Chapter 12
Retrieval of Relevant Web Pages by a New Filtering Method

Sahar Maâlej Dammak
MIRACL Laboratory, Tunisia

Anis Jedidi
MIRACL Laboratory, Tunisia

Rafik Bouaziz
MIRACL Laboratory, Tunisia

ABSTRACT

With the great mass of the pages managed through the world, and especially with the advent of the web, it has become more difficult to find the relevant pages after an interrogation. Furthermore, the manual filtering of the indexed web pages is a laborious task. A new filtering method of the annotated web pages (by a semantic annotation process) and the non-annotated web pages (retrieved from search engine Google) is then necessary to group the relevant web pages for the user. In this chapter, the authors first synthesize their previous work of the semantic annotation of web pages. Then, they define a new filtering method based on three activities. They also present their querying and filtering component of web pages; their purpose is to demonstrate the feasibility of our filtering method. Finally, the authors present an evaluation of this component, which has proved its performance for multiple domains, and they discuss the use of the extended Boolean retrieval method in the new filtering method.

INTRODUCTION

The search of the relevant pages on the Web becomes a very difficult task facing the considerable increase in the number of available Web pages, the diversity of their structures and their contents and the presence of a significant amount of useless information. It is important then to improve this difficult task in order to obtain relevant answers to users.

DOI: 10.4018/978-1-5225-7347-0.ch012

Indeed, the retrieved Web pages after the search on the Web do not often satisfy the user when making his/her interrogation. The user has to sift a long time to locate pages of interest. So, our objective in this chapter is to see how we can improve the retrieval of the relevant pages in the interrogation results, taking into account the annotations made by our semantic annotation process. In fact, with the exponential growth of the internet, it becomes more and more difficult to find the relevant pages. Therefore, a filtering process automation of the annotated and non-annotated Web pages is required.

This chapter provides then two main contributions. First, it synthesizes our previous work. In fact, we have proposed a semantic annotation approach of the Web pages (Maâlej Dammak *et al.,* 2013a, 2014b). This approach describes the Web pages by particular metadata, called "Fuzzy semantic annotations", in the semantic Web environment. The annotations are stored in RDF documents. Our annotation method (Maâlej Dammak *et al.,* 2013b, 2014b) is an enhancement of the first result of the annotation done by the Semantic Radar Plug-in (Bojars *et al.,* 2008) on the Web pages, using an enriched domain ontology, the FOAF (friend-of-a-friend) ontology (Brickley & Miller, 2015) and the SIOC (semantically-interlinked online communities) ontology (Bojars & Breslin, 2018). The additional annotation concerns the concepts of the result of the Semantic Radar. Secondly, this chapter defines a new method for filtering the indexed Web pages with the annotated ones. The new filtering method puts the retrieval Web pages in order, after the interrogation, before their display to the user. In addition, it groups and places the most relevant Web pages at the top of the extracted list. Also, the present chapter is interested in the operationalization of this method in order to automate the clustering of the relevant Web pages. So, we propose a new querying and filtering component for improving the user's search result. On the one hand, this component allows the extraction of the concepts of the domain ontology corresponding to the search area of the user. These concepts help the latter when writing the semantic query for the Web interrogation. On the other hand, the component ensures the automation of the scores calculation for the retrieved Web pages, annotated and/or non-annotated, and the hierarchical classification of these pages in order to regroup them according to their hierarchies. We set two groups in the end of the clustering; a group that contains the most relevant pages and a group that includes the less relevant pages. However, we propose presenting to the user only the group that contains the most relevant pages.

The remainder of this chapter is organized as follows. Section 2 shows the positioning of our work in relation to the literature. Section 3 presents a synthesis of our semantic annotation process of the Web pages and the RDF of the semantic and fuzzy annotation. Section 4 defines our new filtering method for the annotated and non-annotated Web pages. Section 5 describes the principal development parts of our querying and filtering component. Section 6 shows the relevance of our component, through an evaluation of these results. Section 7 discusses the use of the extended Boolean retrieval method in our new filtering method. Finally, Section 8 summarizes the contribution of the chapter and outlines future work.

RELATED WORKS

The search on the Web cannot guarantee the retrieval of the relevant Web pages to the user; the pages retrieved may not satisfy his/her request. The purpose of our filtering process is to improve the search result by the selection of the relevant pages. So, a filtering method of the indexed Web pages (including the annotated ones) that groups the relevant pages is required.