

Problem statement: Build the web crawler to pull product information and links from an e-commerce website (python).

Objective :- To understand the working of web crawler & implement it on local - terminal soft drive port.

Outcome:- At the end of the assignment the students should have

1. Understand how web crawler works.

Infrastructure: Desktop / Laptop system with Linux or its derivatives (Debian, Fedora, Ubuntu etc.) installed.

Software used: Linux & Windows OS.

Theory: A program that searches documents for specified keywords and returns a list of the documents where the keywords were found is a search engine. Although search engines are really a general class of programs,

the term is often used to specifically describe systems like Google, Alta Vista, Excite that enable to search for documents on the World Wide Web and USENET newsgroups.

- Typically, a search engine works by sending out a spider to fetch as many documents as possible. Another program, called an indexer, then reads these documents & creates an index based

(3. word) 20. all from 1991

on the words contained in each document. Each search engine uses a proprietary algorithm to create its indices such that, ideally, only meaningful results are returned for each query.

- They search the internet - based on important words
- They keep an index of the words they find, and where they find them.
- They allow users to look for words or combinations of words found in that index.

- Figure the shows general search engine architecture. Every engine consists of a crawler module to provide the input for its operation.

- crawlers are small programs that browser the web on the search engines behalf, similar to how a human user would follow links to reach different pages.
- The crawlers extract URLs appearing in the mentioned pages, and give this information to the crawler control module.

Web crawlers:

- Web crawlers are program that exploit the graph structure of the web to move from page to page. It may be observed that the noun 'crawlers' is not indicative of the speed of the these programs,

as they can be considered fast. A key motivation for designers web crawlers has been to retrieve web pages and add them or their representations to a local repository. Such a repository may then serve particular application needs such as those of a web search engine.

- In its simplest form, a crawler starts from a seed and then uses the external links within it to attend to other pages.
- The crawler is the means by which web crawler collects pages from the web. It operates by iteratively downloading a web page, processing it and following the links in that page to other web pages, perhaps on the server.

Robot Exclusion:

The robot exclusion standard, also known as the Robots Exclusion protocol or robots.txt protocol, is a convention operating web crawlers and other web robots from accessing all or a part of a website which is otherwise publicly viewable. Robots are often used by search engines to categorize and archive web sites, or by webmasters to protect source code.

The standard is different but can be used in conjunction with Sitemaps, a robot inclusion.

Algorithm:

- 1) Make user interface
- 2) Input the URL of any website
- 3) Establish HTTP connection
- 4) Read HTML page source code
- 5) Extract Hyperlinks of HTML page
- 6) Display the list of hyperlinks.

Conclusion:

Implementation is concluded by stating the basic working of web crawler.

21-10-24