InfoCrawler 4.6

InfoCrawler is a software that allows you to crawl and index various types of documents, accessing data from various resources : Intranets, public WEB sites, News groups, FTP sites, local or remote file systems.

Main Features

Distributed architecture: InfoCrawler was designed from the ground up for distributed architecture, it is a 100% java WEB service, and can be executed permanently on one or more machines. Communicating using XML, its components can be installed on different machines: the administration, the spider, and the indexing engine

Intuitive administration: Using its own WEB based administrating interface, you can administer and monitor the different collections in a very user-friendly manner. The simplicity and flexibility limits the total costs of ownership.

Optimized crawling: Thanks to its multi-threaded architecture, InfoCrawler can spider many collections in parallel, and can have many threads per collection.

Powerful indexing: Using a very powerful indexing engine to index the documents, InfoCrawler can index various file types : HTML files, Microsoft office documents, PDF, XML, and more than 240 other types of documents.

Open technology: InfoCrawler does not use any proprietary technology, URLs are maintained using mySql database, the WEB administration is done using Apache Tomcat and JSP, the communication between the administration and the spider is done using XML, and the spider itself is 100% java.

Flexible: Being compatible with standards like HTML, XML, JSP, Java, and JDBC, InfoCrawler can be integrated easily in large projects.

Unique features: InfoCrawler has some unique features like the JavaScript interpreter, the native XML indexing, the automatic classification, or the intelligent URL management.

Detailed Features and Benefits

Administration Features

Easy index building: Building an index can be as simple as entering in your home page URL. Even more refined indices are effortlessly created through an easy-to-use interface.

Flexible administration : All administrative configuration information can be viewed and changed at any time, and from any location. All user and administrative functions are available through a Web interface. The easy-to-use interface makes server administration quick and efficient.

Flexible administration (2) : The administrator can control different spiders from the same machine, you only need to indicate the server where the spider is running.

Administration protected by password: The administration interface is protected by a crypted password, this password can be changed.

Administration dashboard: InfoCrawler displays a dashboard of all the collections that are defined, with a quick look, the administrator can see in real time the evolution of the different collections (number of URLs, number of indexed documents, errors, and more)

Collection details: By choosing a particular collection, the administrator can watch all the different indicators of that collection in real time. You can see the crawling time, the total number of URLs, the number of URLs that are pending, the number of servers, and much more. You can even see the indicators per server.

URL or document details: The administrator can see all the details concerning a URL or a document, like the last modification date, the description, or the error code.

Customize : Being developed using JSP, the administration screens can be customized easily.

Server Features

Distributed Architecture: InfoCrawler was designed from the ground up for distributed architecture, its components can be installed on different machines: the administration, the spider, and Searchserver

100% Java : The spider server is 100% java, porting from Windows to Linux took less than one hour.

100% XML : Administrating the spider is done using XML, there is no API to learn, you can control the spider by just sending XML commands using any software and from any system.

Multi-threaded spider: A multi-threaded spider means better indexing performance, which means lower load on the system. The administrator can adjust the number of threads allowed, so you get the best performance possible for the amount of bandwidth available.

Error recovery: Automatic recovery from system crashes, maintaining all index updates. No need to worry about losing changes on large indexing runs. Server restarts automatically upon failure.

Network-friendly spider: The administrator can configure a delay between two HTTP requests, so that the WEB server that is being spidered is not overloaded.

No CGI: Because InfoCrawler does not use CGIs, user performance is better, and there is less load on the server.

Per-collection data directories: You can specify data directories on a per-collection basis. You can even specify that a collection keep its data on a different file system from the rest of InfoCrawler's data. This is useful if you want to spread out InfoCrawler's data across several file systems.

Logging: The logging level is completely configurable.

Log rotation: Log files according to your needs. InfoCrawler can open new log files on a daily, weekly, or monthly basis. Easy-to-use log file administrative interface provides instant access to current and archived logs.

Automatic classification: InfoCrawler can analyze the documents and classify them automatically into categories.

Indexer scheduling : You can establish a schedule for each collection to specify that the indexer run only at certain times. This feature is useful if you would like to specify that a collection only index during off-hours.

Crawling Features

Multiple collection types: Collections can be created by spidering across a network, scanning a file system, or following newsgroups.

