On Building a Search Interface Discovery System

Denis Shestakov*

Department of Media Technology
Aalto University, Espoo, Finland-02150
firstname.lastname@tkk.fi

Abstract. A huge portion of the Web known as the deep Web is accessible via search interfaces to myriads of databases on the Web. While relatively good approaches for querying the contents of web databases have been recently proposed, one cannot fully utilize them having most search interfaces unlocated. Thus, the automatic recognition of search interfaces to online databases is crucial for any application accessing the deep Web. This paper describes the architecture of the I-Crawler, a system for finding and classifying search interfaces. The I-Crawler is intentionally designed to be used in the deep web characterization surveys and for constructing directories of deep web resources.

1 Introduction

Since current-day web search engines do not crawl and index a significant portion of the Web, web users relying on search engines only are unable to discover and access a large amount of information from the non-indexable part of the Web. Specifically, dynamic pages generated based on parameters provided by a user via web search forms (or search interfaces) are poorly indexed by major searchers and, therefore, are scarcely presented in searchers’ results. Such search interfaces provide users with an access to myriads of databases which content comprise a huge part of the Web known as the deep Web [20].

Due to the huge volume of information in the deep Web, there has been a significant interest to approaches that allow users and computer applications to leverage this information. For example, works 3172126 discuss how to query the contents of web databases via their search interfaces that, as assumed, have been already discovered. However, the large scale of the deep Web makes this assumption unrealistic. In fact, even national 2223 or specific community-oriented (e.g., bioinformatics community) parts of the deep Web are too large to be fully discovered. Manually created collections, such as the one for the Molecular Biology domain 9, are of great help to corresponding communities, but, because of the size of the deep Web, they uncover just the top of the iceberg of all community-specific resources. Similarly, existing directories of deep web resources (i.e., directories that classify web databases in some taxonomies) have extremely low coverage for online databases. For example, Completeplanet.com,

* Work done while at University of Turku.

© Springer-Verlag Berlin Heidelberg 2010
the largest of such directories, with around 70,000 databases covered only 15.6% of the total 450,000 web databases as of April 2004 [6]. Clearly, currently existing lists of online databases do not correspond to the scale of the deep Web. Besides, a technique for automatic finding search interfaces is of great interest to the people involved in such directories’ building and maintaining.

The dynamism of the Web, when new sources being added all the time and old sources modified or removed completely, is another challenge that requires the automation of the search interface discovery process. One of the problems here is that search interfaces are very sparsely distributed over the Web, even within specific domains. For example, it was shown in [19] that only one of 38 (in average) web sites has at least one web search form, or, according to [5], approximately one of thousand web pages related to movies contains a movie search form.

To summarize, while relatively good approaches for querying the contents of web databases are now available, one cannot fully utilize them as most search interfaces are undiscovered. Thus, the ability to automatically locate search interfaces to web databases is crucial for any application accessing the deep Web. In this paper, we describe a system for the automatic detection of search interfaces and identifying database domains accessible via these interfaces. The proposed system called I-Crawler is specifically designed to be used in the deep web characterization studies as well as for automatic building of web databases’ directories.

Our first contribution is an efficient recognition of non-HTML (i.e., interfaces implemented as Java applets or in Flash) and JavaScript-rich search interfaces. Such interfaces are beginning to prevail on the Web and, hence, handling them is crucial for any search interface discovery system. Additionally, we suggest to divide all forms into two groups based on the number of visible form controls and demonstrate that such separation improves the system accuracy. Our second contribution is an approach combining pre-query and post-query techniques for web database classification.

The rest of the paper is organized as follows. Section 2 discusses related work. In Section 3, we present our motivation and challenges. Section 4 describes our approach on search interface discovery. Section 5 presents the architecture of the I-Crawler system. In Section 6, we report our experiments with the I-Crawler and some preliminary results. Finally, Section 7 concludes the paper.

2 Related Work

Surprisingly, finding of search interfaces to web databases is a challenging problem in itself. Indeed, since several hundred thousands of databases are available on the Web [6], even an expert in a highly specialized domain cannot be aware of most relevant databases.

There are two classes of approaches to identify search interfaces to online databases: pre-query and post-query approaches. Pre-query approaches identify searchable forms on web sites by analyzing the features of web forms. Post-query