the user's needs has been calculated from (1) the frequencies of the keywords in different parts of the documents, (2) the popularity of these documents (measured by the time spent on them by an average user), and (3) the structure of the links to and from related web pages.
In the research reported here, we build on the obvious fact that each user has somewhat different interests that are to a great degree reflected by the contents of the web pages he or she has visited in the past. Alternatively, the user may want to indicate these preferences by one or more web sites that he or she deems relevant. Building on this assumption, we have developed a tool that accepts the hyperlinks obtained by submitting the user's query to an off-the-shelf search engine, downloads the corresponding documents, and then uses the knowledge induced from documents previously visited by the same user to calculate the weights to be assigned to the returned documents.
Finally, the system re-orders the returned documents according to these weights, and returns to the user the new ranking instead of the one recommended by the original search engine.
To justify the hypothesis that the new ranking better reflects the user's needs, we have conducted several experiments with two alternative ways to exploit the user-supplied reference documents. From these, our own LVA algorithm appears to outperform the classical VSA technique known from the discipline of information retrieval both in the quality of ranking and in the computation expenses.

2. Previous Research
In the past few years, several research groups have studied various methods to personalize the outputs of search engines. Some of these groups employed machine learning techniques. To establish the context for our own research, let us briefly summarize this earlier work.
Fan et al. [1] developed a system that personalizes its output, and demonstrated that the personalization improved its utility. Following this idea, literature reports several attempts to improve the behavior of information-retrieval systems by the use of machine-learning techniques. One possibility is reported by Boyan et al. in [2] who use an induction algorithm to find a way to assign weights to the returned documents. These weights are based on the location of the keywords in the document and are used for the re-ordering of the documents. Cui et al. [12] improve the quality of query processing by analyzing the search engine's logs of previous queries (submitted by the