Chapter 4
Issues and Challenges in Web Crawling for Information Extraction

Subrata Paul
Vignan Institute of Technology and Management, India

Anirban Mitra
Vignan Institute of Technology and Management, India

Swagata Dey
MIPS, MITS, Rayagada, India

ABSTRACT

Computational biology and bio inspired techniques are part of a larger revolution that is increasing the processing, storage and retrieving of data in major way. This larger revolution is being driven by the generation and use of information in all forms and in enormous quantities and requires the development of intelligent systems for gathering, storing and accessing information. This chapter describes the concepts, design and implementation of a distributed web crawler that runs on a network of workstations and has been used for web information extraction. The crawler needs to scale (at least) several hundred pages per second, is resilient against system crashes and other events, and is capable to adapted various crawling applications. Further this chapter, focusses on various ways in which appropriate biological and bio inspired tools can be used to implement, automatically locate, understand, and extract online data independent of the source and also to make it available for Semantic web agents like a web crawler.

DOI: 10.4018/978-1-5225-2375-8.ch004

1. INTRODUCTION

Web search engines are today used by everyone with access to computers, and those people have very different interests. But search engines always return the same result, regardless of who did the search. Search results could be improved if more information about the user was considered. Web crawlers are computer programs that scan the web, 'reading' everything they find. Web crawlers are also known as spiders, bots and automatic indexers. These crawlers scan web pages to see what words they contain, and where those words are used. The crawler turns its findings into a giant index. The index is basically a big list of words and the web pages that feature them. So when you ask a search engine for pages about hippos, the search engine checks its index and gives you a list of pages that mention hippos. Web crawlers scan the web regularly so they always have an up-to-date index of the web. Archie is the first search engine, created in 1990. Downloaded directory listings of all files on anonymous FTP sites, and created searchable database (Gupta, 2011). In a generalized web crawler, two different users get different results for the same query, sometime when the transverse links-paths are from different direction. Web search engines have broadly three basic phases. These are crawling, indexing, and searching. The information available about the users' interest is considered in some of those three phases, depending on its nature.

Information retrieval (IR) is finding material of unstructured nature such as text, images, videos, and music. These materials are extracted from large collections usually stored on computers. For decades information retrieval is used by professional searchers, but now-a-days hundreds of millions of people use information retrieval daily. The field of IR also covers document clustering and document classification. Given a set of documents, clustering is the task of coming up with a good grouping of the documents based on their contents. Given a set of topics, and a set of documents, classification is the task of assigning each document to its most suitable topics, if any. IR systems can also be classified by the scale on which they operate. Three main scales are IR on the web, IR on the documents of an enterprise, and IR on a personal computer.

When doing IR on the web, the IR system will have to retrieve information from billions of documents. Furthermore, the IR system will have to be aware of some webs, where its owners will manipulate it, so that their web can appear on the top results for some specific searches. Moreover, the indexing will have to filter, and index only the most important information, as it is impossible to store everything. During the latest years, the Web 2.0 has emerged. With this development, web users not only retrieve information from the web but also add value to the web. If the search engines are capable to retrieve implicit information (such as the number of