

Administration and Troubleshooting Guide



Administration and Troubleshooting Guide

Note

Before using this information and the product that it supports, be sure to read the general information under “Notices and trademarks” on page 63.

This edition applies to version 8, release 4, modification 2 of IBM OmniFind Yahoo! Edition (product number 5724-R21) and to all subsequent releases and modifications until otherwise indicated in new editions.

© **Copyright International Business Machines Corporation 2006, 2007. All rights reserved.**

US Government Users Restricted Rights – Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

ibm.com and related resources	v	Starting and stopping the search engine as a Windows service	36
How to send your comments	v	Starting and stopping the search engine as a Linux service	37
Contacting IBM	v	Changing the administrator password	38
Getting started with IBM OmniFind Yahoo! Edition	1	Changing a lost password and user name	38
How the search engine works: collections, crawlers, indexes, and the search page	2	Backing up and restoring a collection.	38
Creating a collection.	4	Administering the search engine from the command line	39
Crawling Web sites to retrieve documents	7	Removing the search engine	45
Crawling protected sites	10	Troubleshooting crawler, search, and system problems	47
Accessing Web sites through a proxy server	11	Monitoring the Web crawler.	47
Configuring how the Web crawler interacts with Web servers	12	Web crawler cannot crawl or cannot retrieve documents.	48
Crawling file systems to retrieve documents	13	Symptoms.	48
Managing the search experience	21	Resolving the problem.	48
Customizing the search page	21	Checking the status of a crawled document	51
Managing metadata for search queries	22	File crawler problems	52
Monitoring search results.	23	Symptoms.	52
Adding synonyms	24	Resolving the problem.	52
Adding featured links	26	Users cannot find the appropriate documents	52
Adjusting search results ranking	28	Symptoms.	53
Language processing options	30	Diagnosing the problem	53
Enabling n-gram segmentation for Chinese, Japanese, and Korean	30	Resolving the problem.	53
Deleting documents from the collection	32	Checking for errors in the system logs	55
Enabling or disabling document cache	33	Glossary	57
Managing the system	35	Notices and trademarks	63
Starting and stopping the search engine	35	Notices	63
Stopping the search engine from the administration console.	35	Trademarks	65
Stopping the search engine from the operating system	36	Index	67

ibm.com and related resources

Product support and documentation are available from [ibm.com](http://www.ibm.com).

Support and assistance

Product support is available on the Web. Click Support from the product Web site at:

OmniFind Yahoo! Edition

<http://www.ibm.com/software/data/enterprise-search/omnifind-yahoo/support.html>

PDF publications

You can view the PDF files online using the Adobe Acrobat Reader for your operating system. If you do not have the Acrobat Reader installed, you can download it from the Adobe Web site at <http://www.adobe.com>.

See the following PDF publications Web sites:

Product	Web site address
IBM OmniFind Discovery Edition	http://www-1.ibm.com/support/docview.wss?rs=3035&uid=swg27008552
IBM OmniFind Enterprise Edition	http://www-1.ibm.com/support/docview.wss?rs=63&uid=swg27007911
IBM OmniFind Yahoo! Edition	http://www.ibm.com/support/docview.wss?rs=3193&uid=swg27010191

How to send your comments

Your feedback is important in helping to provide the most accurate and highest quality information.

Send your comments by using the online reader comment form at https://www14.software.ibm.com/webapp/iwm/web/signup.do?lang=en_US&source=swg-rcf.

Contacting IBM

To contact IBM customer service in the United States or Canada, call 1-800-IBM-SERV (1-800-426-7378).

To learn about available service options, call one of the following numbers:

- In the United States: 1-888-426-4343
- In Canada: 1-800-465-9600

For more information about how to contact IBM, see the Contact IBM Web site at <http://www.ibm.com/contact/us/>.

Getting started with IBM OmniFind Yahoo! Edition

IBM® OmniFind™ Yahoo! Edition provides a simple yet powerful search engine for Web sites and file systems. The search engine is easy to set up and customize, and with it you can quickly provide users of your own Web site with a powerful way to find the information that they are looking for.

Before your users can search for content, you must create a collection (the first one is created for you during installation) and fill the index with documents from sources that you select.

As the search engine administrator, to set up and customize a unique search experience, you can:



Create separate collections to allow different groups of users to search separate sets of content

For example, you can create a collection for employees to search for documents on your intranet, and another collection for search on human resources documents. You can have as many as five collections.



Specify which Web sites and file systems contain the documents that you want to make searchable

As soon as you specify one or more Web sites or file systems, the search engine starts retrieving and analyzing documents so those documents are available for search.



Manage the search experience

Find out what users are searching for, whether they are finding what they want, and view the most popular queries. To make documents even easier to find, you can:

- Define synonyms, which help users find documents even if queries use different wording than what is in the documents.
- Define featured links, which provide you a way to ensure that specific documents appear at the top of the results for particular queries.
- Adjust ranking, which helps you influence the way results are ordered.



Customize the search page

You can customize fonts, logos, and colors of the search page. Also, the search page is seamlessly integrated with Yahoo! search so that users can easily expand their search by querying Yahoo! Search. Users can search the Web, news, images, video, and use other Yahoo! search options.



Manage the system

The administration console helps you manage your system by displaying warning and error logs. You can also do tasks such as change passwords and check document status.



Write custom applications

Application developers can extend the capabilities of the search engine by writing applications to index content from additional sources, such as databases or content management systems. Application developers can also write custom search pages by using their favorite programming language. For information about the APIs, see the *Programming Guide and API Reference*.



Troubleshoot crawler, search, and system issues

Installation and upgrade instructions for IBM OmniFind Yahoo! Edition are available on the Web at <http://omnifind.ibm.yahoo.net/> and <http://www.ibm.com/support/docview.wss?rs=3193&uid=swg27010188>

How the search engine works: collections, crawlers, indexes, and the search page

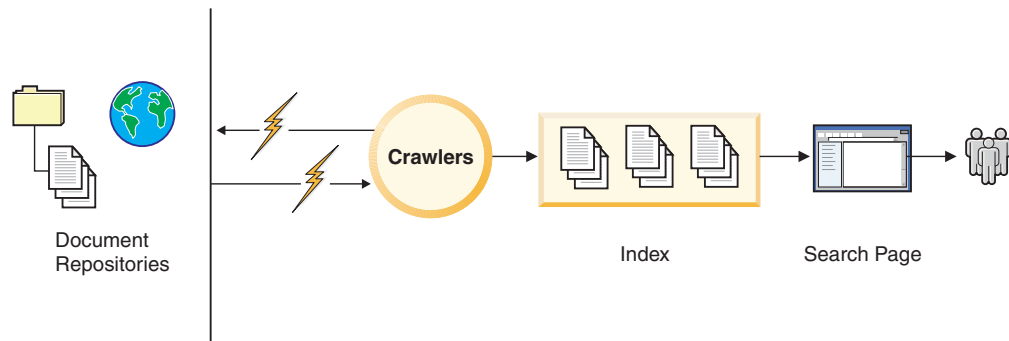
After you specify which sources you want to collect documents from, the crawlers immediately start crawling those sources and returning documents to the index. Shortly after the documents are added to the index, the documents are available for search.

Your company users can open the search page and search for documents. To ensure that users are finding the documents that they need, you should monitor the crawlers and query activity. You can monitor the system, the queries, and the crawlers from the administration console.

The search results are produced by the following process:

1. Documents that you want to be able to search are crawled.
2. Those documents are processed and added to the index.
3. After documents are added to the index, they are available for users to search.

The following diagram shows how data flows through the search system:



Crawlers

Crawlers are software programs that retrieve documents from document sources. Crawlers are typically designed to retrieve documents from one specific document source, such as Web sites or file systems.

Crawlers retrieve the document links and the search engine adds those links to the index. Crawlers find documents by following links or directory hierarchies.

IBM OmniFind Yahoo! Edition uses two types of crawlers:

- The Web crawler for Internet and intranet sites
- The file system crawler for local file systems

Web crawlers

Web crawlers are often called *spiders* or *robots* because of the way they behave. The Web crawler does the following tasks:

- Starts crawling one or more starting Web site addresses (URLs). Starting from the main, or most general URL, the Web crawler retrieves pages and then extracts all HTML links.
- Retrieves all the links that are found and extracts links from the new pages again and continues recursively.
- Tracks pages that it already visited so that it does not always revisit the same content.

For example, a Web crawler starts crawling the site `http://www.example.org/`. However, this site has many more pages (documents) such as `www.example.org/personnel/private/records`. The Web crawler starts at the top level or most general Web site address and crawls down to the more specific Web pages.

However, unlike a file directory system, the HTTP protocol does not provide a method to show all Web site addresses under a specific Web page. The Web crawler must extract all content from a Web site by following all HTML links systematically to discover all pages.

Web crawlers, unlike file system crawlers, crawl continuously. After a Web crawler crawls an entire Web site, the crawler becomes idle. The crawler will automatically start again after a predefined time (about 36 hours). However, if the Web crawler notices that documents are being updated more often every 36 hours, the crawler will crawl those sites more often. You do not need to restart the crawler.

File system crawlers

A file system crawler is designed specifically to crawl local file system directories or directories that are mounted locally, such as `C:\My Documents` (Windows®) or `/data/document/` (Linux®).

For example, on Windows, you can crawl a shared network folder that is mapped as a drive, such as `\\fileserv\shared\documents\` that is mapped to `F:\`. However, you cannot crawl a shared network folder directly.

On Linux, you can crawl any directory that is mounted on the file system.

Unlike the Web crawler, when the directory crawler finishes crawling all levels of the directory, it stops. If you want to recrawl the directory, you must start the directory crawler again.

Collection

A collection is an index and its associated crawlers, synonyms, featured links, and configuration settings.

An empty collection called *Default* is included during installation. You can change the name of the default collection later, or specify a different collection to be the default for search.

You can add collections that have crawlers for different sets of documents for search, depending on the needs of your users. For example, you can create a

collection for human resources managers, and another collection for all employees to search your intranet.

Index

An index is a set of crawled documents that can be searched.

The index in the Default collection initially has no documents. Therefore, you must specify the sources to crawl. After those sources are crawled, they are parsed and added to the index. Shortly after they are added to the index, you can do searches from the search page.

Search page

The search page is the graphical user interface that you use to search for documents in an index.

You can configure the appearance of the search page to suit your needs. For example, you can change logos, fonts, colors, banners, and so on.

You can also create a search page interface by using the application programming interfaces (APIs). You can build custom applications with the APIs to send and receive search requests, add documents to an index, and so on.

Creating a collection

The search engine is installed with one default collection. You can add as many as four more collections, up to a total of five collections.

You might have different groups of search users who need to search different sets of documents, or you might need to restrict some set of documents to a selected group of users, such as your Human Resources department. You can create a separate collection for each group of search users.

Tip: You can change the name of a collection later.

To create a collection:

1. On the Collection Status page, click **Create New Collection**.
2. On the Create Collection page, specify a name for the new collection. The collection name can include any characters that are valid in a directory name on the operating system on which the search engine is running.
3. Optional: Select **Cache Documents** if you want documents that are included in the index to be cached locally.
4. Optional: Select **N-Gram Segmentation** to enable n-gram segmentation processing. This option is available only if your browser is configured to display Chinese, Japanese, or Korean.
5. Click **OK**.

After the new collection is created, the Collection Status page for the new collection opens. You can then add Web sites or file directories to crawl for inclusion in the index.

If you have more than one collection, specify a default collection on the Manage Collections page or on a Collection Status page for the selected collection. The search page will search the default collection when first opened, or users can select another collection to search.

Crawling Web sites to retrieve documents

You tell the search engine what documents that you want to make searchable by specifying what Web sites to crawl. When you add Web sites, the crawler automatically starts crawling the site and retrieving documents for search by adding them to the index.

Web crawling is a continuous, automatic process and runs in the background. You can continue to work in the administration console while the Web crawler runs.

Restriction: If you host a Web site on the same server as the search engine server, you must use the external name of your server in the address. Do not use localhost. For example, change the address `http://localhost/docs` to `http://myintranet.example.org/docs`. If the Web pages are crawled with `http://localhost/` in the URL, then your users cannot open the pages when they click a search result. Their browsers will try to open a page on their own computers.

The system displays the most recent collection that you worked on by default. To select another collection, click the **Collections** tab, then click the tab for the appropriate collection.

If you need to add more Web sites or exclude parts of sites:

1. On the Collection Status page, click **Web Crawler**, then **Add Web Site**.
2. On the Add Web Site page, enter the full starting URL. For example, enter `http://www.example.org`.
3. Optional: Modify the URL in the **Include URLs that match these patterns** box. This address is automatically generated based on the URL that you entered as a starting URL.
4. Optional: Specify which parts of the Web site that you want to exclude. Add one URL per line in the Exclude box. If you want to exclude Web documents according to the pattern of a URL, add an asterisk (*) anywhere in the URL. The most common reason to exclude parts of a Web site from crawling is to avoid returning restricted or duplicated content in search results. For example, if you exclude `http://www.example.org/*printer*`, then `http://www.example.org/products/` is crawled, but `http://www.example.org/forprinteronly` is not crawled.
5. Click **Save**. The crawler continues to crawl with these changes.

