PersonalSearcher:
An Intelligent Agent for Searching Web Pages

Daniela Godoy and Analía Amandi
ISISTAN Research Institute, Facultad de Ciencias Exactas
Universidad Nacional del Centro de la Prov. de Buenos Aires
Campus Universitario - Paraje Arroyo Seco – (7000) – Tandil, Bs. As., Argentina
1Also Facultad de Ingeniería, Universidad Nacional de La Pampa
Calle 9 esq. 110 – (6360) – General Pico, La Pampa, Argentina
{dgodoy,amandi}@exa.unicen.edu.ar

Abstract. The volume of information on the Internet is constantly growing. This fact causes that the search of interesting information becomes a time-consuming task. Generally, a user must revise a big number of uninteresting documents and consult several search engines before finding relevant information. A personalized agent, called PersonalSearcher, that assists the user in finding interesting documents in the World Wide Web is presented in this paper. This agent carries out a parallel search in the most popular Web search engines and filters their result, listing to the user a reduced number of documents with high probability of being relevant to him. This filtering is based on a user profile that the agent builds by observing the user behavior on the Web. The agent uses a textual case-based reasoning approach in order to detect specific subjects that the user is interested in and organizes them in a hierarchy that defines the user profile.

1 Introduction

The information available through Internet is constantly growing. This fact causes that the search of interesting information becomes a time-consuming task since this activity involves the analysis and separation of interesting pages from a great set of candidate pages.

Search engines are the most widely spread tools for searching web pages. Users provide a set of words to these engines and wait for a set of pages related to those words. This mechanism based on words that act as keywords in the searching process is easy to use. However, this simplicity for expressing search goals generally produces low levels of precision in the response from these engines.

In this context, users have to dedicate a considerable amount of both time and effort to browse a ranking list of documents. Generally, this list contains a lot of uninteresting documents and just a few really relevant ones.

For example, we can imagine a user looking for web pages about software agents. This user makes a query using the keyword agents chosen from many other words that refer to this subject (i.e. softbots). Traditional tools return to the user documents about software agents, travel agents, insure agents, etc. all at the same time. A personalized system able to contextualize the user consult according to his
preferences and subjects of interest could be preferred over traditional search engines. It could, for example, filter out documents about travel agent and insure agents for our user.

In this sense, personal agents have been developed to help the management of the increasing volume of information. They are intelligent assistants that make different tasks on behalf of the user to find, filter and access to a great amount of information from different sources, and finally present a reduced and potentially relevant part of this information to their users. These personalized agents use different learning mechanisms to capture users’ interests and habits over time.

We present in this article an intelligent agent that learns about users’ interests by observing users’ behavior while they are carrying out regular activities on the Web. By a content-based analysis of the information extracted by observation, this agent is able to deduce the subjects that a user is interested in, and according to them filters the resulting list of web pages of a traditional search.

Our agent, named PersonalSearcher, builds user profiles using a technique for dynamic classification based on textual case-based reasoning. In this article, we present this agent and, particularly, our technique for dynamic classification that allows the agent to filter pages according to personal interests.

The article is organized as follows. Section 2 introduces the functionality of our PersonalSearcher agent. Section 3 treats the construction of a user profile. Section 4 shows some evaluations for PersonalSearcher. Section 5 compares our agent and technique with related works. Finally, conclusions are presented.

2 Agent Overview

Each agent, instance of PersonalSearcher, monitors Web activity of his associated user. This monitoring is made in order to collect documents, which are interesting to the user. For each document read by a user on the standard browser, the agent observes a set of given characteristics in order to determine its relevance degree for that user. These observed characteristics are basically the time consumed on the reading, its length, and so on.

The documents classified as interesting are analyzed to obtain other characteristics, which describe the subject treated on them. For achieving this goal, a textual case-based reasoning approach is used. In this approach, the main characteristics of textual documents are extracted to represent them as cases. Our case-based reasoner deals with these cases in order to learn interesting subjects for the user. At the same time, it organizes them building a subject hierarchy, which determines the user profile for such user.

Users interact with their PersonalSearcher expressing their information needs by keywords as usual. The agent in turn posts these keywords to the most popular search engines (Altavista, Infoseek, Excite, etc.), obtaining a set of documents covering a widely portion of the Web.

The relevance degree of each document in relation to the user profile is computed by the PersonalSearcher to determine the convenience of suggesting the document. Only documents that surpass a given threshold of relevance as regards the user profile are sent back to the user as a result to his query. Figure 1 illustrates this process.