FiPaWeb: method to reduce the complexity of web search

María R. Romagnano[1], Silvana V. Aciar[1], Martín G. Marchetta[2]

[1] Instituto de Informática, Fac. de Cs. Exactas, Físicas y Naturles, UNSJ, San Juan, Argentina
{maritaroma,saciar}@iinfo.unsj.edu.ar
[2] FI - UNCu, Centro Universitario S/N, Mendoza, Argentina
mmarchetta@fing.uncu.edu.ar

The rapid evolution of technology and the widespread use of the Internet provide information from a lot of sources of information. These sources have some problems such as heterogeneity of information; lack of structure, the sources does not have a single format; they are not always available; they are distributed over the web and information is not always reliable [1]. There are complex domains, where these problems are exacerbated by the accumulation and variety of information handling and users may become overwhelmed trying to do a simple query. Trying to get answers according to their needs, the user could spend days reading each of the thousands of results provided by the search engine or just randomly choose one of the earliest and perhaps not entirely convincing. The user should find the desired information without much effort.

The above problems led to a review of the literature on the subject, formalize the problem and propose a method to seek information at the web, group sources according to the services offered and provide the user with accurate and unified information, reducing time and complexity in the search.

There are several related works, as for example in [2] the anchor text is used for recovery. In [3] the k-means method and tag information were used to group pages. In [4] a system that gathers web pages using a fuzzy clustering algorithm was proposed.

In this work the FiPaWeb method is proposed to group the pages; and at the same time keep the overlap between them; without losing information.

The method has five stages. In the first stage the web search is performed. In the second stage the relevant pages are selected. In the third stage the value of each service is determined for each page. In the fourth stage the pages are grouped based on established criteria. Finally, in the fifth stage, the system accesses the groups that contain only the information requested by the user in a query.

This method can be applied to a number of important domains such as education, tourism, health, etc. In this instance, a case study was conducted in the domain of tourism. The data used in the tests were collected from tourism sites in the province of San Juan.

The measure of time complexity was used to compare the performance of FiPaWeb with other algorithms. The results obtained were: K-Means: 12,800, Fuzzy C-Means: 102,400 and FiPaWeb: 10,240.

Our main contribution is an algorithm that quickly and easily groups similar pages to provide users with a unified view of the information requested. With this method a retail number of responses was achieved in less time.

The definition of a threshold to determine the amount of page to be displayed when user makes a query is proposed as future work. It is also expected to apply the method to other domains.

1. Baldi, P., Frasconi, P. and Smyth, P. “Modeling the Internet and the Web: Probabilistic Methods and Algorithms”.Publicado por John Wiley & Sons, Ltd. Año 2003. ISBN: 0-470-84906-1. Pags.21-22.
2. Eiron, N. and McCurley, K. “Analysis of anchor text for web search” Proceedings of SIGIR’03. Pags. 459-460. ISBN 1-58113-646-3. http://dl.acm.org/citation.cfm?id=860550
3. Ramage, D., Heymann, P., manning, C. D. and Garcia Molina, H. “Clustering the tagged web”. Proceeding of WSDM’09. Pags. 54-63
4. Shelke, M, Sadavarte, K., Dhurjad R. and Pandit, N. “Improved web page clustering using words and tags”. 1° International Conference on Recent Trends in Enginneering & Technology, Mar-2012. http://www.ijecscse.org/papers/SpecialIssue/comp1/49.pdf