The search engine starts making documents available for search while it is still crawling. It might take up to several minutes until the first documents are available for search.

You can check that the documents were crawled and added to the index by using the Document Status page.

On the Collection Status page, you can monitor the crawl rate and how many pages were crawled since the search system was last started.

The crawl rate, query response times, and other indicators that you see on the Collection Status page are not exact measurements. They are general indicators of what is happening in your search system.

When the Web crawler recrawls

The Web crawler revisits every crawled Web page to determine if the Web page has been updated or removed. The Web crawler adjusts its schedule to recrawl more or less often depending on how often documents change. The initial recrawl period is 36 hours. If documents change infrequently, the recrawl period expands, up to a maximum of 48 hours. If documents change frequently, the recrawl period becomes shorter, to a minimum of 24 hours. However, RSS feeds are recrawled every hour.

Changes to any of the following settings trigger the Web crawler to recrawl earlier:

- User-Agent string
- Included or excluded Web addresses
- HTTP proxy server configuration
- HTTP basic or form-based authentication

Even if the authentication or proxy server configuration for a crawled Web page changes, the page remains in the index until the next recrawl.

You cannot configure the Web crawler to recrawl on a particular schedule, but you can start a complete recrawl by selecting **Start full recrawl** on the **Start Crawling Web sites** button.

Web crawler rules

The search engine must retrieve and analyze documents before users can search for them. The process of retrieving documents is called crawling.

You can specify which Web pages the Web crawler should retrieve and which ones should be excluded from the crawl.

Including or excluding Web pages defines the boundaries that the Web crawler can crawl. You define this *crawl space* by providing pattern-matching rules that are based on the URLs.

A default set of rules is automatically calculated. However, you can modify these rules, and the Web crawler automatically adjusts the crawl space.

The two types of rules are:

Include

Control which URLs should be crawled.

Exclude

Controls which URLs should not be crawled.

Every link that the Web crawler encounters is evaluated based on these rules, and the Web crawler crawls a site according to the result of the evaluation.

The patterns in the rules are expressed by using the wildcard character (*). For example, the URL pattern `http://*.example.org/*` matches the following Web sites:


```
http://www.example.org/  
http://intranet.example.org/  
http://hr.intranet.example.org/  
http://www.example.org/products/  
http://www.example.org/products/detail.jsp?id=12345
```

The same URL pattern (`http://*.example.org/*`) does not match:

```
https://www.example.org/  
http://www.example.com/  
http://example.org/
```

The first URL uses `https`, not `http`. The second URL uses `.com`, not `.org`. The third URL has no period (`.`) before the word `example`. Therefore, none of the URLs match the pattern.

A URL must match at least one of the include rules if the site is to be crawled. If a URL matches both the include and exclude rules, the most specific rule is the one that takes precedence, for example:

Include rule:

```
http://www.example.org/*  
http://www.example.org/secure/public/*
```

Exclude rule:

```
http://www.example.org/secure/*
```

The Web crawler follows a link to `http://www.example.org/secure/public/account.html` because the include rule `http://www.example.org/secure/public/*` is more specific than the exclude rule `http://www.example.org/secure/*`.

If you update the configuration by modifying the rules, the crawler will crawl accordingly. For example, if `http://www.example.org/` is used as the starting URL with no specific rule provided, the system automatically generates the following include rule: `http://*.example.org/*`.

If you start crawling but find that the URL pattern is too broad (for example, because it includes `http://intranet.example.org`), you can modify it to include just `http://www.example.org/*`. The Web crawler removes all documents that are already crawled from the index that do not match this pattern. This operation is not immediate and can take some time depending on the number of pages being removed.

Also, the Web crawler does not crawl beyond a depth of more than 15 subsections (or slashes) in the Web address.

Tuning the crawl space

You might not have extensive knowledge about the content on the Web sites that you are crawling, so you can start with a broad crawl space definition and then monitor the types of URLs that the Web crawler is finding. You can then add exclude rules to remove unnecessary content.

It is important to narrow the crawl space as much as possible because it can increase search quality significantly. Users are more likely to find the right document and not be bothered by a large number of irrelevant results.

Check for:

- Different rendering of the same content. For example, the following exclude rule `http://www.example.org/products/*&view=printable` removes the printable rendition of a page.
- A page that is used to send comments about any page, passed as a parameter, for example, `http://www.example.org/comment.do?*`.
- A subsection of a Web site that contains a large amount of unmanaged, uninteresting, or non-authoritative content, for example, `http://www.example.org/forum/*`.

Crawling protected sites

You must provide authentication information, either basic or form-based, to crawl protected Web sites.

Prerequisite: Before you can enter authentication information for a Web site, you must have previously added the Web site to your list of sites to crawl.

To set up authentication so that you can crawl protected Web sites:

1. On the Collection Status page, click **Web Crawler**.
2. On the Web Crawler page, click the **Password-Protected Web Sites** button



for the Web site that you need to set up.

3. From the drop-down list, select the type of authentication:

If you select:	Provide the following type of information, for example:
HTTP basic authentication	User name and password and then continue with the next step.
HTML form-based authentication	<p>Form name (optional) Example: loginPage</p> <p>Form action Example: <code>http://www.example.org/authentication/login.do</code></p> <p>HTTP method: POST or GET Example: POST</p> <p>Form parameters (optional) Example: <code>userid</code> and <code>myuserID</code></p>

If you configure form-based authentication, the crawler will submit the form with the specified parameters, obtain a cookie as a token of authentication, and use that cookie for all subsequent requests. The crawler might have problems if the form changes. In the future, if you have problems crawling the site that requires form-based authentication, ensure that the information that you added is consistent with the form.

You can extract the information that you need for form-based authentication from the HTML page that contains the form:

- a. Open a Web browser and go to the form that is used for authentication.
- b. Use the view source feature of your browser to view the HTML source. Find the `FORM` tag for the form.

Tip: There can be multiple forms in one page. Be sure to find the one that is used for authentication.

- c. Within that form, find all the INPUT or TEXTAREA tags, including those that are not displayed by the Web browser (type="hidden").
- d. Extract the information to configure the Web crawler authentication

The following example shows the information in the form element from the <http://www.example.org/authentication/login.jsp> Web site:

```
<form name="loginPage" action="login.do" method="post">
  Login: <input id="userid" name="username"/><br>
  Password: <input id="password" type="password"
           name="password"/><br>
  Male: <input type="radio" checked="checked" name="gender"
         value="male"><br>
  Female: <input type="radio" name="gender" value="female"><br>
  <input type="hidden" name="login-form-type" value="pwd"/>
  <input type="submit" value="Submit" name="submitButton"/>
</form>
```

If this form information came from the Web site that you are trying to crawl, you can derive the following information that you need to set up form-based authentication:

Form name

loginPage

Form action

<http://www.example.org/authentication/login.do>

Form method

POST

Form parameters

Name	Value
userid	<i>provide the user ID</i>
password	<i>provide the password</i>
gender	<i>male or female</i>
login-form-type	pwd
submitButton	Submit

4. Click **Save**.

Accessing Web sites through a proxy server

If you installed the search engine on a computer that requires an HTTP proxy server to access Web sites, you must provide the host name or IP address, port, and authentication information of the proxy server.

To specify HTTP proxy server information:

1. On the Collection Status page for the selected collection, click **Web Crawler**.
2. On the Web Crawler page, click **Edit proxy server settings**.
3. Add the following information in each field:
 - Enter a full server name or an IP address, for example, `myproxy.mysite.com` or `127.0.0.1`.
 - Enter the port number on which your proxy server accepts connections, for example: `8088`

- Optional: If your proxy server requires a user name, enter a user name that has permission to use the proxy server.
 - Optional: If your proxy server requires a password, enter the password that authenticates the user name that you entered.
4. Click **Save**. This change takes place immediately even if you are currently crawling the site.
 5. Optional: To test your proxy server settings, enter a URL for a valid Web site and click **Test**.
 6. Click **Close**.

Configuring how the Web crawler interacts with Web servers

You can configure the way the Web crawler interacts with Web servers by specifying an e-mail address for notifications, a User-Agent string, and a crawl delay.

It is strongly recommended that you specify an e-mail address as a courtesy to Web site administrators.

To configure the Web crawler:

1. On the Collection Status page, click **Web Crawler**, then click **Edit Crawler Settings**.
2. Enter an e-mail address, for example, `myname@domain.net`.
Web site administrators can trace the e-mail address in their Web server logs, which can be used if they think that crawling is impacting normal operation of their Web server.
3. Enter a User-Agent string. The User-Agent string identifies the Web crawler to the Web server that is being crawled. The Web site administrator can allow or forbid specific crawlers based on the crawler's User-Agent string. Contact the Web site administrator to agree on a User-Agent string.
4. Specify a crawl delay in milliseconds.
The crawl delay controls the speed of the crawler so that its activity does not overwhelm Web servers. It sets an interval of time that the Web crawler waits between two requests to a given Web server. The Web crawler can generate many requests to a Web server, which can cause the site to fail or slow its normal operation. If the crawler makes too many requests too quickly, the Web server administrator might deny the crawler access.
The default crawl delay of 500 milliseconds (two requests per second) should put only a minimal load on Web sites that you crawl. A crawl delay of 0 seconds (no delay) might put excessive load on crawled Web sites.
5. Click **Save**.

Crawling file systems to retrieve documents

You can crawl only those file systems that are accessible to the search engine server.

Restriction: The file system crawler can crawl only local directories. Directories on a different server from the search engine server must be mounted locally (for example, a mapped network drive on Windows). To the crawler, mounted drives appear to be local. You cannot specify a file system using a Universal Naming Convention (UNC) directory path such as \\hostname\directory on Windows.

On the Collection Status page, you can monitor the activity of the crawler. If file systems are not being crawled, ensure that you can access the file system. You can crawl only file systems that are accessible to the search engine server.

The crawl rate, query response times, and other indicators that you see on the Collection Status page are not exact measurements. They are general indicators of what is happening in your search system.

If you never specified a directory to crawl, you can enter a directory to crawl in the Collection Status page.

The system displays the most recent collection that you worked on by default. To select another collection, click the **Collections** tab, then click the tab for the appropriate collection.

To crawl additional file systems:

1. On the Collection Status page, click **File System**, then **Add Directory**.
2. In the Add Directory page, enter or browse for a directory path. For example, enter C:\documents or /data/document/, depending on the operating system. Directory names are case sensitive.
3. Optional: Enter specific subdirectories that you want to exclude from crawling. To exclude documents from crawling based on a directory pattern, add an asterisk (*) anywhere in the directory path. You can use one or more wildcard characters. Wildcard characters other than the asterisk (*) are not supported.
4. Click **Save**. The crawler immediately starts crawling and runs in the background. You can continue to work in the administration console while the crawler runs.

When the file system crawler recrawls

Unlike the Web crawler, the file system crawler will not crawl the file system again unless you manually restart it. If documents in that directory change and you want to update the index with the new documents, you must start the file system crawler again from the Collection Status page or the File System Crawler page.

File system crawler rules

The file system crawler will crawl the entire directory that you specify and exclude any subdirectories that you specify. When it finishes crawling, it stops.

When you add a directory, the crawler crawls the entire directory, including all of its subdirectories. You can exclude files or subdirectories from the crawl by specifying exclude patterns. An exclude pattern is a path that possibly contains a wildcard character (*):

- A path matches a pattern if it can be obtained by replacing all occurrences of the wildcard character (*) in the pattern with arbitrary strings (including empty strings).
- If a subdirectory matches an exclude pattern, it is excluded from the crawl and so are all (nested) subdirectories and files that it contains. In other words, excluding a subdirectory removes the entire subtree under it.
- If a file name matches an exclude pattern, it is excluded from the crawl.

The following examples show how subdirectories can be excluded. For the examples, assume that you have the following directories and files:

```
C:\documents\personnel\All_Members.doc
C:\documents\personnel\All_Members.txt
C:\documents\personnel\All_Contractors.txt
C:\documents\personnel\LAB\SVL_DEV\DEV_Member.txt
C:\documents\personnel\LAB\SVL_QA\QA_Member.txt
C:\documents\personnel\LAB\YSL_DEV\DEV_Member.txt
C:\documents\personnel\LAB\YSL_QA\QA_Member.txt
C:\documents\personnel\LAB\YSL_QA\Contractors\contact.txt
```

The starting directory to crawl is C:\documents\personnel\.

Example 1

If the excluded pattern is *.txt, the crawler crawls only the directory and file C:\documents\personnel\All_Members.doc.

