CourseCrawler Product User Documentation

Contents

1 Introduction **3**

2 Installation **3**

- 2.1 Permissions 3
- 2.2 Directory Structure 3
- 2.3 MySQL Database 4
- 2.4 Set Up Java-DB Interface 4
- 2.5 Compile the Program 5
- 2.6 Start the Program 5

3 Administration / Configuration **5**

- 3.1 Logging In 6
- 3.2 Logging Out 6
- 3.3 Kill (Shut Down) Server 6
- 3.4 Administrative Options 7
- 3.5 Dates Table 7
- 3.6 Change Password 7
- 3.7 Current Site List 8
  - 3.7.1 Deleting a Site 8
  - 3.7.2 Editing a Site 8
- 3.8 Add Site 8

4 Usage **10**

4.1 Ranking 10

5 Crawler **11**

6 Determining Pre/Post Conditions - Supplemental **11**

7 FAQ **12**

7.1 How do I Recover a Lost Password? 12

7.2 What are the System Requirements? 13

7.3 Do I Need Other Software to use CourseCrawler? 13

7.4 How do I Install the Software? 13

7.5 How do I Start the Software? 13

7.6 How do I Turn it Off? 14

7.7 What do I do if the Program Crashes? 14

7.8 Is Logging Supported? 14

7.9 Can Logging be Turned On/Off? 14

7.10 Can I Customize the Software? 14

7.11 How Secure is the Product? 15

7.12 What should I do if I need help? 15

8 License **15**

9 Credits **15**

1 Introduction

This document will instruct you on how to install and configure the CourseWebServer product. The process should go fairly smoothly. This software is distributed **AS IS**. As such, please do **NOT** contact the authors for assistance in setup, maintenance, modification, etc. The final release is just that - **final**.

2 Installation

2.1 Permissions

There is no reason that you would want other users to be able to read either the source or bytecode files. Your Database information will be stored in plain text inside the .java files, and can be retrieved from the .class files with relative ease. As such, it is strongly recommended that you change the permissions of this directory and all subfiles so that only the owner has access. To do this, navigate to the directory that the files are in and type **chmod 600 *** and hit ENTER.

2.2 Directory Structure

In order for the program to run successfully, you will have to create a subdirectory called "logs" by typing **mkdir logs** and hitting ENTER in the directory that the program will be executed in. This is necessary even if you do not intend to use the logging feature.

2.3 MySQL Database

In order to use this product, you must have MySQL installed and running, with a database for this product to use. This product was written using MySQL Version 4.0.17, although it should run on other versions. It is strongly recommended that you allocate a separate database for this product, however it may be possible to use an existing database.

If you have any questions regarding installation, configuration, etc. of MySQL please consult *http://www.mysql.com/* for assistance.

Please make note of your database server (may be localhost), database name, username and password when creating the database.

Once you have created the database, you must configure it for the product. Navigate to the directory that the product is in. Log into MySQL (The command should be **mysql -h db_server -u db_username db_name -p**) replacing db_* with the appropriate value. Enter the password when prompted and hit ENTER. Once successfully logged in, type **\. CourseCrawlerSetup.sql** and hit ENTER. The database is now configured. Exit MySQL by typing **exit** and hitting ENTER.

2.4 Set Up Java-DB Interface

Open the file **cc_db.java** in your favorite text editor. Near the top of the file there will be a section of code as follows:

```
// Database server DNS address
String dbHost = "your_db_server";
// Database Name
String dbName = "your_db_name";
```

```
// Database Username
String username = "your_db_username";
// Database Password
String password = "your_db_password";
```

Replace the text **your_db_server** with your Database Server's DNS address, **your_db_name** with your Database Name, **your_db_username** with your Database Username and **your_db_password** with your Database Password. Save the file and exit.

2.5 Compile the Program

This program was developed using Java 1.4.2 and although it should be compatible with future versions, compatibility cannot be guaranteed. Compile the program by typing **make** and hitting ENTER. (This requires that you have the make utility installed.) This will run the included makefile, compiling the program.

2.6 Start the Program

Start the program by typing **nohup java CourseWebServer port_number** and hitting ENTER, where port_number is replaced by the port that you desire the program to run on. The default for web (http) content is port 80.

3 Administration / Configuration

Please note that only one administration session can be active at a time. If a person logs in while another person is in a session, the first session will be invalidated.

