Chapter LXII
Deep Web:
Databases on the Web

Denis Shestakov
Turku Centre of Computer Science, Finland

INTRODUCTION

Finding information on the Web using a web search engine is one of the primary activities of today's web users. For a majority of users results returned by conventional search engines are an essentially complete set of links to all pages on the Web relevant to their queries. However, current-day searchers do not crawl and index a significant portion of the Web and, hence, 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 are not indexed by search engines and cannot be found in searchers' results. Such search interfaces provide web users with an online access to myriads of databases on the Web. In order to obtain some information from a web database of interest, a user issues his/her query by specifying query terms in a search form and receives the query results, a set of dynamic pages which embed required information from a database. At the same time, issuing a query via an arbitrary search interface is an extremely complex task for any kind of automatic agents including web crawlers, which, at least up to the present day, do not even attempt to pass through web forms on a large scale.

Content provided by many web databases is often of very high quality and can be extremely valuable to many users. For example, the PubMed database (http://www.pubmed.gov) allows a user to search through millions of high-quality peer-reviewed papers on biomedical research, while the AutoTrader car classifieds database at http://autotrader.com is highly useful for anyone wishing to buy or sell a car. In general, since each

searchable database is a collection of data in a specific domain it can often provide more specific and detailed information that is not available or hard to find in the indexable Web. The following section provides background information on the non-indexable Web and web databases.

BACKGROUND

Conventional web search engines index only a portion of the Web, called the publicly indexable Web, which consists of publicly available web pages reachable by following hyperlinks.

Figure 1. Indexable and non-indexable portions of the Web and deep Web

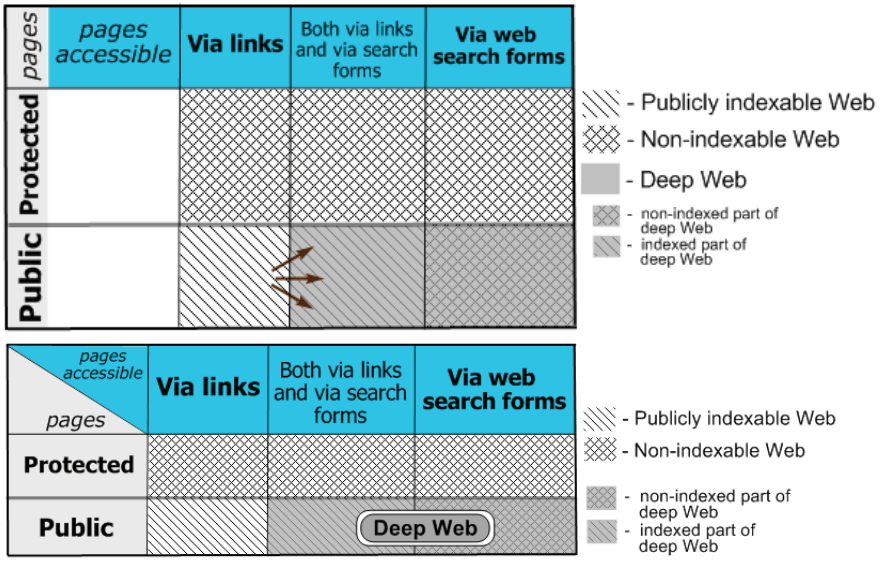

Figure 2. User interaction with web database