Example 2

If the excluded pattern is C:\documents\personnel\LAB*, the crawler crawls the following directories and files:

```
C:\documents\personnel\All_Members.doc
C:\documents\personnel\All_Members.txt
C:\documents\personnel\All_Contractors.txt
```

Example 3

If the excluded pattern is C:\documents\personnel\LAB*_QA\, the crawler crawls the following directories and files:

```
C:\documents\personnel\All_Members.doc
C:\documents\personnel\All_Members.txt
C:\documents\personnel\All_Contractors.txt
C:\documents\personnel\LAB\SVL_DEV\DEV_Member.txt
C:\documents\personnel\LAB\YSL_DEV\DEV_Member.txt
```

Example 4

If the excluded pattern is *Contractor*, the crawler crawls the following directories and files:

```
C:\documents\personnel\All_Members.doc
C:\documents\personnel\All_Members.txt
C:\documents\personnel\LAB\SVL_DEV\DEV_Member.txt
C:\documents\personnel\LAB\SVL_QA\QA_Member.txt
C:\documents\personnel\LAB\YSL_DEV\DEV_Member.txt
C:\documents\personnel\LAB\YSL_QA\QA_Member.txt
```

Supported file types

The search engine can process many different file types, such as Lotus® Word Pro®, Microsoft® Word, compressed files, and documents in markup languages such as HTML and XML.

The search engine supports the file types in the following table. Some file extensions vary depending on the version of the product that creates the file type.

Table 1. Supported file types

File extension	File type
123	Lotus 1-2-3® and Lotus 1-2-3 for SmartSuite®
DOC	Microsoft Word (after 2000)
HTML	HyperText Markup Language
JTD, JTI, JFW, JVW	Ichitaro
LWP	Lotus Word Pro
MPP	Microsoft Project
PDF	Portable Document Format
PPT	Microsoft PowerPoint
PRZ	Lotus Freelance
QPW	Quattro Pro
RTF	Microsoft Rich Text Format
SXC	OpenOffice Calc and StarOffice Calc
SXI	StarOffice Impress and OpenOffice Impress
SXW	StarOffice Writer and OpenOffice Writer
TAR	Tape Archive File (Each file in the archive is indexed individually.)
TAR.GZ	Tape Archive File that is compressed with GZIP (Each file in the archive is indexed individually.)
TXT	Microsoft WordPad (File types can vary.)
VSD	Microsoft Visio
WRI	Microsoft Write
WS	WordStar and WordStar 2000
XLS	Microsoft Excel
XML	Extensible Markup Language
ZIP	Compressed Archive File (Each file in the archive is indexed individually.)

Note: Some versions of a file type might not be supported. For example, Lotus WordPro SmartSuite 96 files are not supported on non-Windows operating systems.

Crawling databases to retrieve documents

You can include in your collection documents that are stored in a database by crawling supported local or remote databases. The database crawler connects by using JDBC drivers.

You can select which database tables to crawl and within each table the columns to crawl. If a primary key is specified for a table, the primary key appears as a unique identifier. If no primary key is specified in the table, you can select a column to be the unique identifier. The unique identifier is included in the Uniform Resource Identifier (URI) that identifies each crawled record.

The crawler automatically retrieves predefined metadata fields and includes them in the index. The crawler also includes column names as metadata in the index. You can change the field name for each column that is displayed to users by specifying a field name that is different from the column name. Users can search by using the predefined metadata fields, column names, and field names as metadata. If you specify a field name that is different from the column name, search users can search on the field name as metadata.

Restriction: You can crawl only supported databases that you can connect to by using a Java™ Database Connectivity (JDBC) driver.

Restriction: After a database is crawled for the first time, it will not be crawled again unless you start a recrawl manually.

Tip: If you do not know details about the database that you want to crawl, such as the correct JDBC driver, ask your database administrator for help.

The system displays the most recent collection that you worked on by default. To select another collection, click the **Collections** tab, then click the tab for the appropriate collection.

To crawl a database:

1. On the Collection Status page for the selected collection, click **Database Crawler**, then **Add Database**.
2. On the Add Database page, select:

Option	Description
Use an existing JDBC driver	Select a driver from the drop-down list. This option is available only if a JDBC driver is already configured.
Configure a new JDBC driver	Specify the driver name and path.

3. Click **Next**.
4. Specify a URL to connect to the database and a user name and password authorized to access the database.
5. Optional: Click **Test Connection** to test that the crawler can connect to the database.
6. Click **Next**.

7. Select a table from the database to crawl and click **Next**. You can add only one table at a time.
8. Specify the following information, as required for your needs:
 - Select columns to crawl from the database table.
 - Specify a field name. If you specify a field name that is different from the column name, the column name is mapped to the field name, and search users can search on the field name as metadata.
 - Always Search: Specify any columns that will always be searched regardless of whether the search query includes the field name. The metadata value is not broken into its smallest textual components.
 - Require Field Search: Specify any columns will be searched with any keyword, but the query must specify the field name. The values of fields that require field search are always broken into their smallest textual components and analyzed for linguistic variants.
 - Require Exact Match: Specify any columns for which query terms must match the field value exactly. Field values that require an exact match are not broken into basic text units but are added to the index as the entire value, and are case-sensitive. A search query must include the field name. If a column that is used as a unique identifier also requires an exact match you cannot change the exact match selection later.
 - If the table does not have a primary key defined, select at least one column to use as the unique identifier. The unique identifier is used as part of the URI that identifies crawled documents in the index.
 - You can also specify an SQL WHERE statement to limit the data that is crawled in this table. For example, specifying `EMPID > 500` causes the crawler to include employee number 600, but not employee number 400.
9. Click **Finish**.

Enabling search on additional database properties

You can enable and define additional properties of a database crawler, such as language or date fields, by editing a crawler table record.

Creating a database crawler by selecting tables from the database applies some default values to table properties. You can edit the table record to modify additional properties.

To edit a database table:

1. On the Database Crawler page, click the **Edit** button for the selected table.
2. Select one or more of the following options:
 - Click **Columns** and select or clear the columns to be crawled.
 - Click **Language** and choose whether to enable automatic language detection, or select a default language to use if automatic detection is not selected or the document language cannot be detected.
 - Click **Document Content** and select one database column to be crawled and analyzed for the content of documents stored in the column instead of data. If you know that a certain column contains documents, such as Word files or photographs, instead of only data, you can have the content of the documents analyzed for words and metadata.
 - Click **Date Fields** and select a date field in this table to use as the document date when returning search results. Optionally select a date field to use to determine when documents will be recrawled.

Database content types

The content types supported by the database crawler vary with each database.

Supported data types

The database crawler treats binary and large object (LOB) data types as documents that can be crawled for content and metadata. All other data types are crawled as data.

The database crawler supports the following data types.

Table 2. DB2® data types

Data Type	Crawled as data	Crawled as content
CHAR	X	
CHAR FOR BIT DATA		X
GRAPHIC	X	
VARCHAR	X	
VARCHAR FOR BIT DATA		X
VARGRAPHIC	X	
LONG VARCHAR	X	
LONG VARCHAR FOR BIT DATA		X
LONG VARGRAPHIC	X	
BLOB		X
CLOB	X	X
DBCLOB	X	X
INTEGER	X	
BIGINT	X	
DOUBLE	X	
REAL	X	
SMALLINT	X	
DATE	X	
TIME	X	
TIMESTAMP		
XML	X	X

Table 3. Oracle data types

Data Type	Crawled as data	Crawled as content
CHAR	X	
NCHAR	X	
VARCHAR2	X	
NVARCHAR2	X	
NUMBER	X	
LONG	X	
RAW		X

Table 3. Oracle data types (continued)

Data Type	Crawled as data	Crawled as content
BLOB		X
CLOB	X	X
NCLOB	X	X
FLOAT	X	
DATE	X	
TIMESTAMP	X	
TIMESTAMP WITH TIMEZONE	X	

Table 4. Microsoft SQL Server data types

Data Type	Crawled as data	Crawled as content
CHAR	X	
NCHAR	X	
VARCHAR	X	
NVARCHAR	X	
TEXT	X	X
NTEXT	X	X
SMALL INT	X	
TINY INT	X	
BIGINT	X	
DECIMAL	X	
FLOAT	X	
NUMERIC	X	
REAL	X	
DATETIME	X	
SMALL DATETIME	X	
BINARY		X
VARBINARY		X
TIMESTAMP		X
IMAGE		X

Managing the search experience

You can improve the quality of search results by defining synonyms, adding featured links, changing language settings, and adjusting ranking.

If users cannot find the information that they need, check whether a document exists that contains text that matches the query. Determine whether the document contains the terms users have in their queries and whether the document was crawled.

To check documents and to improve the search experience:

- Monitor search results. Review the query logs and reports to find out whether users are getting the results that they want.
- Check document status to find out whether a document was crawled, will be crawled, or is in the index.
- If users are not finding what they need, you can do one of the following tasks to improve search results:
 - Add synonyms
 - Add featured links
 - Adjust ranking
- If your collection includes documents that you do not want users to search, you can delete documents from the collection.

Tip: If you accidentally crawled sites or files that you do not want users to see, you can disable the search from the Collection Status page. You can then restart the search after you remove the documents by either deleting the Web site or file system or by excluding Web site subdomains or file subdirectories.

After you do any of these tasks to improve search results, you should do several searches with different queries to ensure that your changes produce the search results that you want.

Customizing the search page

You can change fonts, logos, colors, text, and other parts of the search page.

The preview page that you see in the customization editor functions exactly like the search page that your users see. You can use it to preview your layout by entering queries and seeing results.

To customize the search page:

1. From the **System** tab, click **Customize Search Page**.
2. In the Customization Editor page, click the section where you want to make changes:
 - **Page Layout**
 - **Results Layout**
 - **Yahoo! Features**
3. Within a section, click **Edit** to open a window where you can make changes.
 - a. Optional: Click **Upload** to upload images from your local file system.

Note: You can delete images that you upload, but you cannot delete images that were delivered with the search engine.

- b. Optional: Click **Choose Color** to select a color from the palette.
4. Optional: Click **Show Preview** at any time to see how your edits affect the appearance of the Search page.
5. Click **Save & Close** to save your changes.

Important: When you click **Save & Close**, your changes are applied immediately. All users who use the search page will see the changes immediately.

Managing metadata for search queries

Metadata is data about a document that is not part of the content of the document. You can create, edit, or map metadata fields to use in search queries.

Users of your search engine can search by using keywords for content that might appear in the title or text of a document. Users can also search by specifying values for metadata fields.

To be used in a search query, information associated with a document must be crawled and parsed before being added to an index. A basic set of common metadata is automatically identified by the search engine, and you can specify additional metadata for each collection.

The search engine supports a subset of metadata as defined by the Dublin Core Metadata Initiative (<http://dublincore.org>). Dublin Core metadata is identified with the prefix dc.

To create or edit a metadata field:

1. On the Collection Status page, click **Metadata**.
2. Click one of the following options:
 - **Add Metadata Field** to create and define a new metadata field.
 - **Edit Metadata Field** to view a list of metadata fields and to edit a field. You can delete a field by clicking the delete icon.
3. Click **Save**.

Metadata properties

Metadata properties can be specified and edited, with some restrictions.

The editable metadata properties are:

Name in Source

The name of the metadata field as it appears in the source.

Field names must begin with a letter and cannot include the following restricted characters: () : ^ @ [] " { } ~ * ? \ < > = !

Type Metadata fields can have one of three data types:

- **Text:** The field value is treated as text, even if it contains numbers. For example, a search query that includes a date will return metadata fields of type date, but not of type text.
- **Decimal:** The field value is treated as a number, but not as a date.

- **Date:** The field value is treated as a date, but not as a number. If a search query includes a date, only metadata fields with the date type will be returned.

Field Name

The name of the metadata field as it appears to search users and as it is stored in the index. The field name has the same restrictions as the name in source.

Always Search

Any columns that will always be searched regardless of whether the search query includes the field name. The metadata value is not broken into its smallest textual components.

Require Field Search

Any columns will be searched with any keyword, but the query must specify the field name. The values of fields that require field search are always broken into their smallest textual components and analyzed for linguistic variants.

Require Exact Match

Any columns for which query terms must match the field value exactly. Field values that require an exact match are not broken into basic text units but are added to the index as the entire value and are case-sensitive. A search query must include the field name.

Monitoring search results

Viewing query statistics can help you understand what queries are popular and whether queries are returning results.

You can view some query information on the Collection Status page. To view more query statistics:

1. On the Collection Status page, click **Query Statistics**.

Tip: If you enter search queries and later check the query statistics in the Query Statistics window, you might need to refresh the browser to see the most current information.

