Chapter 9

A New Approach for Building a Scalable and Adaptive Vertical Search Engine

H. Arafat Ali
Mansoura University, Egypt

Ali I. El Desouky
Mansoura University, Egypt

Ahmed I. Saleh
Mansoura University, Egypt

ABSTRACT

Search engines are the most important search tools for finding useful and recent information on the Web today. They rely on crawlers that continually crawl the Web for new pages. Meanwhile, focused crawlers have become an attractive area for research in recent years. They suggest a better solution for general-purpose search engine limitations and lead to a new generation of search engines called vertical-search engines. Searching the Web vertically is to divide the Web into smaller regions; each region is related to a specific domain. In addition, one crawler is allowed to search in each domain. The innovation of this article is adding intelligence and adaptation ability to focused crawlers. Such added features will certainly guide the crawler perfectly to retrieve more relevant pages while crawling the Web. The proposed crawler has the ability to estimate the rank of the page before visiting it and adapts itself to any changes in its domain using.

INTRODUCTION

The World Wide Web (WWW) is a huge source of distributed data that is increasingly growing with dynamic nature. Millions of Web pages are added daily and others are deleted. The numerous size and dynamic nature of the Web makes it difficult to search efficiently. Many problems
arise when trying to build an information retrieval system on the Web because the data is distributed, unstructured, heterogeneous, as well as, highly volatile with a large volume and poor quality. Today, three search services are available which are directories, meta-search engines, and search engines.

Search directories, also called Web portals or taxonomies, organize Web pages into a tree-like topic hierarchy. General topics can be subdivided into more specified topics or categories. The tree-like structure of the Web directories allows non-expert users to find useful information easily (Zhdanova & Fensel, 2005). Some of the most popular portals on the Web today are: Yahoo, LookSmart, and Open Directory Project. Directories have two main drawbacks (i) the taxonomies are manually populated; hence, only a small portion of the Web could be covered and (ii) The directory structure is defined according to the knowledge of the human constructors, which means that different hierarchies could be built for the same page collection by different constructors.

A meta-search engine is a system that sends the user query to a several search engines via a number of interface agents, then collects the results from the different search engines, and present them to the user. Meta-search engines are powerful since the best search engine covers only about 16% of the Web (Gulli & Signorin, 2005).

Search engines are information retrieval systems that help users to find what they want on the Web. The user sends his query to the search engine in a form of keywords. Then, the search engine searches its database and retrieves the pages relevant to the search query. Finally, the query result is introduced to the user in the form of a ranked list of relevant pages. Most search engines rely on crawlers to traverse the Web to collect pages, pass them to the indexer, and then follow links from one page to another. Web crawlers have the ability to index thousands of pages per day. This overcomes the limitations of the Web portals. They also keep track of changes made to pages visited earlier. No page is added to the search engine database until the crawler visits that page. For the purpose of illustration, the basic components of a typical search engine are shown in Table 1 and Figure 1.

Search engines suffer from many problems such as low precision and recall, freshness problem (Cho & Garcia-Molina, 2000), poor retrieval rate, long list of results which consume user time and effort, a huge amount of rapidly expanded information which causes a storage problem, and finally, a large number of daily hits which means most search engines are not able to provide

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crawler</td>
<td>Also called robot, spider, or Web worm; they are used to retrieve Web pages, read them, pass them to the indexer, then follow links to the next page (Marios, Athena, &amp; Loizos, 2005).</td>
</tr>
<tr>
<td>Indexer</td>
<td>It receives the pages retrieved by the crawler, analyzes the various elements of each page like title, headings, body text, then extracts the main features of the page, finally dumping the retrieved features into the database (Qiu, Shao, &amp; Zatsman, 2003).</td>
</tr>
<tr>
<td>Database</td>
<td>Stores information retrieved from each page the indexer analyze.</td>
</tr>
<tr>
<td>Query manager</td>
<td>It has four basic functions: (i) it receives the query and reformulates it into a suitable database query, (ii) it retrieves relevant pages from a database, (iii) it ranks results according to the user query, and (iv) it performs analysis on the retrieved pages including categorization and text summarization (Jianjun, David, &amp; Yuan, 2000).</td>
</tr>
<tr>
<td>User interface</td>
<td>This is the part that the users see. It allows the user to enter his query, sending the query to a database via a query manager and finally displays the search result the user.</td>
</tr>
</tbody>
</table>