The product is designed for minimal administration by one person. Parallel administration sessions are not supported. Session validity is based on IP address and time of last login as well as some internal parameters. As such it is possible to maintain a session from the same computer after a reboot, etc. although it is strongly suggested that you log out when you are finished every time.

3.1 Logging In

Open your favorite web browser and navigate to **http://host:port/admin** where host is replaced by the DNS name or IP of the computer that the program is running on, and port is the port chosen above. As web browsers will default to port 80, if you are running the program on that port, you can use this form: **http://host/admin** for administration. The default password is **admin**. Log in, and you can begin to configure the product to your specifications.

3.2 Logging Out

To log out click the **Log Out** link found near the top or bottom of the administration page. This will close your session.

3.3 Kill (Shut Down) Server

To turn off the server, click the **Kill Server** button. Please note that there is no way to restart the server from the web - it will have to be done in the terminal on the machine upon which it is installed.

3.4 Administrative Options

Here you can change server settings and see crawl/parse status. The Logging checkbox determines whether the server generates logs. Logs are written as plain-text to the directory logs/ and are stored by date. The Login Timeout box allows you to change how long administration sessions are valid after the user logs in. The Crawl Frequency box allows you to select how often the crawler and parser run. The two status boxes show the current status of the Crawler and Parser.

3.5 Dates Table

Here you are shown the timestamp for the current administration session (when it started) and the dates for the last crawl and last clean. If these values do not exist in the server, a horizontal rule is displayed. Here you can chose to manually clean the database. Cleaning the database will remove any entries not found within the number of days entered of the last crawl. It is suggested that you chose a value at least as large as the interval between crawls.

3.6 Change Password

The first that that you should do after logging in for the first time is to change the password. The default password of **admin** is hardly the most secure. To do this, click the **Change Password** button. You will be prompted to enter the current password, then to type and confirm your new password. Be careful to use a password that you can remember, as passwords are very difficult to retrieve. (See section 7.1) Please note that passwords are restricted to alphanumeric characters only.

3.7 Current Site List

This displays a summary of the sites that are being crawled (Name and URL). You can permanently remove a site from the list by clicking on the **Delete** button (you will receive a confirmation dialog) and you can view or edit a site's information by clicking the **Edit** link.

3.7.1 Deleting a Site

To delete a site, click the corresponding **Delete** button in the Current Site List.

3.7.2 Editing a Site

Please see section 3.8, Add Site, for more information. Editing a site is almost identical to adding one except that the values will start filled in rather than blank. For more details, please see section 3.8.

3.8 Add Site

Here you will define the parameters for parsing a given site. It must be noted that in the parse terms, the contents of double-quotation marks inside of HTML tags is treated as a wildcard. For instance, if you have a page that placed <**a name="some-varying-string"**><**b**> you would enter <**a name=""**><**b**> as the tag to be searched for. Please note that the crawler treats any combination of whitespace (newlines, tabs, spaces, etc.) as a *single* space character. Please enter your parse terms accordingly. Further documentation regarding how to determine what to use for pre/post conditions can be found in Section 6.

The *Site Name* is the name to be displayed in the Source column of the search results page.

Site URL is the base URL for the site. Please note that in order for the crawler to correctly gather all information, this page must link directly or indirectly to all pages that you wish to parse in this particular site entry.

The *Allow Higher* checkbox (unchecked by default) tells the crawler whether to traverse links to higher levels or subdomains of the same domain. Please note that the crawler will not visit pages outside of the originating domain.

The *Pre-Word HTML* defines what the crawler should search for while parsing pages in the site to find words to be defined. To determine the appropriate value you will have to view the page's source HTML to ascertain the largest consistent pattern leading up to words.

The *Pose-Word HTML* is the only text value that can be left blank. This is for the case where the markup between the word and definition are uniformly the same. As the name implies this is the markup immediately following a word in the page's source HTML.

Much like the last two entries, *Pre-Definition HTML* is the markup found immediately before the definition corresponding to the word as determined by Pre-Word and Post-Word HTML. This value cannot be empty.

Post-Definition HTML contains the markup signifying the end of the definition. Be especially careful here as if the markup is found inside of the definition, the definition will be cut short.

4 Usage

To use the program to search, open your web browser and navigate to **http://host** where host is the DNS name or IP of the machine that the program is running on. If you wish to limit results to exactly what you entered, uncheck the **Include Similar** checkbox. Enter the term that you wish to search for and click the search button.