2. In the Query Statistics page, select a report from the drop-down list.

Option	Description
Most Popular Queries	Displays queries in order of popularity, and the quantity of results returned. The Results Clicked column displays the number of search results for each query result that users followed, and the Top Result Clicked column displays the search result that users most frequently followed for that query. Click the magnifying glass icon to view a detailed report of sessions that included that query.
Most Popular Queries Without Results	Displays queries in order of popularity that did not return any results.
Most Popular Queries Without Clicks	Displays queries in order of popularity that returned search results that users did not follow.

Option	Description
Most Popular Results Clicked	Displays search results that users followed, in order of popularity, with the queries that returned those results.
Query Logs	Displays all search queries in reverse chronological order. You can view the entire log or save it to your hard drive.

Query reports show information about the most popular queries with results and without results. You can see the number of results for each query and how many times that query was entered (count). With this information, you can determine how fast results are returned, when users use the search engine, and what keywords they use.

You should monitor what users are looking for to ensure that they are finding the correct documents. You should do a search with the most popular queries and review the results. If the results are not satisfactory, add synonyms, featured links, or adjust ranking to get more precise results.

You can adjust the maximum allowed query response time by specifying a time interval on the Collection Settings page. The default value is 3000 milliseconds (3 seconds).

Adding synonyms

Synonyms are different words with the same meaning. Sometimes, you might find that users are not finding the information that they are looking for with particular queries because the words in the queries are different from the words that are used in the documents.

You can improve the chances that users will find the correct document by defining synonyms that the search engine will use to expand what it searches for when one of the synonyms appears in a query.

For example, *films*, *movies*, or *moving pictures* might all be used to mean the same thing in documents or user queries. If a movie Web site uses only the word *movies* and a user searches with the query *films*, the user will not find the movie Web site. You can tell the search engine to expand the query for *films* to also search for *movies* or *moving pictures* by entering the following synonyms:

```
films
movies
moving pictures
```

Synonyms are especially helpful when users need to find documents that contain abbreviations or acronyms. For example, some users might use the query *NASA*. Other users might use *national space administration*. If you define the queries *NASA*, *national space administration*, *national aeronautics and space administration* as synonyms, you help users find the appropriate documents.

To add synonyms:

1. On the Collection Status page, click **Synonyms**, then click **Add Synonyms**.
2. In the **Synonyms** field, enter one or more words that have similar meanings. Add one word or phrase per line. Do not use punctuation, such as commas, semicolons, or quotation marks, around or between the synonyms.

3. Click **Save**.
4. To add another group of synonyms, click **Add Synonyms** again.

Synonyms are available immediately after you click **Save**. Also, note that the search engine highlights occurrences of synonyms in the search results.

You should check that the synonyms work by entering a query in the search page for each synonym and checking the results.

Related troubleshooting information

“Users cannot find the appropriate documents” on page 52

Importing synonyms

You can import synonyms from an XML file that you created in another instance of the search engine or created in an editing tool.

The file must use the following XML format:

```
<?xml version="1.0" encoding="UTF-8"?>
<synonymgroups version="1.0">
  <synonymgroup>
    <synonym>car</synonym>
    <synonym>automobile</synonym>
    <synonym>truck</synonym>
  </synonymgroup>
</synonymgroups>
```

To import synonyms:

1. On the Collection Status page, click **Synonyms**, then **Import Synonyms**.
2. Browse for or enter the XML file name.
3. Click **Import**.

The synonyms will display in the Synonyms page in the administration console.

If you get an error, check the format of the XML file.

Exporting synonyms

You can export synonyms that you create in the search engine to an XML file.

You must define at least one synonym before you can export synonyms.

You can also export synonyms to an XML file, edit the file, then import it to another instance of the search engine.

To export synonyms to an XML file:

1. On the Collection Status page, click **Synonyms**, then **Export Synonyms**. The **Export Synonyms** button appears only if you have defined at least one synonym.
2. Use the save function in your browser to save the file.

When you export synonyms, the search engine creates one file that contains all groups of synonyms.

Adding featured links

To help users find specific Web site or file system documents more quickly, you can add a featured link that appears at the top of the search results and has a different appearance than the other results on the page.

For example, if you have a Web site that describes different types of Siamese cats and users have trouble finding this site or they need it often, you can set up a featured link to make this site easier to find.

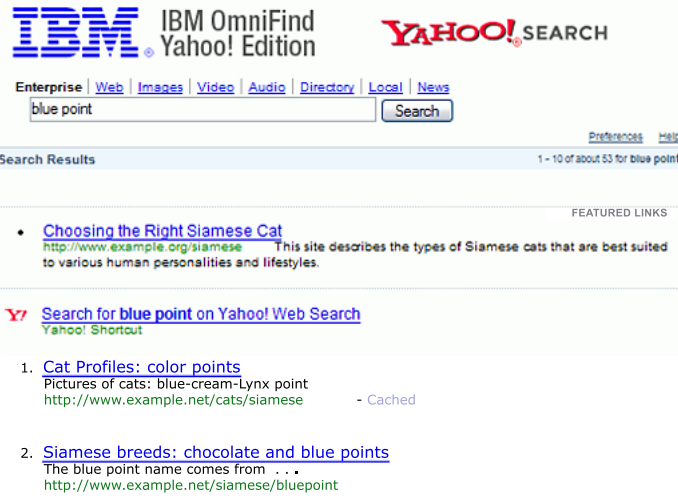
To add a featured link:

1. On the Collection Status page, click **Featured Links**, then **Add Featured Link**.
2. Enter the featured link information:

Featured link information	Example
<p>Queries</p> <p>For a featured link to be returned in the search results, the words in a query must exactly match (except for order of the terms) the keywords that you specify here. For example, the queries <i>blue point</i>, <i>blue +point</i>, <i>Blue Point</i>, and <i>point blue</i> return the same featured link. However, the queries <i>blue point cat</i> will not return the featured link. The search engine ignores wildcard characters (*) and exclude signs (-).</p> <p>Enter one query per line. Do not use punctuation. Enter as many queries as necessary.</p>	<p>siamese balinese cat burmese cat himalayan cat blue point chocolate point</p>
Web Address (URL)	<p>http://www.example.org/siamese</p> <p>file://hostname/C:/cats/siamese.doc</p>
Featured Link Label	Choosing the Right Siamese Cat
Featured Link Summary (optional)	This site describes the types of Siamese cats that are best suited to various human personalities and lifestyles.

3. Click **Save**.

After you define the featured link, you can check that it works by entering one of the queries in the search page. For example, for the Siamese cat example, if you enter the query *blue point* in the search page, you see the following featured link:



Related troubleshooting information

“Users cannot find the appropriate documents” on page 52

Importing featured links

You can import a featured link from an XML file that you created in another instance of the search engine or created in a text editing tool.

The file must use the following XML format:

```
<?xml version="1.0" encoding="UTF-8"?>
<quickLinks version="1.0">
  <quickLink url="http://www.example.org/siamese"
    ID="5c6f5758010f35291f720-7ff3">
    <title>Choosing the Right Siamese Cat</title>
    <summary>This site describes the types of Siamese cats</summary>
    <keywords>
      <keyword value="blue point"></keyword>
      <keyword value="siamese"></keyword>
      <keyword value="balinese cat"></keyword>
    </keywords>
  </quickLink>
</quickLinks>
```

The quickLink element in the featured links XML file supports an optional ID attribute. If the ID attribute is present, it must contain a unique ID for the XML file. If the ID attribute is not present, the search engine assigns an ID to the XML file when it is imported. The ID attribute is written to the XML file when the featured links are exported in the administration console.

During import, if a featured links XML file already exists in the search engine with the same ID as another XML file, the featured link in the search engine is replaced by the one that is defined in the XML file. You should allow the search engine to generate an ID for the featured links XML file.

To import an XML featured links file:

1. On the Collection Status page, click **Featured Links**, then **Import Featured Links**.
2. Browse for or enter the file name.
3. Click **Import**.

The featured link will appear in the Featured Links page in the administration console.

If you get an error, check the format of the XML file.

Exporting featured links

You can export featured links that you create in the search engine to an XML file.

You can also export featured links to an XML file, edit the file, and import it into another instance of the search engine.

To export featured links to an XML file:

1. On the Collection Status page, click **Featured Links**, then **Export Featured Links**. The **Export Featured Links** button appears only if you have defined at least one featured link.
2. Use the save function in your browser to save the file.

Adjusting search results ranking

You can adjust the way search results are ranked by enabling or disabling specific ranking factors.

Adjust the ranking only if you are sure that search results are not adequate for your users. The search engine is optimized to return the best search results for most types of queries. Before you adjust ranking, try adding synonyms or featured links to return results that users need.

Tip: Before you adjust ranking, determine what documents are good results for the user query and that these documents are being indexed and contain the query terms. Then, decide on a set of queries to use for a test search. Each time that you adjust a ranking factor, enter the queries and check the results. Remember that adjusting ranking factors can improve the search results for some queries but adversely affect the search results for other queries.

To adjust document ranking:

1. On the Collections Status page, click **Ranking**.
2. Enable or disable one or more of the ranking factors:
 - Document modification date
 - URL or file path depth
 - Web links analysis

Tip: Disable one ranking factor at a time. Then, enter several queries and check the results. You can revert to the default ranking factors by clicking the **Load Defaults** button.

3. Click **Save**.

Related troubleshooting information

“Users cannot find the appropriate documents” on page 52

How search results are ranked

Search results are returned based on four ranking factors: document modification date, URL or file path depth, Web links analysis, and keyword match. Keyword

match carries the most weight and cannot be disabled. The search engine uses a preset ranking that is appropriate for most Web site and file system searches.

Ranking is relative, not absolute. The search engine must weigh the relative ranking factors to return results. Enabling or disabling a ranking factor does not guarantee a particular ordering of results. For example, if you enable only the date-based ranking, an older document might still be returned before a newer document, if the older document has more keyword matches than the newer one.

The order in which documents are returned is determined by a combination of the ranking factors. Therefore, each of these factors can only increase the likelihood that a particular document appears higher in the search results. But the combination of the other factors typically outweighs the effect of a single factor.

The search engine identifies all documents that match the query and ranks them according to a combination of ranking factors. The ranking determines in which order results are returned to the search user. The search uses the following ranking factors:

Keyword match

Keywords are the terms that users enter to do a search. The search engine determines for every result how closely it matches the query. The more occurrences of the search terms a document has and the closer these occurrences are to each other, the higher the result will likely show in the list of results.

Document modification date

Documents that are newer are more likely to be returned before documents that are older.

Ranking by modification date can be especially important for documents in directories. However, document dates can be typically unreliable for Web site documents because the last modification date that is returned by some Web servers does not reflect the true age of the documents.

The document modification date ranking factor is enabled by default.

URL or directory path depth

URL or directory path depth refers to the length or number of slashes in a URL or the number of subdirectories in file paths. For example, the URL `www.example.org/personnel/private/records` is deeper (has more slashes) than the URL `www.example.org/personnel`. The file directory `C:\My Company\personnel\private\records` is deeper than `C:\My Company\personnel`.

By enabling URL or directory path depth ranking, documents that have shorter URLs or directory paths are more likely to be returned before documents that have longer URLs or directory paths. For example, a document with the URL `http://www.example.org/soccer` will be ranked higher than a document with the URL `http://example.blog.com/team/members/gwen/soccer.htm`. This ranking factor is especially helpful for Web documents because it favors home pages and entry pages over content that is hidden deep inside a site.

The URL or directory path depth ranking factor is enabled by default.

Web links analysis

A document that has many links that point to it is more likely to be returned before a document with fewer links that point to it. The search engine assumes that documents that are frequently linked to are typically more interesting. Documents that are not frequently linked to are deemed less interesting.

The Web links analysis ranking factor is enabled by default.

Language processing options

A dictionary pack is included with the search engine, which can help to process non-English documents and queries. Also, you can enable n-gram support for languages such as Chinese, Japanese, and Korean.

Dictionary pack for processing queries

You will see a list of languages that are available for processing queries in the Preferences window from the search results page. See the help from the search results page for more information about query language selections.

Spelling suggestions and language selections

The search engine returns documents that contain linguistic variations of the query terms. The variations of a word depend on the language of the query. By default, the search engine assumes that the query is in the language specified by the browser locale. In the Preferences page from the search page, users can select a different query language. The language that users select in the Preferences page overrides the browser settings.

In addition to finding variations of query terms, the query language is also used for spelling suggestions.

N-gram segmentation

As an alternative to dictionary-based word segmentation, you can enable n-gram segmentation support for languages such as Chinese, Japanese, and Korean. The option to enable or disable n-grams is available only if your browser language setting is Chinese, Japanese, or Korean.

Enabling n-gram segmentation for Chinese, Japanese, and Korean

You can enable n-gram segmentation to do specialized linguistic analysis on Chinese, Japanese, and Korean languages. You can enable n-gram segmentation only when you create a collection.

Prerequisite: You can enable n-gram segmentation only if your browser language setting is Chinese, Japanese, or Korean.

