DOWSER : Discovery of Web Sources by Evaluating Relevance

Romain Noël*,**, Alexandre Pauchet*, Bruno Grilheres**,
Nicolas Malandain*, Stéphan Brunessaux**, Laurent Vercouter*

*LITIS MIU
INSA ROUEN
St Etienne du Rouvray, FRANCE
firstname.lastname@insa-rouen.fr,
**CASSIDIAN
Val-de-Reuil, FRANCE
firstname.lastname@cassidian.com

Abstract. The constant growth of the Web in recent years has made more difficult the discovery of new sources of interest on a given topic. In particular for intelligence analysts which are confronting the search of hard-to-find pages on specific topics with traditional Information Retrieval tools. In this paper, we describe a new Web source discovery system called DOWSER (Discovery Of Web Sources Evaluating Relevance). The goal of this system is to provide users with new relevant sources of information according to their needs without using search engines. We study the interest of exploiting a user profile to lead a focused crawling process in order to avoid collecting and indexing all accessible Web documents. The user's information needs are not specified using keywords, but using user's Web pages of interests represented by DBPedia resources (Bizer et al., 2009). A series of experiments were conducted on the Web and they provided an empirical evaluation. Results of these user experiments are presented in this paper.

1 Introduction

The explosive growth of the world-wide-web has resulted in a huge amount of information available on Internet. In the domain of intelligence analysis, experts find it difficult to discover new sources of information using traditional Information Retrieval tools. A source of information is considered as a Web resource: a website, a section of a website, or a Web page. Due to the heterogeneity of these sources, automatic discovery of information becomes a complex task. Moreover, relevant sources searched by intelligence analysts can be hard-to-find websites such as websites which are not indexed by any search engine. A relevant source can be defined as a Web site or a section of a Web site providing on-topic Web page(s) matching the user needs. Crawlers and search engines have to deal with the size of the Web and with the deep Web, and they also have to consider the user needs in order to find relevant sources.