Web crawling for business intelligence

Field of Use

Web crawler is a program that **surfs the Internet** in a designated fashion which could be automated, methodical or in an orderly way, and **helps you** extract the information you need from the web.

This type of applications is useful for media & marketing domains, business intelligence and program testing.

Technologies Used

Java RMI, Hibernate, Java Persistence API, Struts, Win API, C++, XPCOM, XUL

Client

The company is in business of analyzing global web traffic trends and identifying cyber-security threats in real time

Industry

Business intelligence & information security

Project Summary

The solution is a web crawling technology that allows gathering and extracting large amount of data from a variety of online sources. The crawling process is started in both automated and manual manner, emulates human-user behavior, and includes auto-filling of contact forms. The end users have controls for the active management of the crawling process. They are working with the application through convenient interfaces for data search and analytics.
Technologies in Detail

Actimind has developed in-house service for gathering and extracting data from malicious websites.

The web crawling application finds target on any website using a number of criterion, and then extracts and stores all the necessary information from the web page including its content, screenshots, HTTP headers, etc.

Access points for data extraction are added manually, or can be found by the application from multiple sources – emails, advertising on Facebook, etc.
The web crawling system includes:

**Crawling Engine** – a plug-in for Firefox browser written in Java/Javascript. It allows the browser managing and gathering all the necessary data. This engine is also used for searching tasks to gather on the web sites such as Facebook.

**Features:**

- Processing server and client redirect pages of different types
- Identifies and goes through javascript-links and elements with event handlers
- Emulates user behavior and analyzes a page from their point of view, ignoring hidden elements
- Identifies and fills in HTML forms (using Roboform plug-in)
- Identifies and loads any resources requested by the web page, including dynamic ones
- Catches pop-up windows and opens new windows/tabs
- Adaptive pattern of website crawling is choosing the most probable way of an average user behavior – going through links taking into account the history of opening links and the appearance of links and forms.
- Guaranteed data retention when a browser crashes
- Ability to stop processing a task, and continue it in another Firefox instance

**Process Manager** – java service that supports a pool of running Firefox processes.

**Features:**

- Runs new Firefox instances
- Manages memory consumption and closes/restarts the Firefox browser (if needed)
- Stops not responding browser instances
- Closes Firefox when there are no tasks to complete

**Task Manager** – java service that distributes tasks on data mining in the Crawling Engine. The service communicates with the engine through Java RMI.

**Scheduler** – java service that collects sources for data mining and creates new tasks.

**Features:**

- Takes mail via POP3/IMAP protocols, analyzes and decides upon data mining targets
- Visits Facebook pages and gathers the targeted data in the adverts
- Analyzes context bound adverts on several web sites and in search results of Google, Yahoo, Bing

**MySQL DB** – stores data mining tasks and results, history of previous tasks. The latter is used when tasking decisions.

**WebServer** – Struts 2 web application.
Features:

- Adds new tasks manually
- Watches tasks accomplishment and analyzes the results
- Sets up the Scheduler – email and Facebook accounts, frequency of data mining, etc.

Analyzing Tools – a set of applications for processing the data gathered.

Features:

- Data indexing for a fast-access retrieval, using the dtSearch
- Finds duplicates via screenshots and content, ignoring changing content such as banner ads, current date/time, etc.
- Half-automated distribution of gathered data by categories using resembling screenshots and content

This crawling system works in the production and provides:

- transparency due to using a real browser, proxy server and emulating human behavior
- high trafficability of random web sites
- defining and passing though the HTML-forms of any complexity
- daily processing of thousands of web sites on one server
- ability to distribute the work load among the servers (when needed)