To enable n-gram segmentation:

1. Create a new collection.
2. Select **N-gram Segmentation**.
3. Click **Save**.

Processing for Chinese, Japanese and Korean documents

You can process documents with dictionary-based segmentation or with n-gram segmentation.

For a search engine, getting good search results depends mostly on the techniques that are used to process text. For most languages, the first step in text processing, after the text is extracted from the document, is to identify the individual words in the text. This is generally referred to as segmentation. For most languages, white space (blanks, end-of-lines, and certain punctuation) can be used to recognize word boundaries. However, Chinese, Japanese, and Korean do not use white space between characters to separate words, so other techniques must be used.

IBM OmniFind Yahoo! Edition has two methods to handle Chinese, Japanese, and Korean:

- Dictionary-based word segmentation (also called morphological analysis)
- N-gram segmentation

Dictionary-based word segmentation

This technique uses a language-specific dictionary to identify words in the sequence of characters in the document.

- Advantage: Provides more precise search results because the dictionaries are used to identify word boundaries
- Disadvantage: Can miss specific matching results

N-gram segmentation

This technique avoids the problem of identifying word boundaries, and instead indexes overlapping pairs of characters. Because the search engine uses two characters, this is also called bigram segmentation.

- Advantage: Always returns all matching documents that contain the search terms
- Disadvantage: Can return documents that do not match the query

By default, the system has a pre-configured index that uses n-gram segmentation for Chinese, Japanese, and Korean. You can recrawl the index to use dictionary-based word segmentation. Note that you will lose all previously indexed data, so it is better to make decide which method to use before starting to crawl and index documents.

To show how both types of linguistic processing work, examine the following text in a document: election for governor of Kanagawa prefecture. In Japanese, this text contains eight characters. For this example, the eight characters are represented as A B C D E F G H. A sample query that users might enter could be election for governor, which is four characters and are represented as E F G H. Note that the document text and the sample query share similar characters.

If you use n-gram processing:

After the document is crawled, the search engine segments the text election for governor of Kanagawa prefecture into the following sets of characters: AB BC CD DE EF FG GH

The sample query election for governor is segmented into the following sets of characters: EF FG GH. If users search with the sample query

election for governor, the document will be found by the query because the tokens for both the document text and the query appear in the same order.

When you enable n-gram segmentation, you will likely see more results but possibly less precise results. For example, in Japanese, if you search with the query Kyoto and a document in your index contains the text City of Tokyo, the query Kyoto will return the document with the text City of Tokyo. The reason is that City of Tokyo and Kyoto share two of the same Japanese characters.

If you do not use n-gram processing:

After the document is crawled, the search engine segments the text election for governor of Kanagawa prefecture into the following sets of characters: ABC DEF GH.

The sample query election for governor is segmented into the following sets of characters: EF GH. The characters EF do not appear in the tokens of the document text. Note that even though the document does not have EF, it does have DEF.

The document text contains DEF, but the query contains only EF. Therefore, the document is less likely to be found by using the sample query.

When you do not enable n-gram segmentation, you will likely see more precise results but possibly fewer results.

Deleting documents from the collection


You can delete individual Web sites, file directories, or all documents from the collection's index.

Deleting Web sites from the collection

You can delete Web sites from the collection. You might need to wait several minutes to several hours before the documents are removed.

To delete a Web site from the collection:

1. On the Collection Status page, click **Web Crawler**.

2. Click the **Delete** button  next to the Web site that you want to delete.

The search engine requires some time depending on the number of crawled documents for that Web site to remove the documents after you delete a Web site. Some of those documents might still be available for search immediately after you delete the site.


To check that a document is still in the index, click **Document Status** and enter a URL for a specific document.

Deleting directories from the collection

You can delete directories from the collection. You might need to wait several minutes or several hours before the documents are removed depending on the number of documents in the directory.

To delete a directory from the collection:

1. On the Collection Status page, click **File System Crawler**.

2. Click the **Delete** button  next to the file system that you want to delete.

Some of those documents might still be available for search immediately after you delete the directory. The index requires some time to remove the documents after you delete a directory, depending on the number of crawled documents in that directory.


To check that a document is still in the index, click **Document Status** and enter a directory for a specific document.

Deleting databases or tables from the collection

You can delete databases or database tables from the collection. You might need to wait several minutes or several hours before the documents are removed depending on the number of documents in the database.

To delete a database or table from the collection:

1. On the Collection Status page, click **Database Crawler**.

2. Click the **Delete** button  next to the database or table that you want to delete.

Some of the documents might still be available for search immediately after you delete the database or table. The index requires some time to remove the documents after you delete a database or table, depending on the number of crawled documents in that database.

To check that a document is still in the index, click **Document Status** and enter the URL for a specific document.

Enabling or disabling document cache

The document cache contains copies of all of the documents in the index. The search engine can return a cached version of each document that is returned in a search result.

The search engine can also convert binary documents into HTML documents for viewing using the View as HTML link on the search page. This allows users who do not have the appropriate plug-ins to see the text in the binary documents.

Caching makes it possible to retrieve documents even if they are temporarily unavailable on the original Web site or directory. Not all parts of the document can be retrieved if the original site or directory is unavailable. For example, images are not cached. If the original server is not available when the cached document is displayed, the images might be broken.

Disabling document caching increases indexing speed and saves significant disk space. However, the search results will not show the Cached or View as HTML link if document caching is disabled.

If you disabled caching and re-enable caching later, the index will have a mix of documents that are cached and not cached. Some documents might not display the Cached or View as HTML link.

By default, document caching is enabled.

To disable document caching:

1. On the Collection Status page, click **Collection Settings**.
2. Click **Delete and Disable Document Cache**.

Managing the system

You can start and stop the server and administration console, start and stop crawlers, change passwords, and review system errors.

You can also configure the search engine to run as a Windows or Linux service, which allows the search engine to be started automatically when the system is started, back up and restore the system, and administer the system from the command line.

Starting and stopping the search engine

To start the search engine, you start the server, then log in to the administration console.

Prerequisite: You must enable cookies in your browser to use the administration console and the search page.

If you have never logged into the administration console, you are prompted to create a user name and password.

To start the search engine:

1. Start the server:

Operating system	Action
Linux	Run the following command: INSTALL_ROOT/bin/startup.sh
Windows	Click Start → All Programs → IBM OmniFind Yahoo! Edition → Startup . You can also run the startup.vbs command from the INSTALL_ROOT\bin directory.

To run in console mode, so that the startup screen and administrator console do not appear, append **-console** to the **startup** command.

2. Confirm that the server is started by logging in to the administration console.

If the administration console does not open after you start the server, try the following actions:

- Linux: Run the command; INSTALL_ROOT/Shortcuts/admin.desktop
- Windows: Click **Start** → **All Programs** → **IBM OmniFind Yahoo! Edition** → **Administration Console**.
- Open a Web browser and go to the following URL: `http://localhost:port/admin`
where *port* is the port that you selected during installation. You can omit the port number if you accepted the default port (80), for example, `http://localhost/admin`.

Stopping the search engine from the administration console

You can stop the search engine from the administration console.

After shutting down the search engine, you cannot use the administration console or the search page.

To stop the search engine from the administration console:

From the **System** tab, click **Shut Down**, then click the **Shut Down** button.

Stopping the search engine from the operating system

To stop the search engine from the operating system, you log out of the administration console, then shut down the server.

Important: Ensure that you shut down the server before you turn off your computer. If you turn off your computer before the search engine shuts down completely, you might lose documents that are being added to the index. You might also corrupt the index.

To stop the search engine from the operating system:

Run the following command for your operating system:

Operating system	Action
Linux	Run the command: <code>INSTALL_ROOT/bin/shutdown.sh</code>
Windows	Click Start → All Programs → IBM OmniFind Yahoo! Edition → Shutdown . You can also run the <code>shutdown.vbs</code> command from the <code>INSTALL_ROOT\bin</code> directory.

If you stop and restart the search engine, the Web crawler automatically resumes crawling.

Starting and stopping the search engine as a Windows service

You can start and stop the search engine through the standard Windows services interface. For example, you can configure the service so that the search engine is started automatically when you restart the computer.

To set up the search engine as a Windows service, you must be logged in as a user with Windows administrator authority.

If you selected the option to set up the search engine as a service when you installed IBM OmniFind Yahoo! Edition, the service was set up without associating a user name and password. For greater security, you can specify a user name and password for the service.

The service is installed as an automatic service by default. To change the service to manual, use the Windows Services applet.

Use one of the following commands to install or remove the Windows service:

- To install the Windows service, enter the following command:

```
INSTALL_ROOT\bin\OmniFindWinService.exe -install
```

- To install the Windows service and specify a user name and password for running the service, enter the following command:
`INSTALL_ROOT\bin\OmniFindWinService.exe -install user_name password`
- To remove the Windows service, enter the following command:
`INSTALL_ROOT\bin\OmniFindWinService.exe -uninstall`
 The search engine will no longer be started and stopped through standard Windows service operations.

Starting and stopping the search engine as a Linux service

You can add the search engine to the Linux `inittab` file, which enables the search engine to run like a service and be started automatically when you restart the computer.

To set up the search engine as a Linux service, you must be logged in as the root user.

If you did not select the option to set up the search engine as a service when you installed IBM OmniFind Yahoo! Edition, you can use this procedure to set up the service at a later time. How you install and remove the service depends on whether the Linux Standard Base package is installed.

1. Copy the **ibm-omnifind** script from the `INSTALL_ROOT/bin` directory to the `/etc/init.d` directory.
2. To install or remove the search engine as a service:
 - If the Linux Standard Base package is installed on your Linux system:

Option	Description
To install the search engine as a service	Run the following command: <code>/usr/lib/lsb/install_initd /etc/init.d/ibm-omnifind</code>
To remove the service	Run the following command: <code>/usr/lib/lsb/remove_initd /etc/init.d/ibm-omnifind</code> The search engine will no longer be started and stopped through standard Linux <code>inittab</code> operations.

- If you use Red Hat Enterprise Linux and the Linux Standard Base package is not installed:

Option	Description
To install the search engine as a service	Run the following command: <code>/usr/lib/lsb/install_initd /etc/init.d/ibm-omnifind</code>
To remove the service	Run the following command: <code>/sbin/chkconfig --del ibm-omnifind</code> The search engine will no longer be started and stopped through standard Linux <code>inittab</code> operations.

Changing the administrator password

Use the administration console to change the administrator password.

To change your password:

1. Click the **System** tab, then **Change Password**.
2. Enter your current password and new password.
3. Click **Save**.

Restriction: You cannot change your user name from the administration console.

Changing a lost password and user name

If you forget your password or you want to change your user name, you can delete the `authentication.xml` file.

The `key.txt` and `authentication.xml` files should have access protection so that they cannot be read by other users. Ensure that these files are readable only by the user name under which the search engine is run, but not readable by others.

To remove the `authentication.xml` file and create another user name and password:

1. Log out of the administration console.
2. Stop the server.
3. Change to the following directory:

Operating system	Directory
Linux	INSTALL_ROOT/config
Windows	INSTALL_ROOT\config

4. Delete the `authentication.xml` file.
5. Start the search server.
6. Open the administration console and log in with a new user name and password.

Backing up and restoring a collection

You can back up a collection, including data in the index, by manually saving the collection directory. You can then restore the backed up copy of the collection. You can also back up a search page customization, and the authentication information.

On Linux, the default installation directory, `INSTALL_ROOT`, is `/IBM/OmniFindYahooEdition`.

On Windows, the default installation directory, `INSTALL_ROOT`, is `C:\Program Files\IBM\OmniFindYahooEdition`.

To back up or restore the collection:

1. Stop the server.
2. Back up or restore the collection.
 - To back up the collection, create a copy of the `INSTALL_ROOT\config\collections\Collection_Name` directory.

- To restore the collection, copy the backed up directory over the exact location where you backed up from, for example, *INSTALL_ROOT*\config\collections\Default directory.

The backup and restored directory paths must be exactly the same.

On Linux, the default installation directory (*INSTALL_ROOT*) is *opt/IBM/OmniFindYahooEdition*. On Windows, the default installation directory (*INSTALL_ROOT*) is *C:\Program Files\IBM\OmniFindYahooEdition*.

3. Back up or restore the search page customization.
 - To back up the search page customization, create a copy of the *INSTALL_ROOT*\config\customization directory.
 - To restore the search page customization, copy the backed up directory over the exact location where you backed up from, for example, *INSTALL_ROOT*\config\customization.
4. Back up or restore the administrator user name and API password.
 - To back up the administrator user name and API password, create a copy of the *INSTALL_ROOT*\config\authentication.xml file.
 - To restore the administrator user name and API password, copy the backed up file over the file that you backed up from, for example, *INSTALL_ROOT*\config\authentication.xml.
5. Start the server.

