Chapter 10

Information Customization using SOMSE: A Self-Organizing Map Based Approach

Mohamed Salah Hamdi
Ahmed Bin Mohammed Military College, Qatar

ABSTRACT

Information overload on the World-Wide Web is a well recognized problem. Research to subdue this problem and extract maximum benefit from the Internet is still in its infancy. Managing information overload on the Web is a challenge and the need for more precise techniques for assisting the user in finding the most relevant and most useful information is obvious. Search engines are very effective at filtering pages that match explicit queries. Search engines, however, require massive memory resources (to store an index of the Web) and tremendous network bandwidth (to create and continually refresh the index). These systems receive millions of queries per day, and as a result, the CPU cycles devoted to satisfying each individual query are sharply curtailed. There is no time for intelligence which is mandatory for offering ways to combat information overload. What is needed are systems, often referred to as information customization systems, that act on the user’s behalf and that can rely on existing information services like search engines that do the resource-intensive part of the work. These systems will be sufficiently lightweight to run on an average PC and serve as personal assistants. Since such an assistant has relatively modest resource requirements it can reside on an individual user’s machine. If the assistant resides on the user’s machine, there is no need to turn down intelligence. The system can have substantial local intelligence. In an attempt to circumvent the problems of search engines and contribute to resolving the problem of information overload over the Web, the authors propose SOMSE, a system that improves the quality of Web search by combining meta-search and unsupervised learning.

DOI: 10.4018/978-1-60566-970-0.ch005
INTRODUCTION

We live in an age of information abundance. The technology research firm IDC (http://www.idc.com/) determined that the world generated approximately 161 exabytes (i.e., around 161 billion gigabytes) of new information in 2006 (Bergstein, 2007); that’s many thousands of times the size of the U.S. Library of Congress. In July 2008, Google announced that it had indexed one trillion (as in 1,000,000,000,000) unique URLs and estimated that the web was growing at a rate of several billion pages per day (Google, 2008).

The plenitude of information, not its scarcity, defines the world we live in now. Historically, more information has almost always been a good thing. However, as the ability to collect information grew, the ability to process that information did not keep up. Today, we have large amounts of available information and a high rate of new information being added, but contradictions in the available information, a low signal-to-noise ratio (proportion of useful information found to all information found), and inefficient methods for comparing and processing different kinds of information characterize the situation. The result is the “information overload” of the user, i.e., users have too much information to make a decision or remain informed about a topic. Seeking to shift between dealing with information scarcity and information abundance is necessary. Locating information is not the problem; locating relevant, reliable information is the real issue.

Information overload on the World-Wide Web is a well recognized problem. Research to subdue this problem and extract maximum benefit from the Internet is still in its infancy. Managing information overload on the Web is a challenge and the need for more precise techniques for assisting the user in finding the most relevant and most useful information is obvious. With largely unstructured pages authored by a massive range of people on a diverse range of topics, simple browsing has given way to filtering as the practical way to manage Web-based information. Today’s online resources are therefore mainly accessible via a panoply of primitive but popular information services such as search engines.

Search engines are very effective at filtering pages that match explicit queries. Search engines, however, require massive memory resources (to store an index of the Web) and tremendous network bandwidth (to create and continually refresh the index). These systems receive millions of queries per day, and as a result, the CPU cycles devoted to satisfying each individual query are sharply curtailed. There is no time for intelligence which is mandatory for offering ways to combat information overload.

Search engines rank the retrieved documents in descending order of relevance to the user’s information needs according to certain predetermined criteria. The usual outcome of the ranking process applied by a search engine is a long list of document titles. The main drawback of such an approach is that the user is still required to browse through this long list to select those that are actually considered to be of interest. Another shortcoming is that the resultant list of documents from a search engine does not make distinctions between the different concepts that may be present in the query, as the list inevitably has to be ranked sequentially. The problem lies mainly in the presentation of the list of document titles. These documents are usually listed serially irrespective of the similarity or dissimilarity in their contents - that is, it does not make distinctions between the different concepts. Thus, two documents appearing next to each other in the list may not necessarily be of a similar nature and vice versa. As the list of documents grows longer, the amount of time and effort needed to browse through the list to look for relevant documents increases.

What is needed are systems, often referred to as information customization systems (Hamdi, 2006a, 2006b, 2007a, 2007b, 2008a, 2008b, 2008c), that act on the user’s behalf and that can rely on existing information services like search