Multiple protocols: The spider can index across Hypertext Transfer Protocol (HTTP), File Transfer Protocol (FTP), Usenet (NNTP), secure HTTP (HTTPS), and file protocol (FILE)

Multiple document types: InfoCrawler supports Plain text, HTML, XML, RTF, Microsoft Word, Microsoft Excel, Microsoft PowerPoint, Adobe Acrobat (PDF), and more than 240 other document formats.

Multiple root URLs: An index can be built by using more than one URL for its root. You can adjust the depth of links to for the spider to index by specifying the number of hops from the root, so you can build an index to be as small or as large as you want.

Flexible URL filters: URL filters give administrators the power to include or exclude content. So when you have a collection that includes any particular URL, you can still filter out given directories, sites, or file types within that same URL.

Robots.txt support: InfoCrawler allows for full robots.txt support according to the robots.txt standard. Administrators can customize the user agent to meet their needs. The crawler recognize also the tags “NO INDEX“ and “NO FOLLOW“.

Proxies and Firewalls: The spider can operate through network proxies and firewalls. Indexed sites can be in another part of the network or even outside the secure network, such as public Web servers.

Password authorization: If desired, the spider can access content protected by passwords, allowing authorized users to search all material on the site.

Usenet collections: Create collections that subscribe to Usenet groups. Simply point InfoCrawler at the Usenet server and indices will automatically be created and updated..

Intelligent duplicate elimination: As documents are retrieved by the spider, they are checked to see if the same content is already in the index. If this is the case, only one copy will be maintained. The spider knows when a significant change has been made, and whether the document still qualifies as a duplicate.

Intelligent obsolete recognition: The system can synchronize the index with the Web server at any interval. Obsolete documents are deleted, and new documents are added, all of which is real-time updated to the collection.

Deleting URLs in real-time: When you remove a URL it is updated to the collection in real-time. You can remove a single URL, or all URLs matching a specific query. Obsolete URLs are recognized and deleted, as are duplicate URLs. Adaptive spidering means more up-to-date information.

Spider control: InfoCrawler allows you to control precisely how much information you want to gather, you can specify a maximum number of links, a maximum number of documents to index, the minimum/maximum size of downloaded files, the depth of links with respect to the base URL, and many other features.

Flexible file types: You can configure precisely the types of files that you want to spider, either by using the mime-type or by using the file extension.

Connection timeouts: Specify connection timeout times for all server network connections made by InfoCrawler from a simple administrative interface.

Site fetching : By default the spider download the documents, index them, then delete them, only indexes are kept and the reference to the corresponding documents. You can also fetch the hole site and reproduce its arborescence locally.

Java script interpreter: Unlike other crawlers, InfoCrawler does not search for “hard coded” URLs, because most URLs are built dynamically. Instead of that, InfoCrawler has its own java script interpreter so that is can execute the code and extract the correct URLs.

Indexing Features

META Search: Create your own searchable field by using the HTML META tag name, a corresponding column will be automatically created in InfoCrawler so that you can not only search this specific META but also display it in a result list.

XML field searching: Element names can be mapped to InfoCrawler field names. This allows searches to be focused on a part of an XML document. Using an administrative interface you can create different XML mappings.

Stop word exclusion: InfoCrawler installs many stop word files that corresponds to different languages.

Proximity indexing : You can choose to index data using different proximity schemes: character, word, phrase, paragraph, or all.

Periodicity: You can configure the indexing to be immediate or periodic, you can also configure the indexing period.

Thesaurus expansion: Administrators can create custom thesauri, so your organization's special vocabulary can be integrated into search. So for an automobile Web site, a query for both "hood" and "bonnet" will find similar results.

Multiple languages: InfoCrawler supports English as a standard. Additional component provides lexical analysis and localization for Dutch, French, German, Italian, and Portuguese, Spanish, Swedish, Danish, Finnish, Norwegian, Traditional & Simplified Chinese, Japanese and Korean, plus other languages.

Dictionaries: By default InfoCrawler uses standard dictionaries, but you can add your own custom dictionary.

Requirements

Supported platforms: Windows NT4, Windows 2000, Unix, Linux

Memory :128 MB minimum, 512 MB or more is recommended .

Disk space : 136 MB for the server, 70 MB for the administration.