Changing the name of a collection

You can change the name of the default collection or any collection after it is created.

When the search engine is first installed, an initial collection is created named *Default*.

To change the name of a collection:

1. On the Collection Status page, click **Rename Collection**
2. Specify a name for the collection.
3. Click **Save**.

The new collection name is displayed on the collection tab.

Clearing a collection

You can clear all the contents from a collection and start over.

You can clear a collection of all its content, including the index, without deleting the collection. Clearing the collection does not delete existing crawlers, synonyms, or featured links.

To clear a collection:

1. On the Collection Status page, click **Clear Collection**.
2. On the Clear Collection page, click the **Clear Collection** button.

Administering the search engine from the command line

You can use the **configTool** command to administer the system from the command line.

For example, you can change the server ports, specify directory paths and message logging options, and generate an authentication token for communicating with the search server. You can also use the **configTool** command to view current system settings.

The **configTool** command is in the *INSTALL_ROOT/bin* directory. The command and all parameters are case-sensitive.

View help for the configuration tool

You use the **configTool help** command to display usage information for the configuration tool commands.

View help for the configuration tool

configTool help

Configure server ports

You use the **configTool configureHTTPListener** command to specify different port numbers for the search server.

Configure server ports

```
configTool configureHTTPListener -configPath path -locale locale  
-adminHTTPPort port_number -searchHTTPPort port_number
```

Parameters

-configPath path

The fully qualified path to the configuration directory, such as *INSTALL_ROOT/config*.

-locale locale

Optional. The 5-character locale code, such as *en_US*, *de_DE*, or *zh_TW*. If omitted, the server locale is used.

-adminHTTPPort port_number

The HTTP port number for the administration console. This value is set initially when the search engine is installed. If you do not configure a port for the search application, this port is the search server port.

-searchHTTPPort port_number

Optional: The HTTP port number for the search application.

Configure collection-level options

You use the **configTool configureParams** command to specify various options for administering collections.

Configure collection-level options

```
configTool configureParams -configPath path -locale locale -logPath path  
-temDirPath path -defaultDataPath path -installPath path -logLevel message_type
```

Parameters

-configPath *path*

The fully qualified path to the configuration directory, such as *INSTALL_ROOT/config*.

-locale *locale*

Optional. The 5-character locale code, such as *en_US*, *de_DE*, or *zh_TW*. If omitted, the server locale is used.

-logPath *path*

Optional. The fully qualified path to the directory to use for log files.

-temDirPath *path*

Optional. The fully qualified path to a directory to use for temporary space.

-defaultDataPath *path*

Optional. The fully qualified path to the search engine data directory.

-installPath *path*

Optional. The fully qualified path to the search engine installation directory.

-logLevel *message_type*

Optional. The types of messages to be logged. Valid values are FINE, FINER, FINEST, INFO, OFF, SEVERE and WARNING.

View the current port configuration

You use the **configTool printAdminHTTPPort** command to see the port numbers that the search server is currently configured to use.

View the current port configuration

```
configTool printAdminHTTPPort -configPath path -locale locale
```

Parameters

-configPath *path*

The fully qualified path to the configuration directory, such as *INSTALL_ROOT/config*.

-locale *locale*

Optional. The 5-character locale code, such as *en_US*, *de_DE*, or *zh_TW*. If omitted, the server locale is used.

View configuration data for all collections

You use the **configTool printAll** command to see the collection-level options that are currently configured for all of the collections in the search system.

View configuration data for all collections

```
configTool printAll -configPath path -locale locale
```

Parameters:

-configPath *path*

The fully qualified path to the configuration directory, such as *INSTALL_ROOT/config*.

-locale *locale*

Optional. The 5-character locale code, such as *en_US*, *de_DE*, or *zh_TW*. If omitted, the server locale is used.

View the value of the current token

You use the `configTool printToken` command to see the seed value that was used to create the current authentication token.

View the value of the current token

```
configTool printToken -configPath path -locale locale
```

Parameters

-configPath *path*

The fully qualified path to the configuration directory, such as `INSTALL_ROOT/config`.

-locale *locale*

Optional. The 5-character locale code, such as `en_US`, `de_DE`, or `zh_TW`. If omitted, the server locale is used.

Starting and stopping a crawler by using scripts

You can use the `manageCrawler` script to start and stop the crawlers.

The crawler management script allows you to start and stop a crawler from the command line.

Important: The crawler management script is not a scheduling tool. You must use the appropriate scheduling tool to invoke the crawler script.

For Windows

The script file is `manageCrawler.bat`.

For Linux

The script file is `manageCrawler.sh`.

The scripts are in the `INSTALL_ROOT/bin` directory.

The syntax of the crawler script is:

```
manageCrawler -c configuration_file -h dest_url -a action -i collection_name -t  
crawler_type -p api_password -o output_file
```

To start or stop a crawler by using a script:

- Run the script from the command line, specifying appropriate parameters.
- Run the script from a scheduling program, specifying appropriate parameters.

For example, the following script stops a Web crawler, and sends the resulting output to `output.txt`:

```
manageCrawler -h http://localhost:8888 -a stop -i Default -t web -p "6eKvCms=" -o output.txt
```

.

Crawler management script options

You must specify options to control the crawler management script.

All parameters for the crawler management scripts are required, except `-c`, `-o`, and `-?`.

- c *global_config*
Optional. The fully qualified path to the config.xml file.
- h *dest_url*
The destination server URL. For example, http://localhost:8080.
- a *action*
start: Starts the crawler.
stop: Stops the crawler.
- c *collection*
The name of the collection that the crawler belongs to, for example, Default.
- t *crawler_type*
file: File system crawler
web: Web crawler
- p *api_password*
The API password. You can retrieve this value from the administration console.
- o *output_file*
Optional. The output file. Directs the messages generated from requests to the specified file. If no file is specified, the script output is directed to the standard error and standard output files. For example, the output of the following script appears in the output.txt file in C:\temp:
manageCrawler -h http://localhost:8888/ -a start -i Default
-t file -p "6eKvCms=" -o c:\temp\output.txt
- ? Optional. Shows information about how to use this command.

Output file example

The output of the script `manageCrawler -h http://localhost:8888 -a stop -i Default -t web -p "6eKvCms=" -o c:\temp\output.txt` appears as follows in the output.txt file:

```
C:\OmniFindYahooEdition\bin>manageCrawler -h http://localhost:8888 -a stop -i Default -t web -p "6eKvCms="
-o c:\temp\output.txt
January 24, 2007 16:20:45.807 PST Sending stop web crawler request to: http://localhost:8888/api/admin
January 24, 2007 16:20:47.349 PST HTTP Return Code: 200
January 24, 2007 16:20:47.349 PST Status: stop web crawler request successful.
```

Changing the search server ports

You can use the `configTool httpListenerConfigTool` command to change the search server ports.

When the search engine is installed, one HTTP port is configured to handle administration and search requests. You can change the search server port or specify different ports for the administration console and search application.

To view the currently configured port values, use the `configTool printAdminHTTPPort` command.

The `configTool` command is in the `INSTALL_ROOT/bin` directory. The command and all parameters are case-sensitive.

To configure the search server port:

1. Run the `configTool configureHTTPListener` command from the command line and specify the appropriate parameters.

For complete information about the syntax for the **configTool** command, enter `configTool help`.

2. After you change a port number, restart the search engine server and the search and administration applications.

Examples

The following command configures the HTTP port for the search engine server:

```
configTool configureHTTPListener -configPath opt/IBM/OmniFindYahooEdition/  
config -adminHTTPPort 8888
```

The following command displays the current search engine server port:

```
configTool printAdminHTTPPort -configPath c:\Program Files\IBM\  
OmniFindYahooEdition\config
```

Configuring collection-level options

You can use the **configTool configureParams** command to change various collection-level options.

For example, you can configure the types of messages to be logged, the log file directory, a directory to use for temporary space, and the search server installation and data directories.

To view the options that are currently configured for all collections in the search system, use the **configTool printAll** command.

The **configTool** command is in the *INSTALL_ROOT/bin* directory. The command and all parameters are case-sensitive.

To configure collection-level options:

1. Run the **configTool configureParams** command from the command line and specify the appropriate parameters.
For complete information about the syntax for the **configTool** command, enter `configTool help`.
2. For the changes to become effective, restart the search server.

Examples

The following command specifies that the French locale is to be used, that log files are to be created in the `opt/search/log` directory, and that the highest level of message detail is to be logged:

```
configTool configureParams -configPath opt/IBM/OmniFindYahooEdition/config  
-locale fr_FR -logPath opt/search/log -logLevel FINEST
```

The following command displays the options that are currently configured for all collections:

```
configTool printAll -configPath c:\Program Files\IBM\OmniFindYahooEdition\  
config
```

Removing the search engine

You can run a program to remove IBM OmniFind Yahoo! Edition from your system.

On Windows, you can also use **Add or Remove Programs** from the Control Panel.

To remove OmniFind Yahoo! Edition from your system:

1. Go to the following directory:

Operating system	Directory
Linux	<i>INSTALL_ROOT</i> /_uninst
Windows	<i>INSTALL_ROOT</i> _uninst

2. Run the following command:

Operating system	Action
Linux	uninstaller.bin
Windows	Double-click uninstaller.exe

3. Optional: Decide whether you want to keep or delete your configuration and indexed data. You can reuse this data if you install the search engine again.

Troubleshooting crawler, search, and system problems

You can resolve typical problems with crawlers, passwords, starting the search engine, and other problems.

These troubleshooting topics can help you answer questions such as:

- Why did a crawler not crawl documents?
- Why does an index not contain the correct content?
- Why are my users not finding the documents that they need?

Monitoring the Web crawler

You can monitor Web crawler activity to ensure that the crawler is retrieving the Web documents that you want.

You can see which documents were crawled last. This information can help you determine whether the Web crawler crawled a specific site. If the Web crawler did not crawl a site, you can also see an HTTP trace for the crawler. It can show you if the Web site is preventing crawlers from crawling.

Follow these guidelines to ensure that the Web crawler runs effectively:

- Monitor the activity of the web crawler regularly. Verify that the metrics match your expectations (number of pages crawled, in particular).
- Exclude Web sites or parts of sites to remove documents that are not interesting, can “pollute” search results, or are redundant.
- Specify an e-mail address and a User-Agent string in the Edit Crawler Settings page.
- Consider notifying Web site administrators that you are planning to crawl their Web sites and of the increased site load.

You can design a formal process for your organization to provide feedback about the Web crawler activity so that Web site administrators can contact you if they do not want their site to be crawled or to be searchable.

You can also provide tips for Web site administrators to improve the quality of search for their site by ensuring that information that people are looking for is available on the Web site and not be excluded by using a robots.txt file, creating meaningful titles, and using meta tags.

To monitor Web crawler activity:

1. Ensure that the Web crawler is running.
2. On the Collection Status page, click **Web Crawler**.
3. Click **View last crawled URLs**. Review the list of URLs. If a Web site is not being crawled, check the URL. Type the URL in a browser to find out whether the address goes to the site that you expect. Or if you think a document is missing, check that a document is available by using the Check Document Status page.
4. Click **View HTTP trace**. Review the communication between the Web crawler and the target server. The HTTP trace allows you to view the communication between the Web crawler and the target Web server. In particular, you can see information that is contained in HTTP headers such as cookies, authentication,

and proxy information. For more information about HTTP headers, see the W3C Web site at <http://www.w3.org/Protocols/rfc2616/rfc2616-sec14.html>.

Web crawler cannot crawl or cannot retrieve documents

Because of the diversity of Web sites and the dynamics of Web-based applications, the Web crawler can encounter various problems.

Symptoms

You might not be aware of these problems because it can be difficult to identify normal situations as opposed to real problems.

Resolving the problem

To avoid or mitigate potential problems, be sure monitor Web crawling activity regularly.

Crawler cannot retrieve documents due to network connections

If the Web crawler does not seem to be able to retrieve any documents, you should first check the status of the network.

Symptoms

If the Web crawler crawls only a few sites and stops, your network connection might not be working properly.

Resolving the problem

Open a Web browser on the computer where the search engine is running and try to access the Web site that you want to crawl. If you cannot access the site, check the following components of your network:

Firewall (particularly software-based firewall)

Ensure that you authorize the `java.exe` process to connect to the network.

DNS servers

Ensure that the DNS server is working properly.

Network connection

Ensure that the network is working properly.

Crawler cannot crawl due to a redirected URL

The Web crawler follows redirects. However, the redirect might point to a different server that is not included in the crawl space, in which case it cannot be crawled.

Symptoms

If the Web crawler crawls only a few sites and stops, the Web crawler might have encountered a redirected URL that it cannot follow.

Resolving the problem

Open a browser and enter the URLs that you want to crawl. If the site is redirected and is not currently defined in your crawl space, copy the new Web address. Then, in the Web Crawler window in the Collection Status page, add that URL.

