Hopfilter: An Agent for Filtering Web Pages Based on the Hopfield Artificial Neural Network Model

Juan Manuel Adán-Coello[1], Carlos Miguel Tobar[1], Ricardo Luís de Freitas[1], and Armando Marin[2]

[1] PUC-Campinas, Rod. Cx. P. 317, CEP 13012-970, Campinas, SP, Brazil
{juan, tobar, rfreitas}@puc-campinas.edu.br
[2] Senac Ribeirão-Preto, Ribeirão Preto, SP, Brazil
amarin@sp.senac.br

Abstract. With the expansion of the Internet, the amount of information available is continuously reaching higher growth rates. This fact leads to the necessity of developing new advanced tools to collect and filter information that meets users' preferences. This paper presents an agent that uses automatic indexing, concept space generation, and a Hopfield artificial neural network to filter web pages according to users' interests. The experiments that were conducted to evaluate the agent show that it has very satisfactory precision and coverage rates.

1 Introduction

With the expansion of the Web, users face increasing difficulties to fulfill their information needs. The unstructured nature of the data stored in the Web and its dynamic nature contribute to this scenario, which requires users to check frequently for new documents of interest. This situation has motivated the development of personal software agents for continuous information gathering and filtering.

Information needs change from user to user. Therefore, information filtering systems have to be personalized, playing the real role of personal assistants. Such personalized information filtering system has to satisfy three requirements:

1. Specialization: the system selects only documents relevant to the user;
2. Adaptation: information filtering is an iterative process that is performed for long periods of time, during which user's interests change;
3. Exploration: the system should be able to explore new domains, in order to find new information potentially interesting to the user.

A number of different models have been implemented for information retrieval and filtering. Typically, these implementations consist of three main components: document representation, user's interest representation, and algorithms used to match user's interests to documents representations [1] [2].

This paper presents and evaluates the architecture of Hopfilter, a personal agent that mines Web information sources and retrieves documents according to user's

R. Cooper and J. Kennedy (Eds.): BNCOD 2007, LNCS 4587, pp. 164–167, 2007.

interests. It is organized as follows. In section 2 the structure of the Hopfiter agent is presented. Section 3 presents the results of some experiments conducted to evaluate the agent. Section 4 closes the paper with some final remarks.

2 The Architecture of Hopfilter

The filtering agent is composed by User interface (UI), Web interface (WI), Document Preprocessing (DPP), Automatic indexing (AI), Generation of the Space of Concepts (GSC), Artificial Neural Net (ANN). The UI and WI modules interface with the user and with the Web, as the names suggest, and are not the focus of this paper.

The filtering agent can operate in two modes: "concept space generation" and "document filtering". During the concept space generation mode, the DPP, AI, and GSC modules are used. The document filtering mode involves the DPP, AI, and ANN modules. During the document filtering mode, a concept space (CS) for the considered domain must be available. Each mode of operation is briefly described below together with each module.

2.1 Automatic Indexing

When a document is indexed, the result is a list of terms or indexes that represents the document content. AI consists of three operations: stopword removal, work stemming, and term formation.

Removing stopwords. After identifying the words in the input document, using the DPP module, the words that are not relevant for characterizing the document content are removed. To assist this process it is used a dictionary with some 46,000 entries, which can be manually marked by the user as stopwords. All input document words not found in the dictionary are kept in a table for posterior analysis. They can be included in the dictionary if desired.

Word stemming. This step purpose is to reduce the number of cognate words to be indexed. The implemented algorithm is adapted from the Lancaster Stemming Algorithm [3] for the Portuguese language, which only removes words suffixes.

Term formation. A term can be formed by one, two or three adjacent words. For each term, it is computed a term frequency, *tf*, that represents the number of times the term appears in the document. When the agent operates in the concept space generation mode, it is also calculated the document frequency for the term, *df*, that represents the number of documents, in the collection, where the term appears.

2.2 Generation of a Concept Space

The objective of the GSC module is to calculate asymmetric coefficients of similarity for each pair of terms, generating a matrix containing the terms and the respective coefficients, or relationship degrees. This matrix, also called similarity matrix, represents the concept space.