Text Content Approaches in Web Content Mining

Víctor Fresno Fernández
Universidad Rey Juan Carlos, Spain

Luis Magdalena Layos
Universidad Politécnica de Madrid, Spain

INTRODUCTION

Since the creation of the Web until now, the Internet has become the greatest source of information available in the world. The Web is defined as a global information system that connects several sources of information by hyperlinks, providing a simple media to publish electronic information and being available to all the connected people.

In this context, data mining researchers have a fertile area to develop different systems, using Internet as a knowledge base or personalizing Web information. The combination of the Internet and data mining typically has been referred as Web mining, defined by Kosala and Blockeel (2000) as "a converging research area from several research communities, such as DataBase (DB), Information Retrieval (IR) and Artificial Intelligent (AI), especially from machine learning and Natural Language Processing (NLP)"

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services; traditionally focused in three distinct ways, based on which part of the Web to mine: Web content, Web structure and Web usage. Brief descriptions of these categories are summarized below.

- **Web Content Mining:** Web content consists of several types of data, such as textual, image, audio, video, and metadata, as well as hyperlinks. Web content mining describes the process of information discovery from millions of sources across the World Wide Web. From an IR point of view, Web sites consist of collections of hypertext documents for unstructured documents (Turney, 2002); from a DB point of view, Web sites consist of collections of semi-structured documents (Jeh & Widom, 2004).
- **Web Structure Mining:** This approach is interested in the structure of the hyperlinks within the Web itself—the interdocument structure. The Web structure is inspired by the study of social network and citation analysis (Chakrabarti, 2002). Some algorithms have been proposed to model the Web topology, such as PageRank (Brin & Page, 1998) from Google and other approaches that add content information to the link structure (Getoor, 2003).
- **Web Usage Mining:** Web usage mining focuses on techniques that could predict user behavior while the user interacts with the Web. A first approach maps the usage data of the Web server into relational tables for a later analysis. A second approach uses the log data directly by using special preprocessing techniques (Borges & Levene, 2004).

BACKGROUND

On the Web, there are no standards or style rules; the contents are created by a set of very heterogeneous people in an autonomous way. In this sense, the Web can be seen as a huge amount of online unstructured information. Due to this inherent chaos, the necessity of developing systems that aid us in the processes of searching and efficient accessing of information has emerged.

When we want to find information on the Web, we usually access it by search services, such as Google (http://www.google.com) or AllTheWeb (http://www.alltheweb.com), which return a ranked list of Web pages in response to our request. A recent study (Gonzalo, 2004) showed that this method of finding information works well when we want to retrieve home pages, Websites related to corporations, institutions, or specific events, or to find quality portals. However, when we want to explore several pages, relating information from several sources, this way has some deficiencies: the ranked lists are not conceptually ordered, and information in different sources is not related. The Google model has the following features: crawling the Web, the application of a simple Boolean search, the PageRank algorithm, and an efficient implementation. This model directs us to a Web page, and then we are abandoned with the local server search tools, once the page is reached.

Nowadays, these tools are very simple, and the search results are poor.

Other ways to find information is using Web directories organized by categories, such as Yahoo (http://www.yahoo.com) or Open Directory Project (http://www. dmoz.org). However, the manual nature of this categorization makes the directories' maintenance too arduous, if machine processes do not assist it.

Future and present research tends to the visualization and organization of results, the information extraction over the retrieved pages, or the development of efficient local servers search tools. Next, we summarize some of the technologies that can be explored in Web content mining and give a brief description of their main features.

Web Mining and Information Retrieval

These systems retrieve contents with as much text as multimedia; the main feature is that access to information is accomplished in response to a user's request (Fan et al., 2004; Wang et al., 2003). Techniques inherited from NLP are added to these systems.

Text Categorization on the Web

The main goal of these methods is to find the nearest category, from a pre-classified categories hierarchy to a specific Web-page content. Some relevant works in this approach can be found in Chakrabarti (2003) and Kwon and Lee (2003).

Web Document Clustering

Clustering involves dividing a set of n documents into a specific number of clusters k, so that some documents are similar to other documents in the same cluster and different from those in other clusters. Some examples in this context are Carey, et al. (2003) and Liu, et al. (2002).

MAIN THRUST

In general, Web mining systems can be decomposed into different stages that can be grouped in four main phases: *resource access*, the task of capturing intended Web documents; *information preprocessing*, the automatic selection of specific information from the captured resources; *generalization*, where machine learning or data-mining processes discover general patterns in individual Web pages or across multiple sites; and finally, the *analysis phase*, or validation and interpretation of the mined patterns. We think that by improving each of the phases, the final system behavior also can be improved.

In this work, we focus our efforts on Web pages representation, which can be associated with the information-preprocessing phase in a general Web-mining system. Several hypertext representations have been introduced in literature in different Web mining categories, and they will depend on the later use and application that will be given. Here, we restrict our analysis to Web-content mining, and, in addition, hyperlinks and multimedia data are not considered. The main reason to select only the tagged text is to look for the existence of special features emerging from the HTML tags with the aim to develop Web-content mining systems with greater scope and better performance as local server search tools. In this case, the representation of Web pages is similar to the representation of any text.

A model of text must build a machine representation of the world knowledge and, therefore, must involve a natural language grammar. Since we restrict our scope to statistical analyses for Web-page classification, we need to find suitable representations for hypertext that will suffice for our learning applications.

We carry out a comparison between different representations using the vector space model (Salton et al., 1975), where documents are tokenized using simple rules, such as whitespace delimiters in English and tokens stemmed to canonical form (e.g., reading to read). Each canonical token represents an axis in the Euclidean space. This representation ignores the sequence in which words occur and is based on the statistical about single independent words. This independence principle between the words that coappear in a text or appear as multiword terms is a certain error but reduces the complexity of our problem without loss of efficiency. The different representations are obtained using different functions to assign the value of each component in the vector representation. We used a subset of the BankSearch Dataset as the Web document collection (Sinka & Corne, 2002).

First, we obtained five representations using well-known functions in the IR environment. All these are based only on the term frequency in the Web page that we want to represent, and on the term frequency in the pages of the collection. Below, we summarize the different evaluated representations and a brief explanation.

1. **Binary:** This is the most straightforward model, which is called *set of words*. The relevance or weight of a feature is a binary value *{0,1},* depending on whether the feature appears in the document or not.
2. **Term Frequency (TF):** Each term is assumed to have an importance proportional to the number of times it occurs in the text (Luhn, 1957). The