Personalized Search on the World Wide Web

Alessandro Micarelli¹, Fabio Gasparetti¹, Filippo Sciarrone¹, and Susan Gauch²

¹ Department of Computer Science and Automation
   Artificial Intelligence Laboratory
   Roma Tre University
   Via della Vasca Navale, 79 - 00146 Rome, Italy
   {micarel, gaspare, sciarro}@dia.uniroma3.it

² Information & Telecommunication Technology Center
   University of Kansas
   2335 Irving Hill Road, Lawrence Kansas 66045-7612
   sgauch@itc.ku.edu

Abstract. With the exponential growth of the available information on the World Wide Web, a traditional search engine, even if based on sophisticated document indexing algorithms, has difficulty meeting efficiency and effectiveness performance demanded by users searching for relevant information. Users surfing the Web in search of resources to satisfy their information needs have less and less time and patience to formulate queries, wait for the results and sift through them. Consequently, it is vital in many applications - for example in an e-commerce Web site or in a scientific one - for the search system to find the right information very quickly. Personalized Web environments that build models of short-term and long-term user needs based on user actions, browsed documents or past queries are playing an increasingly crucial role: they form a winning combination, able to satisfy the user better than unpersonalized search engines based on traditional Information Retrieval (IR) techniques. Several important user personalization approaches and techniques developed for the Web search domain are illustrated in this chapter, along with examples of real systems currently being used on the Internet.

6.1 Introduction

Recently, several search tools for the Web have been developed to tackle the information overload problem, that is, the over-abundance of resources that prevent the user from retrieving information solely by navigating through the hypertextual space. Some make use of effective personalization, adapting the results according to each user's information needs. This contrasts with traditional search engines that return the same result list for the same query, regardless of who submitted the query, in spite of the fact that different users usually have different needs. In order to incorporate personalization into full-scale Web search tools, we must study the behavior of the users as they interact with information sources.

© Springer-Verlag Berlin Heidelberg 2007
There are three information access paradigms that users undertake each time they need to meet particular information needs on the Web hypertextual environment: searching by surfing (or browsing), searching by query and recommendation. Recommendation-based systems suggest items, such as movies, music or products, analyzing what the users with similar tastes have chosen in the past [67, 58], see Chapter 12 of this book [12] for details.

In searching by surfing, users analyze Web pages one at a time, surfing through them sequentially, following hyperlinks. This is a useful approach to reading and exploring the contents of a hypertext, but it is not suitable for locating a specific piece of information. Even the most detailed and organized catalogs of Web sites, such as YAHOO! DIRECTORY\(^1\) and the OPEN DIRECTORY PROJECT\(^2\), do not always allow users to quickly locate the pages of interest. The larger the hypertextual environment is, the more difficulty a user will have finding what he is looking for.

The other dominant information access paradigm involves querying a search engine, an effective approach that directly retrieves documents from an index of millions of documents in a fraction of a second. This approach is based on an classic Information Retrieval (IR) model [71] wherein documents and information needs are processed and converted into ad-hoc representations. These representations are then used as the inputs to some similarity function that produces the document result list. Further details about this basic approach can be found in Chapter 5 [55] and 2 [29] [55] of this book.

Information Retrieval has always been characterized by relatively stable information sources and sequences of possibly unrelated user queries. It is usually considered distinct from the Information Filtering (IF) process [59], where the user needs are stable and there are large volumes of dynamically generated collections of documents. The user’s interests in IF change relatively slowly with respect to the rate at which information sources become available. The Web is a highly dynamic environment, with information constantly being added, updated and removed, therefore IF prototypes seem to be the most appropriate choice on which to build Web search systems. Nevertheless, IF mostly employs complex representations of user needs and the time needed to perform the retrieval process, that is, matching the incoming stream of information with the model of user’s interests, is quite long. This slow response is one of the reasons why IF prototypes have not become a widespread tool to retrieve information from the Web. For a closer examination of the most important user modeling techniques developed for IF, see Chapter 2 of this book [29].

In the last few years, attention has focused on the adaptation of traditional IR system to the Web environment, and related implementations of personalization techniques. The former task is accomplished by periodically collecting newly-created documents through re-crawling, keeping the search system’s internal document index updated. This chapter discusses the second topic, personalization techniques and their implementation in real systems.

The two paradigms, searching by query and browsing, coexist: most of the times, browsing is useful when the user does not know beforehand the search domain keywords. Often, the user actually learns appropriate query vocabulary while surfing. Be-

---

1. http://dir.yahoo.com