Chapter 13
Harvesting Deep Web Data through Produser Involvement

Tomasz Kaczmarek
Poznań University of Economics, Poland

Dawid Grzegorz Węckowski
Poznań University of Economics, Poland

ABSTRACT

Acquiring the data from the deep Web is a complex process, which requires understanding of Website navigation issues, data extraction, and integration techniques. Currently existing solutions to automate it are not ready to cover the whole deep Web and require skills and knowledge to be applied in practice. However, several systems were created, which approach the problem by involving end users who are able to bring the data from the deep Web to the surface while creating solutions for their own information needs. The authors study these systems in the chapter from the end user perspective, investigating their interfaces, languages that they expose to end users, and the platforms that accompany the systems to involve end users and allow them to share the results of their work.

INTRODUCTION

Produsers category is a group of highly involved, creative users of the Web applications, which are also content and functionality creators. The concept was derived from the 'prosumers' term, which describes exceptionally well-informed and critical consumers, that contribute to product development. The notion of 'produsage' was created to distinguish a group of Web users, that engage in "...collaborative and continuous building and extending of existing content in pursuit for further improvement" (Bruns, 2006). There is a number of characteristic features of these users and their activity, discussed in further publications on produsers (Bruns, 2007; Ritzer & Jurgenson, 2010), such as evolution of content that they create, collaborative effort, or their approach to intellectual property. One area where creativity of produsers is employed is deep Web harvesting. Deep Web consists of the databases, that are accessible via Web interface to humans, but poorly indexed by regular search engines and, in consequence, not available through regular Web search. It contains valuable information that is not available on its reverse – surface Web. In order to acquire this information (prices and stock amounts for prod-

DOI: 10.4018/978-1-4666-4313-0.ch013

ucts, statistical data, bibliographical information and many other types of information) significant knowledge and effort is necessary, that exceeds beyond querying established Web search engines. First, the sources need to be identified or found, then the user has to understand and be able to use the query interface for a given source, and only afterwards, he is able to obtain the Web pages containing actual data, which often require further processing to be useful. Therefore getting the data from the deep Web is a complex process, which requires understanding of Website navigation issues, data extraction, and integration techniques. Due to lack of fully automated tools in the style of search engines, it has to be carried out manually to a large extent. Researchers in the area are constantly looking for solutions to decrease the complexity and provide convenient interfaces to solve the problem. Produsers, with their drive to make their creations accessible and reusable, and higher awareness of the technical issues that need to be solved, engage in surfacing the data and make it more available for a wider audience. Several systems were created, that approach the problem by involving produsers and their higher-than-standard abilities, together with their need for the data which are otherwise inaccessible. We study these systems together with their interfaces and underlying formalisms (to assess the level of knowledge and expertise required to use them) as well as motivation model for the users to take part in such endeavours.

DEEP WEB STRUCTURE

The Deep Web notion (a.k.a. the Hidden Web) refers to Web pages that are not directly accessible by the usage of URLs, but are rather dynamically generated upon HTML form submitting (Madhavan et al., 2008). Web page retrieval from the Deep Web involves filling the form with desired values, which will influence the content of delivered Web page. Apart from the result processing, the challenging part is the automatic determination of HTML form values, that can generate useful outcome.

As it was shown in the studies (He et al., 2007), the Deep Web is very extensive and versatile. In 2007 it was estimated to embrace over 300,000 sites, 450,000 databases and 1,250,000 interfaces, and still expanding at high rate, e.g. increasing 3 – 7 times between 2000 and 2004. The Deep Web pages are distributed across wide range of subject areas, with significant share of e-commerce sites. Although the non-commerce sites are gradually being hidden behind HTML forms. The Deep Web pages are mostly structured, providing the data objects in attribute-value pairs. This feature comes from the back-end structure of Deep Web sites, that use databases running in relational or objective paradigm. As the generated Web pages are the result of database queries, which provides data in highly structured manner, the Web pages design is noticeably influenced by the data structures. This is reflected in table-like layouts or database-style tuples on the Web pages. Also the structure of Deep Web sites tend to be quite shallow (He et al., 2007), about 94% of the Deep Web databases is located not deeper than on the 3rd level of a Website.

The ability of Web crawlers to index Deep Web pages is limited, although some efforts are made in this area. The problem is that new approaches had to be developed, and the large base of research work on surface Web is inapplicable to the deep Web due to differences in using basic building blocks of the Web - pages and links. The classical approaches to Web search were treating the Web as a repository of documents, and hyperlinks in the documents were used to traverse the Web to get access to more documents. Later on, the nature of the Web, being a graph of interlinked documents, was used to improve the search results. Analysing graph structure allowed to improve the accuracy of search. The Web graph structure can be easily