Semi-automatic Creation and Maintenance of Web Resources with webTopic*

Nuno F. Escudeiro and Alípio M. Jorge

LIACC, Faculdade de Economia, Universidade do Porto
nfe@isep.ipp.pt,
amjorge@fep.up.pt

Abstract. In this paper we propose a methodology for automatically retrieving document collections from the web on specific topics and for organizing them and keeping them up-to-date over time, according to user specific persistent information needs. The documents collected are organized according to user specifications and are classified partly by the user and partly automatically. A presentation layer enables the exploration of large sets of documents and, simultaneously, monitors and records user interaction with these document collections. The quality of the system is permanently monitored; the system periodically measures and stores the values of its quality parameters. Using this quality log it is possible to maintain the quality of the resources by triggering procedures aimed at correcting or preventing quality degradation.

1 Introduction

Web characteristics, such as dimension and dynamics [17], place many difficulties to users willing to explore it as an information source. Moreover, information retrieved from the Web is typically a large collection of documents. A query in Google for "Artificial Intelligence" gives, today, a list of 95.000.000 results. Organizing this information conveniently improves the efficiency of its exploitation. To take advantage of the value contained in this huge information system there is a need for tools that help people to explore it and to retrieve, organize and analyze relevant information.

The satisfaction of an information need on the Web is usually seen as an ephemeral one-step process of information search (the traditional search engine paradigm). The user is usually not assisted in the subsequent tasks of organizing, analyzing and exploring the answers produced. Vivisimo (http://vivisimo.com) and Tumba! [25] are exceptions where the retrieved documents are automatically (and immediately) clustered according to syntactic similarity, without any input from the user, other than the keywords of the search query itself. We believe that it is also important to give the user the possibility of specifying how he or she requires the retrieved documents to be organized.

Another important aspect is the existence of persistent information needs. This is the case of many professionals, such as scientists, who need frequent updates about

* Supported by the POSC/EIA/58367/2004/Site-o-Matic Project (Fundação Ciência e Tecnologia), FEDER e Programa de Financiamento Plurianual de Unidades de I & D.

M. Ackermann et al. (Eds.): EWMF/KDO 2005, LNAI 4289, pp. 82 – 102, 2006.

their area of activity. It is also the case of many societies, (professional or other) engaged in constantly providing up-to-date information on a given topic to their members in the form of a web portal. In this case the information is kept as a web resource by a team of editors, who select and edit the published documents. In some cases, the editor and the web end-user, the person who consults the resource, are the same person. Persistent information needs on a specific topic may be answered by tools that keep an eye on the web, automatically searching for documents on that topic and that are able to present them to the end-users as expected by them. Editors have the role of expressing the end-user's needs and preferences.

In this paper we propose *webTOPIC*, a methodology that assists editors in the process of compiling resources on the Web, with the following characteristics:

- allow for the broad specification of any topic, including its ontological structure, by a team of editors;
- enable the effective exploration of large document collections by the end-user;
- maintain quality, as perceived by end-users, at acceptable levels without requiring explicit effort by the end-user or the editor;
- detect and adapt to drift in end-user needs and to changes in information sources.

From the editor's point of view, webTOPIC is a tool for specifying, collecting and organizing document collections that satisfy some specific and persistent information needs of the (end-)users. Once the editor has specified an information need, webTOPIC compiles resources following these specifications. A *resource* is a document collection satisfying a specific information need. We will refer to each specific user information need as a *topic*. Each resource is an instance of a topic.

The user interacts with the methodology during two distinct phases: in the first phase the user defines the topic and specifies its characteristics (editor's role), in the second phase the user explores the resources that are being compiled by the system (end-user's role). The first phase is concentrated on a short period of time. The specification of a particular information need includes, among others, a taxonomy, which describes the ontological structure the user is interested in, and a set of exemplary documents. In the second phase, which occurs while the user maintains interest on the topic, the system follows the evolution on end-user preferences, automatically and incrementally building and keeping resources aligned with end-user current interests.

In the rest of the paper we start, in section 2, by describing and comparing our approach to previous related work. Then, in section 3, we describe the webTOPIC methodology. We refer to its architecture and then describe its most relevant aspects, including the resource acquisition phase, document pre-processing, learning, resource presentation and exploration and resource quality. Section 4 describes the experiments we have conducted to evaluate our semi-supervised document classification method. In section 5 we present our conclusions and directions for future work.

2 Resource Compilers

An automatic resource compiler is a system that, given a topic, seeks and retrieves a list of the most authoritative web documents, as perceived by the system, for that topic [4]. This is a very broad definition, under which many distinct types of systems