Chapter 6
Method to Reduce Complexity and Response Time in a Web Search

María R. Romagnano
Universidad Nacional de San Juan, Argentina

Silvana V. Aciar
Universidad Nacional de San Juan, Argentina

Martín G. Marchetta
Universidad Nacional de Cuyo, Argentina

ABSTRACT

Living in times of technological changes that alter our daily activities, involving tasks such as reading the newspaper, following the weather, scheduling a trip, are usually executed after perusal of the gigantic repository of information, commonly known as the World Wide Web. However some problems are still associated with the information found in such a vast amount of information: heterogeneity, availability, distribution, quality and quantity of irrelevant information. Recent work has suggested different ways of grouping similar information sources, trying to give solutions to these problems. However, some domains are more complex than others. For example, a person looking for tourist information, is generally overwhelmed by visiting various websites. This paper proposes the implementation of a method to retrieve and group web information sources, depending on the services they offer; thereby allowing the user to get accurate answers; thus reducing the time and complexity in the search.

INTRODUCTION

The World Wide Web has become a major source of information for virtually every area of interest. So daily tasks such as reading the newspaper online, following the weather, planning a trip, or just communicating with other people, is done faster, more efficiently, even economically and in real time; through web information sources. However, these web sources can present some problems: *Heterogeneity*; because

of the diversity of information found, *Lack of structure*; in terms of which sources do not have a single format. *Availability*; because at one point in time, some may be available but not in another. *Distribution*; in the sense that they are scattered around the web. *Quantity*; in terms of the huge number of sources that can be accessed; and *Quality*, refers to which information on the web is not always reliable (Cabrera, 2009, pp. 4-5; Mendez, Chavarros & Moreno, 2007, p. 397).

Some domains are more complex than others, where these problems are increased by the accumulation and variety of information handled and because the number of sources consulted is considerable. When a user searches for information on the web, especially one who may not be expert in performing the searches, this individual may obtain a large number of results whereby the response time is greater. This may be due to users who infrequently, or perhaps never make use of the logical operators and/or boolean operators. Moreover, some users tend not to use the advanced search, which allows filtering the results by date, location, or author.

Therefore, in trying to get answers according to individual needs, the user could spend days reading each of the thousands of results or just randomly choose one of the first which perhaps is irrelevant to their interests or is not completely convincing. After this experience, the web user is discouraged from searching for information again in this great digital library. The ideal situation would occur when the user locates the products, services and/or the precise information with minimal effort on his part.

The drawbacks cited motivated the proposal of one method that retrieves information on the web, that groups information sources according to the services it offers and that provides accurate information to the user, according to his needs; thereby reducing time and complexity in the search.

The method is validated by means of an application case in the tourism area.

The second section of this paper reviews the bibliography consulted. The third section mentions related work. Section four presents the new method of information retrieval FiPaWeb (Filtering of web pages). Section five describes an application field related to tourism. Finally, section six presents the conclusions and future research.

BIBLIOGRAPHIC REVIEW

The IR (Information Retrieval) helps users to find relevant and necessary information from a large collection of text documents. Technically, the IR studies the acquisition, organization, storage, retrieval and distribution of information (Liu 2007, pp.183-235).

The work of Han and Kamber (2006, p. 615) mentions that information retrieval is the area dedicated to the organization and retrieval of information from a large amount of text-based documents. Although these have been developed in parallel with database systems; they differ from the latter in some presented problems, such as de-structuring of documents, searchable by approximation based on keywords and the notion of relevance.

To carry out a subsequent analysis of the retrieved information, web pages can be grouped according to a specific criterion. This grouping can be performed using supervised learning techniques such as classification; or unsupervised learning techniques such as clustering (Baldi, Frasconi, & Smyth 2003, p. 17).

The classification is much more than a simple assignment of category labels and an organization of information. Classification is traditionally seen as a supervised learning problem in which a set of labeled data is used to train a classifier; which will then be used to categorize or label future examples (Qi, 2012, p. 3).

Related Content

Named Entity Based Ranking with Term Proximity for XML Retrieval

Abubakar Roko, Shyamala Doraisamy, Azreen Azmanand Azrul Hazri Jantan (2018). *International Journal of Information Retrieval Research (pp. 57-77).*

www.irma-international.org/article/named-entity-based-ranking-with-term-proximity-for-xml-retrieval/198965

Text Processing by Binary Neural Networks

T. Beranand T. Macek (2004). *Intelligent Agents for Data Mining and Information Retrieval (pp. 125-136).*

www.irma-international.org/chapter/text-processing-binary-neural-networks/24159

Semantic Technologies and Big Data Analytics for Cyber Defence

Louise Leenenand Thomas Meyer (2018). *Information Retrieval and Management: Concepts, Methodologies, Tools, and Applications (pp. 1375-1388).*

www.irma-international.org/chapter/semantic-technologies-and-big-data-analytics-for-cyber-defence/198604

A Framework for Automated Scraping of Structured Data Records From the Deep Web Using Semantic Labeling: Semantic Scraper

Umamageswari Kumaresanand Kalpana Ramanujam (2022). *International Journal of Information Retrieval Research (pp. 1-18).*

www.irma-international.org/article/a-framework-for-automated-scraping-of-structured-data-records-from-the-deep-web-using-semantic-labeling/290830

Analysis the Typhoon Eyes of Megi from MTSAT Satellite Cloud Images with 3-D Profile Reconstruction

Yueh-Jyun Lee, Ji-Chyun Liu, Shih-Wen Liu, Yuh-Fong Linand Lung-Chun Chang (2013). *Modern Library Technologies for Data Storage, Retrieval, and Use (pp. 143-151).*

www.irma-international.org/chapter/analysis-typhoon-eyes-megi-mtsat/73771