Crawler cannot follow links due to non-HTML links

The Web crawler follows only basic HTML links (for example, `my link`). It does not follow links that are embedded in JavaScript™, Flash applications, or Java applets.

Symptoms

If the Web crawler seems to work in general, but it cannot find a specific document or a set of documents, it might have encountered a non-HTML link. In particular, sites that use DHTML menus, which is a combination of JavaScript and a cascading style sheet (CSS) file, can cause problems.

The crawler also does not follow links in PDF, Word, or other non-HTML document types.

Resolving the problem

If you own the target Web site, you can create a site map on the Web site, which is a Web page that contains HTML links to all the main pages of the site. The Web crawler should be able to crawl the site map page.

Web crawler cannot retrieve documents due to robots.txt file, meta tags, and no-follow directives

A standard that is observed on the Internet allows Web site administrators to control access to their content by Web crawlers, which are also called robots.

Symptoms

If the Web crawler seems to work in general, but it cannot find a specific document or a set of documents, it might have encountered a robots.txt file that prohibits it from crawling that site.

Diagnosing the problem

Web site administrators provide rules in a file called robots.txt that is at the root of the Web address (for example, `http://www.example.org/robots.txt`). The Web crawler abides by those rules.

The following example shows the contents of a robots.txt file:

```
# /robots.txt file for http://webcrawler.org/  
# mail webmaster@webcrawler.org e-mail notifications
```

```
User-agent: webcrawler  
Disallow:
```

```
User-agent: othercrawler  
Disallow: /
```

```
User-agent: *  
Disallow: /temp  
Disallow: /logs
```

The first paragraph specifies that the robot called webcrawler has nothing disallowed: it can go anywhere.

The second paragraph indicates that the robot called othercrawler disallows all relative URLs that start with the forward slash (/). Because all relative URLs on a server start with a forward slash (/), the entire site cannot be crawled.

The third paragraph indicates that all other robots should not visit URLs that start with /temp or /logs. The asterisk (*) indicates that all crawlers must follow any defined Disallow rules.

Because those rules are defined based on the User-Agent string, you should set up the User-Agent string for the Web crawler.

Resolving the problem

If you need to crawl the site that disallows your crawler, contact the Web server administrator. For more information about the robots.txt file protocol, see the W3C organization Web site at <http://www.w3.org/TR/html4/appendix/notes.html#h-B.4>.

If there is no robots.txt file on the Web server, some document authors can add HTML meta elements, such as `nofollow` and `noindex`, to prevent crawlers from indexing the document or prevent them from following any links in the document. For example, the following meta tag in an HTML document prevents crawlers from indexing the page and following links in the document:

```
<meta name="robots" content="noindex,nofollow">
```

If you own the Web documents that cannot be crawled because of directives in the meta tags, you should change (or ask the HTML document owner to change) the meta tag attributes to allow crawling and indexing.

Web crawler is crawling but not retrieving documents due to crawler traps

The Web crawler can sometimes be caught in *crawler traps*, sometimes called *spider traps*, that are a series of links that cause the Web crawler to go into a loop.

Symptoms

The Web crawler might seem to be retrieving the same content or nearly the same content.

Diagnosing the problem

A typical crawler trap occurs when specific URLs are generated dynamically by simply appending a parameter, causing a recursion. This might happen, for example, with links that are used for alternative rendering (for example, printer-friendly), for comments about a specific page, or to send an e-mail note to someone about a specific page. The following pattern is an example:

```
http://www.example.org/viewevent.do?eventid=1154&printerfriendly
```

```
http://www.example.org/viewevent.do?eventid=1154&printerfriendly&printerfriendly
```

```
http://www.example.org/viewevent.do?eventid=1154&printerfriendly&printerfriendly  
&printerfriendly
```

```
http://www.example.org/viewevent.do?eventid=1154&printerfriendly&printerfriendly  
&printerfriendly&printerfriendly
```

```
. . .
```

Resolving the problem

Because the Web crawler follows every link systematically (based on the Web sites that you include or exclude), you should adjust the Web crawler rules to ensure that it avoids crawler traps.

To solve the problem, change your Web site configuration to exclude the URL pattern that causes recursion. Using the preceding example, you can exclude all Web addresses that are intended for printer-friendly rendering of the page because this content will be duplicated in the index. Exclude the following Web address pattern:

```
http://www.example.org/*printerfriendly*
```

Web crawler is disrupting a Web server's operations

To avoid interfering with the normal operations of a Web server, you can exclude sites from being crawled or increase the crawl delay.

Symptoms

A Web site administrator complains that the Web crawler is disrupting the normal operations of a Web server.

Resolving the problem

Increasing the crawl delay increases the time between crawler requests to the Web server. You can reduce the load that the crawler puts on the Web server by increasing the crawl delay.

You can also notify the Web site administrator to add instructions to the Web sites's robots.txt file to control the Web crawler.

You should specify an e-mail address so that a Web site administrator can contact you if the Web crawler disrupts the operations of the Web server.

Checking the status of a crawled document

If you want to know if a specific document or set of documents was crawled, you can check the document status in the administration console.

Ensure that a crawler has crawled the data source that contains the document or documents that you want to check.

Restriction: You cannot query the status of documents that were added by the addDocument API.

To check document status:

1. On the Collection Status page, click **Document Status**.
2. In the Document Status window, enter a specific URL, URL pattern, specific file path, or file path pattern. To check a pattern, add the asterisk (*) in the middle or end of the URL or file path. For example, enter `http://www.example.org/records`, `http://www.example.org/records*`, `file://localhost/E:/documents/personnel/members.txt`, or `file://localhost/E:/documents/personnel/*`.

Requirement: Addresses or patterns must start with the full protocol, namely `http://`, `https://`, and `file://` prefixes.

3. Click **Check Status**.

To minimize performance impact, the document status check returns only the first ten results. To ensure that the document that you want information about is included in the results, enter a more restrictive pattern.

For example, you have configured a file system crawler to crawl a local file system, `E:/documents/personnel/*`. To check that documents in that location have been crawled, enter `file://localhost/E:/documents/personnel/*` in the URL or file path text box and click **Check Status**.

Consider the following questions:

- Is a specific URL, URL pattern, file, or file directory pattern known to the crawlers?
- Has the crawler attempted to retrieve specific documents yet? See the HTTP code.
- Was the retrieval of specific documents successful? See the HTTP code.
- Was the document properly parsed and added to the index? See the parse code.

Excluded documents erroneously appear in the list of last crawled URLs

After you exclude a source from the crawl space, you might still be able to search for documents from that source.

Symptoms

You exclude a source, such as a Web site or directory, from the list of sources to be crawled. When you search the index, documents from that source appear in the search results.

Resolving the problem

After you specify which sources are to be deleted from the index, the search system must process those deletions.

Wait one day, then try searching for a document from the deleted source to ensure that the crawler is no longer crawling the documents in that source.

File crawler problems

Most problems with crawling directories are typically due to file permissions and directories that are not available.

Symptoms

The file crawler is unable to crawl some directories or files.

Resolving the problem

If you cannot crawl directories or files, ensure that the file system that you are crawling:

- Has read permission for all directories and subdirectories
- Is accessible to the IBM OmniFind Yahoo! Edition server

Users cannot find the appropriate documents

Several tools can help you improve search quality for your users. These tools are available in the Collections page in the administration console.

Symptoms

When you search the index, you see no search results or an unusually low number of results.

Diagnosing the problem

To find out whether users are finding what they need, first review the search statistics, query logs, and query statistics.

Search Statistics

On the Collection Status page, review the number of results and the response times for searches.

Query Statistics

From the Collection Status page, click **Query Statistics**. On the Query Statistics page, review the results for the most popular queries with and without results, and review the query logs.

Resolving the problem

You can improve the quality of the search by doing one or more of the following tasks:

- “Adding synonyms” on page 24
- “Adding featured links” on page 26
- “Adjusting search results ranking” on page 28

Checking for errors in the system logs

Error logs show you messages from the search engine server.

You can click messages to see more information about what caused the message and how to fix the problem.

The administration console displays only error or warning messages.

To view error logs:

1. Click the **System** tab, then **Error Log**.
2. Optional: Download the log by clicking **View Log** if you need to send it to IBM Software Support.

Formatting log files

You can use the log formatting tool for the command line to render the log files easier to read.

The **logformatter** script is in the *INSTALL_ROOT/bin* directory.

You can format the following log files: *Trace.n.log*, *System.n.log*, *QueryLog.n.log*, *migration.n.log*, where *n* represents any number starting from 0. The log files are in the default directory *INSTALL_ROOT/log*.

Although the administration console displays only error and warning messages, when you use the log formatter, informational messages also appear in the output file.

To format a log file:

Run the following command: `logformatter.{bat|sh} -f logfile [-options]`

Options:

-f *log_file*

The fully qualified path for the log file to format.

-l (**lowercase L**) *locale*

Optional. The locale to use when writing the reformatted messages, for example *en_US* or *ja_JP*. The system default locale is used if a locale is not specified.

-o *outputfile*

Optional. The fully qualified path for the output file where the reformatted log messages are to be written using the UTF-8 code page. Standard output is used if an output file is not specified.

-? Optional. Shows information about how to use this command.

-v Optional. Shows additional information (verbose) about the messages, such as showing the log record numbers.

Glossary

Use the glossary to help you understand search system concepts.

access control list (ACL)

In computer security, a list associated with an object that identifies all the subjects that can access the object and their access rights.

authentication

The process of validating the identity of a user or server.

certificate

In computer security, a digital document that binds a public key to the identity of the certificate owner, thereby enabling the certificate owner to be authenticated. A certificate is issued by a certificate authority and is digitally signed by that authority.

certificate authority

A trusted third-party organization or company that issues the digital certificates used to create digital signatures and public-private key pairs. The certificate authority guarantees the identity of the individuals who are granted the unique certificate.

character normalization

A process in which the variant forms of a character, such as capitalization and diacritical marks, are reduced to a common form.

clitic

A word that syntactically functions separately but is phonetically connected to another word. A clitic can be written as connected or separate from the word it is bound to. Common examples of clitics include the last part of a contraction in English (*wouldn't* or *you're*).

collection

A collection is an index and its associated crawlers, synonyms, featured links, and configuration settings.

crawler

A software program that retrieves documents from data sources and gathers information that can be used to create search indexes.

crawling

The activity of software programs that retrieve documents from data sources such as Web sites and file systems. The retrieved documents are then processed by the search engine, indexed, and made available for users to search. See also crawler and crawl space.

crawl space

A set of sources that match specified patterns (such as Uniform Resource Locators (URLs), database names, file system paths, domain names, and IP addresses) that a crawler reads from to retrieve items for indexing.

credential

Detailed information, acquired during authentication, that describes the user, any group associations, and other security-related identity attributes. Credentials can be used to perform a multitude of services, such as authorization, auditing, and delegation. For example, the sign-on information (user ID and password) for a user are credentials that allow the user to access an account.

data source

Any repository of data from which documents can be retrieved, such as the Web, a file system, or a database.

diacritic

A mark indicating a change in the phonetic value of a character or a combination of characters.

directory depth

The length or number of subdirectories in a file system paths. For example, the directory C:\My Company\personnel\private\records is deeper (has more subdirectories) than C:\My Company\personnel. See also URL depth.

featured links

Links with accompanying titles and descriptions that can be configured to appear at the top of the search page whenever users enter specific queries. Featured links have four parts: queries, a Web address, a title, and an optional summary.

field An area into which a particular category of data or control information is entered.

fielded search

A query that is restricted to a particular field.

file system crawler

A type of crawler that retrieves documents from directories.

free-form text

Unstructured text consisting of words or sentences.

free text search

A search in which the search term is expressed as free-form text.

GET command

An HTTP command that requests a file from a Web server.

HTML form-based authentication

An authentication method that uses forms directly in an HTML page to provide user credentials.

HTML meta tags

HTML meta tags can be added to HTML documents, for example, to give instructions to crawlers.

HTTP basic authentication

A standard authentication scheme that is specified in the HTTP protocol and is designed to control access to Web sites. HTTP basic authentication uses only a user name and a password.

HTTP proxy server

A server that acts as an intermediary for HTTP Web requests that are hosted by an application or a Web server. A proxy server acts as a surrogate for the content servers in the enterprise.

index A data structure that references data items to enable a search to find documents that contain the query terms.

IP address

A unique address for a device or logical unit on a network that uses the IP standard.

