Chapter 1
The Evolution of the (Hidden) Web and its Hidden Data

Manuel Álvarez Díaz
University of A Coruña, Spain

Víctor Manuel Prieto Álvarez
University of A Coruña, Spain

Fidel Cacheda Seijo
University of A Coruña, Spain

ABSTRACT

This paper presents an analysis of the most important features of the Web and its evolution and implications on the tools that traverse it to index its content to be searched later. It is important to remark that some of these features of the Web make a quite large subset to remain "hidden". The analysis of the Web focuses on a snapshot of the Global Web for six different years: 2009 to 2014. The results for each year are analyzed independently and together to facilitate the analysis of both the features at any given time and the changes between the different analyzed years. The objective of the analysis are twofold: to characterize the Web and more importantly, its evolution along the time.

INTRODUCTION

Since its origins, the WWW has been the subject of numerous studies. However, one constant has been and continues to be the analysis of its size. Although it is nearly impossible to compute the exact size of the Web, because it is in constant change, everyone agrees that his size is in the order of billions of documents or pages (Gulli & Signorini, 2005). In this way, the WWW could be considered the largest repository of documents ever built.

Due to the large size of the Web, search engines are essential tools for users who want to access relevant information for a specific topic. Search engines are complex systems that allow, among other things: gathering, storing, managing and granting access to the information. Crawling systems are those which perform the task of gathering information. These programs are capable of traversing and analysing the Web in a certain order, by following the links between different pages.

DOI: 10.4018/978-1-4666-8696-0.ch001

The task of a crawling system presents numerous challenges due to the quantity, variability and quality of the information that it needs to collect. Among these challenges, specific aspects can be highlighted, such as the technologies used in web pages to access to data, both in the server-side (Raghavan & Garcia-Molina, 2001) or in the client-side (Bergman, 2001); or problems associated with web content such as Web Spam (Gyongyi & Garcia-Molina, 2005) or repeated contents (Kumar & Govindarajulu, 2009), etc. To get a detailed enumeration it is necessary to analyse the Web in more detail.

This article presents an analysis of the most important features of the Web and its components and also its evolution over a period of time. Particular emphasis is placed on the use of client/server side technologies. It is very important to remark that the Hidden Web is "hidden" just for the existence of some technologies used in web documents that difficult the task of crawler systems for accessing to it.

The analysis focuses on a snapshot of the Global Web for six different years: from 2009 to 2014. The results for each year are analysed independently and together to simplify the evaluation of the features at any given time and the changes between the different analysed years. The objectives of the analysis are twofold: to characterize the Web and more importantly, its evolution along the time, and also to analyze how its changes affect tools such as crawlers and search engines. So, changing trends are presented and explained.

The structure of this paper is as follows. Background section introduces works related with the study and characterization of the Web. Methodology section shows the methodology followed to characterize the Web. Dataset section explains the dataset used. The analysis section discusses the results obtained for each year, and their evolution through the time. Finally, the future research directions section includes possible future works and the conclusions section summarises the results of the paper.

BACKGROUND

The characterization of the Web is a topic widely studied in the supported literature. Baeza-Yates et al. (Baeza-Yates, Castillo & Efthimiadis, 2007) performed a study which analyses various features of the Web at several levels: web page, web site and national domains. On the other hand, there are several studies that are focused on the Web of a particular country. In 2000, Sanguanpong et al. (Sanguanpong, Piamsa-nga, Keretho, Poovarawan & Warangrit, 2000) presented an analysis of various issues related to web servers and web documents in Thailand. Baeza-Yates et al. presented two articles (Baeza-Yates, Castillo, & Lopez, 2005; Baeza-Yates & Castillo, 2000), which were focused more specifically on the characteristics of the Spanish and Chilean Web, respectively. The Spanish Web was also studied by Prieto et al. (Prieto, Álvarez, & Cacheda, 2013), by comparing the analysis of the Spanish Web with the Global Web in a tree-years period. In 2002, Boldi et al. (Boldi, Codenotti, Santini, & Vigna, 2002) presented an interesting article, where the authors have studied different features (content and structure analysis, web graph, etc.) of the African Web. Gomes et al. (Gomes & Silva, 2005), carried out a study to characterise the community Web of the people of Portugal. The authors studied different features such as: the number and domain distribution of sites, the number and size distribution of text documents, the structure of this Web, etc. Years later, Miranda and Gomes (Miranda & Gomes, 2009) performed a study which presented trends on the evolution of the Portuguese Web, derived from the comparison of two characterizations of a web portion performed within a 5 year interval. This study analyses several metrics regarding content and site characteristics. Modesto et al. (Modesto, Pereira, Ziviani, Castillo, & Baeza-Yates, 2005) presented an article, which analyses the features of approximately 2% of the .br