VI

ABSTRACT

Web mining is an important sub-branch of data mining. The data mining technology adopts data integration method to generate Data warehouse. Data warehouse gives Relation Rules, Cluster characters and get the useful Module Prediction and knowledge evaluation. Web data are semi-structure and heterogeneous in nature. Web mining technology is different from pure mining technology which is on database. A Multi-Agent Cooperation Module, which acts as knowledge discovery, exaction machine, Generalization Machine and Analysis Machine has been developed. The communication between Multi-Agent system and Agent Naming Server will enable to gain the useful knowledge in their related task.

The Web crawler is a part of the search engines that fetches web pages along with the other multimedia files from the Web and pre-process the page for extracting information. Web crawlers are also known as robots, spiders, worms, walkers, and wanderers. Web crawling is the process used by search engines to analyze pages from the Web. Web crawlers are essential component for many web applications including search engines, digital libraries, online marketing, and web data mining. MAN (Mobile Agent Name Management) is a three Dimensional column (Agent Naming, Server Location, Birth Time, Death Time, Mining tasks). In Multi-Agents Mining system every Mobile Agent has an Index to an ANS (Agent Naming Server), which accepts Agents searching results and simply processes it. Like router ANS can also notify its neighbor ANS by sending message, which includes valuable information of it. By exchanging information among ANS the valuable information analyzes and cluster. The ANS has a table to maintain the information about what it focuses on.
There are lots of problems in the existing work. Some of them are, not traversing the secure HTTP pages, taking more time to retrieve hidden pages, less precision and less recall. Crawlers are not efficient in mobile applications. Retrieving Related and Non Related web pages are not efficient with respect to time. The challenges of Hidden Web are, crawling Deep Web, Searching the Hidden Web, Updating the Hidden Web pages.

The aim of this thesis is to extract the efficient information from web mining. This thesis addresses the challenges of the existing work on deep web. The Secure HTTP pages are retrieved by a new framework called Hidden GNU WGET (HIGWGET). The results are taken for several keywords to find out the proposed Hidden web page, web crawler’s differ from the traditional web search engine and a better search is found, which includes several secure and hidden pages input in the search results. A new focused crawler analysis model based on improved genetic algorithm is designed. The experimental results show that the focused crawler URL analysis model based on improved genetic algorithm can improve accuracy rate, recall rate effectively. Improved Specific Crawler technique is used to give the Related and Non-Related pages in less time.
This thesis presents the design and implementation of different crawlers for hidden web search. The designed crawlers are used for the best accurate results.

Chapter one presents the details of Introduction to the web Mining. Discusses different types of Web Mining and its sub categories. This chapter also deals with the applications of Web Mining.

Chapter two describes the Hidden Web Search and explanation of the crawler, indexing, query processing. This chapter also dwells upon the classification of different crawlers and also deals with the problems in traditional Crawlers.

Chapter three describes the Survey on Web Crawlers for Searching Hidden Pages. It also explains the advantages and disadvantages of different crawlers.

Chapter four describes how to extract hidden pages in a secure way. The traditional crawlers search only the HTTP pages. This chapter describes new ways to extract hidden content from https pages.

Chapter five describes a framework for Deep Web Crawler. Here Genetic algorithm is used. It also describes how to improve precision and recall.

Chapter six describes how to improve the specific search. Here a new improvised specific crawler is introduced. By using this crawler better results are found.

Chapter seven describes conclusion of the thesis and scope for the further work. The results of all the implemented crawlers are to be extended for the ecommerce and mobile applications as the future work.