Java Database Connectivity (JDBC)

An industry standard for database-independent connectivity between the Java platform and a wide range of databases. The JDBC interface provides a call-level API for SQL-based database access.

key ring

In computer security, a file that contains public keys, private keys, trusted roots, and certificates. See also keystore file.

keystore file

A key ring that contains both public keys that are stored as signer certificates and private keys that are stored in personal certificates.

keyword match ranking

For every result, the search engine determines how closely the result matches a query. The more occurrences of the query terms that a document has and the closer these occurrences of the keywords are to each other, the higher the result is likely to appear in the list of results. Keyword match is the most important factor for returning search results.

lemma

The base form of a word. Lemmas are significant in highly inflected languages such as Czech.

lemmatization

A process that identifies the root form and different grammatical forms of a word. For example, a search for mouse also finds documents that contain the word mice, and a search for go also finds documents that contain going, gone, or went.

lexical affinity

The relationship of search words in a document that are close to each other in meaning. Lexical affinity is used to calculate the relevancy of a result.

Lightweight Directory Access Protocol (LDAP)

An open protocol that uses TCP/IP to provide access to directories that support an X.500 model and that does not incur the resource requirements of the more complex X.500 Directory Access Protocol (DAP). For example, LDAP can be used to locate people, organizations, and other resources in an Internet or intranet directory.

linguistic search

A search type that browses, retrieves, and indexes a document with terms that are reduced to their base form (for example, so that *mice* is indexed as *mouse*) or expanded with their base form (as with compound words).

masking character

A character that is used to represent optional characters at the front, middle, and end of a search term. Masking characters are normally used for finding variations of a term in an index. See also wildcard character.

metadata

Data that describes a particular piece of information and that helps that information be retrieved (by search), browsed (by category), or filtered (by interest). Metadata is often part of a taxonomy or classification scheme.

MIME type

An Internet standard for identifying the type of object that is being transferred across the Internet.

modification date ranking

Documents that are newer are more likely to be ranked higher than documents that are older.

Ranking by modification date can be important for many types of documents. However, document dates are typically unreliable for Web site documents because the last modification date that is returned by many Web servers does not reflect the true age of documents.

n-gram segmentation

A method of analysis that considers overlapping sequences of a given number of characters as a single word rather than using blank space or punctuation to delimit words as in Unicode-based white space segmentation. For example, if $n=2$, the text ABEFD is segmented into the sequence AB BE EF FD.

no-follow directive

A directive in a Web page that instruct robots (such as the Web crawler) to not follow links found in that page.

no-index directive

A directive in a Web page that instruct robots (such as the Web crawler) to not include the contents of that page in the index.

normalization

See character normalization.

normalizer

A character normalization program that scans text and reduces the variant forms of a character, such as capitalization and diacritical marks, to a common form.

parametric search

A type of search that looks for objects that contain a numeric value or attribute, such as dates, integers, or other numeric data types within a specified range.

parser A program that interprets documents that are added to the enterprise search data store. The parser extracts information from the documents and prepares them for indexing, search, and retrieval.

POST command

An HTTP command that sends information to a Web server for processing. The POST method is widely implemented in HTML files for sending forms that contain typed-in data to the server.

query log

A record of query activity. You can review query logs to find out what users are searching for, how long the queries take, and other information.

ranking

The assignment of an integer value to each document in the search results from a query. The order of the documents in the search results is based on the relevance to the query. A higher rank signifies a closer match.

Robots Exclusion Protocol

A protocol that allows Web site administrators to indicate to visiting robots (such as the Web crawler) which parts of their site should not be visited by the robot.

robots.txt file

To prevent crawlers from crawling a server, some Web site administrators

create a file on the Web server that defines an access policy for crawlers. This file, called `robots.txt`, adheres to the Robots Exclusion Protocol.

search cache

A buffer that holds the data and results of previous search requests.

search engine

A program that accepts a search request and returns a list of documents to the user.

search page

A Web page that accepts user queries and displays a list of search results.

search results

A list of documents that match the search request.

Secure Sockets Layer (SSL)

A security protocol that provides communication privacy. With SSL, client/server applications can communicate in a way that is designed to prevent eavesdropping, tampering, and message forgery.

segmentation

The division of text into distinct lexical units. Nondictionary-based processing includes white space and n-gram segmentation, while dictionary-based support includes word, sentence, and paragraph segmentation, and lemmatization.

soft error page

A type of Web page that provides information about why the requested Web page cannot be returned. For example, instead of returning a simple status code, the HTTP server can return a page that explains the status code in detail.

starting directory

The starting point for a file system crawl. If you enter a starting directory such as `C:\mydocuments`, the crawler crawls all documents in that directory, including subdirectories.

starting URL

The starting point for a Web crawl. If you enter a starting URL such as `www.example.org`, the crawler crawls all Web pages (documents) at that site that are reachable by following HTML links from the starting page.

stemming

See word stemming.

stop word

A word that is commonly used, such as *the*, *an*, or *and*, that is ignored by a search application.

stop word removal

The process of removing stop words from the query to ignore common words and return more relevant results.

synonym

Different words with the same meaning. You can improve the chances that users will find the correct document by defining synonyms that the search engine will use to expand what it searches for when one of the synonyms appears in a query.

system log

A record of computer activity as the search engine runs. You can review the system logs to check for problems on your enterprise search system.

text segmentation

See segmentation.

token The basic textual units that are indexed by enterprise search. Tokens can be the words in a language or other units of text that are appropriate for indexing.

tokenization

The process of parsing input into tokens.

tokenizer

A text segmentation program that scans text and determines if and when a series of characters can be recognized as a token.

Unicode-based white space segmentation

A method of tokenization that uses Unicode character properties to distinguish between token and separator characters.

Uniform Resource Identifier (URI)

A compact string of characters that identifies an abstract or physical resource.

Uniform Resource Locator (URL)

The unique address of an information resource that is accessible in a network such as the Internet. The URL includes the abbreviated name of the protocol used to access the information resource and the information used by the protocol to locate the information resource.

URL depth

The length or number of slashes in a Web site address. For example, the Web site address `www.example.org/personnel/private/records` is deeper (has more slashes) than the Web site address `www.example.org/personnel`.

Documents that have greater URL depth are typically less interesting than documents that have lesser depth. See also directory depth.

user agent

An application that browses the Web and leaves information about itself at the sites that it visits. In enterprise search, the Web crawler is a user agent.

User-Agent string

Identifies the Web crawler. One of its uses is in a `robots.txt` file, which can deny access based on the specific User-Agent strings.

Web crawler

A type of crawler that explores the Web by retrieving a Web document and following the links within that document.

Web links analysis ranking

A method of ranking where documents with many links pointing to them are ranked higher than documents with few links.

wildcard character

A character that is used to represent optional characters at the front, middle, or end of a search term.

word stemming

A process of linguistic normalization in which the variant forms of a word are reduced to a common form. For example, words like *connections*, *connective*, and *connected* are reduced to *connect*.

Notices and trademarks

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

IBM World Trade Asia Corporation
Licensing
2-31 Roppongi 3-chome, Minato-ku
Tokyo 106-0032, Japan

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law:

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web

sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Corporation J46A/G4
555 Bailey Avenue
San Jose, CA 95141-1003 U.S.A.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:

Oracle® Outside In Search Export, Copyright © 1992, 2007, Oracle. All rights reserved.

Oracle® Outside In HTML Export, Copyright © 1992, 2007, Oracle. All rights reserved.

Trademarks

This topic lists IBM trademarks and certain non-IBM trademarks.

See <http://www.ibm.com/legal/copytrade.shtml> for information about IBM trademarks.

The following terms are trademarks or registered trademarks of other companies:

Adobe, Acrobat, PostScript and all Adobe-based trademarks are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, other countries, or both.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Other company, product or service names might be trademarks or service marks of others.

Index

A

- administration console
 - overview 1
- administrator commands
 - configTool
 - configureHTTPListener 40, 43
 - configureParams 40, 44
 - help 40
 - printAdminHTTPPort 40, 43
 - printAll 40, 44
 - printToken 40
 - logformatter 55
 - manageCrawler 42
- administrator password
 - changing 38
- authentication file
 - backing up 38
 - restoring after backup 38

B

- backup 38

C

- collection
 - adding 4
 - backing up 38
 - changing the name 39
 - clearing 39
 - crawling file system 13
 - crawling Web sites 7
 - monitoring Web crawler 7
 - overview 3
 - restoring after backup 38
- collection configuration 44
- configTool command
 - configureHTTPListener 43
 - configureParams 44
 - printAdminHTTPPort 43
 - printAll 44
 - syntax 40
- configureHTTPListener, configTool
 - command 40, 43
- configureParams, configTool
 - command 40, 44
- crawl delay
 - configuring 12
- crawl space
 - file system 14
 - Web 8
- crawler
 - file system 3
 - Web 3
- crawler traps 50
- crawlers
 - overview 2
 - scripts 42
 - starting 42
 - stopping 42

D

- directories
 - crawling 13
 - deleting 32
 - excluding 13
- directory crawler problems
 - permissions 52
 - shared directories 52
- directory depth 29
- DNS servers 48
- document cache
 - disabling 33
 - enabling 33
- document date 29
- document language
 - overview 21
- document status
 - checking 51
- document types
 - supported 15

E

- e-mail address
 - configuring 12
- error logs
 - checking 55
 - printing 55

F

- featured links
 - adding 26
 - exporting 28
 - importing 27
 - overview 21
 - XML file 28
 - XML file format 27
- featured links XML file 27
- file crawl space
 - defining 13, 14
- file crawler rules
 - exclude 13, 14
- file formats
 - supported 15
- file system crawler
 - behavior 14
 - excluding 14
 - overview 3
 - recrawling 14
 - security 13
- file systems
 - excluding 13, 14
- firewalls 48
- FORM element
 - form action 10
 - form method 10
 - form name 10
 - form parameters 10

- form-based authentication
 - configuring 10
 - extracting FORM information 10
 - form action 10
 - form method 10
 - form name 10
 - form parameters 10
- formatter, log files 55

H

- HTTP basic authentication
 - configuring 10
- HTTP proxy servers
 - configuring 11

I

- ibm-omnifind script 37
- index
 - adding 4
 - backing up 38
 - overview 4
 - restoring after backup 38
- inittab file 37

K

- keyword match 29

L

- linguistic analysis
 - enabling n-gram segmentation 4
 - n-grams 31
- link analysis 29
- Linux services 37
- logformatter script 55
- logs
 - checking 55
 - directory location 44
 - error 55
 - formatting 55
 - message detail level 44
 - printing 55
- lost password
 - changing 38

M

- manageCrawler script
 - parameters 42
 - sample output 42
 - syntax 42
- maximum document size 44
- meta tags 49
- metadata
 - creating 22
 - editing 22

metadata (*continued*)
properties 22

N

n-grams
analysis 31
enabling n-gram segmentation 4
network connections 48
non-segmented
analysis 31

O

OmniFindWinService.exe 36

P

password
changing 38
changing lost 38
politeness
configuring 12
port configuration 43
printAdminHTTPPort, configTool
command 40, 43
printAll, configTool command 40, 44
printToken, configTool command 40
proxy servers
configuring 11

Q

query language
overview 21
query logs 23
query reports 23
query results
monitoring 23
query logs 23
query reports 23
query statistics
overview 23
query logs 23
query reports 23
quick links See featured links 26

R

ranking factors
date 29
depth 29
keywords 29
link analysis 29
redirected URLs 48
restoring collection after backup 38
restoring index after backup 38
results ranking
adjusting 28
robots 3
robots.txt file 49

S

scripts
configTool 40, 43, 44
logformatter 55
manageCrawler 42
search engine
data flow 2
overview 1
process overview 2
search engine
stopping on Linux 36
shut down 36
starting on Linux 35
starting on Windows 35
stopping from administration
console 36
stopping on Windows 36
search page
customizing 21
overview 4
search page customization
backing up 38
restoring after backup 38
search quality
featured links 21
overview 21
ranking 21
synonyms 21
search query problems
overview 53
search results
calculated 29
date 29
depth 29
keywords 29
link analysis 29
security
file system crawler 13
spiders 3
sponsored links See featured links 26
synonyms
adding 24
exporting 25
importing 25
overview 21
synonyms.xml file 25
XML file format 25
synonyms XML file 25
synonyms.xml file 25

T

troubleshooting
overview 47
redirected URLs 48
redirected Web addresses 48

U

URL depth 29
User-Agent string
configuring 12

W

Web crawl space
defining 8
Web crawler
crawl delay 12
e-mail address 12
monitoring 47
overview 3
politeness 12
settings 12
User-Agent string 12
Web crawler problems
common problems 48
crawler traps 50
disruption of Web server 51
DNS servers 48
firewalls 48
meta tags 49
network connections 48
no-follow directive 49
no-index directive 49
non-HTML links 49
recently crawled URLs 52
redirected URLs 48
redirected Web addresses 48
robots.txt file 49
troubleshooting 48
Web crawler rules
exclude 8
include 8
Web documents
checking status 51
Web sites
crawling 7
deleting 32
excluding parts of sites from
crawling 7
excluding sites 8
including 7
including sites 8
proxy servers 11
Windows services 36



Program Number: 5724-R21

Printed in USA