The results will be displayed. Please help the product distinguish between helpful and unhelpful definitions by clicking on the appropriate button to the right. If no desirable results are displayed, click the **Search Google** button to have Google search for the same query.

There may be no results if the crawler has not finished its first run and that there might be a delay if searching for similar results of a common substring. Please be patient and if need be wait an hour or two and try again.

Please note that all queries must have a length greater than or equal to two characters, and that the % and “ characters are stripped from queries due to implementation specifications.

4.1 Ranking

Results are sorted based on an internal rank. That rank can be adjusted by clicking the **Was this useful?** buttons found to the right of the results. Over a period of time, the better entries will gravitate to the top of the list and vice-versa.

5 Crawler

The crawler is the invisible part of the program that takes the information in the site list and retrieves data from the specified sites. For more information about the crawler please see the technical documentation that accompanies this document.

6 Determining Pre/Post Conditions - Supplemental

It is understandable that it might be quite difficult to determine the proper parse-terms. Perhaps a few examples would be helpful:

Example 1

```
...
   <!--RANDOM COMMENT-->
   <p class=``text''>TERM - DEFINITION</p>
   </body>
...
```

The Pre-Word would be **‘–> <p class = “text”>’**, the Post-Word would be blank, the Pre-Definition would be **‘ - ’** and the Post-Definition would be **‘</p > </body >’**. Notice that we include as much as we can that will not change from page to page across the site so as to avoid false-positives. Also notice that all (multiple) whitespace has been reduced to a single space.

Example 2

```
...
    <table><tr>
         <td>RANDOM CONTENT</td></tr>
     <tr><td><a href=``RANDOM TARGET>TERM</a></td>
         <td>AD CONTENT</td></tr>
     <tr><td colspan=``2''>DEFINITION</td></tr>
    </table>
...
```

Here the Pre-Word would be **‘</td></tr> <tr><td><a href“”>’** (the parser ignores the contents of double-quotationmarks), the Post-Word would be **‘</a ></td> <td>’**, the Pre-Definition would be **‘<tr><td colspan=“2”>’** and the Post-Definition would be **‘</td></tr>’**

Again, the key point is to examine the source of multiple pages on the same site, and to determine the largest chunks of uniform markup in the appropriate places. These will translate directly into the Pre/Post Parse-Terms to use.

7 FAQ

7.1 How do I Recover a Lost Password?

The only way to recover a lost password is directly through MySQL. To do this log into MySQL by executing the command **mysql -h db_server -u db_username db_name -p**, replacing the variables with the appropriate values. Once in, type the command **SELECT admin_password FROM ccrawler_admin;** and hit ENTER and the password will be displayed. Make note of the password and log out by typing **exit** and hitting ENTER.

7.2 What are the System Requirements?

This program was written to run on a Unix/Linux Operating System. The webserver (machine running this program) will not do too much processing under a moderate load, however the MySQL server will experience a significant load. No hardware tests have been performed, but more than 500Mhz and at least 256MB or RAM are recommended for both machines. Also note that the machine running the webserver must be accessible via http to the machines that will be using the service and must have a connection the the internet for the crawler to function properly. It will also have to have network access to the MySQL server.

7.3 Do I Need Other Software to use CourseCrawler?

Yes. This product requires MySQL. It was built on version 4.0.17. For more information see Section 2.3 or the MySQL website at *http://www.mysql.com* Also, you must have MYSQL Connector/J installed. Documentation is available here: *http://dev.mysql.com/doc/connector/j/en/index.html*

7.4 How do I Install the Software?

Please see Section 2.

7.5 How do I Start the Software?

Please see Section 2.6.

7.6 How do I Turn it Off?

The easy way to run off the program is through the Administration Panel. Please reference Section 3.3. It is also possible to kill the process manually if you know the process ID by typing **kill process_id** into the command prompt and hitting ENTER. To determined the process ID, type **ps -aef — grep CourseWebServer** into the command prompt and hit ENTER.

7.7 What do I do if the Program Crashes?

The data for the program is stored in a MySQL Database and should not be affected by a program crash. Simply start the program again by following the instructions found in Section 2.6.

7.8 Is Logging Supported?

Yes. For more specifics please see Section 3.4.

7.9 Can Logging be Turned On/Off?

Logging can be turned on and off in the Administrative Panel. Please see Section 3.4.

7.10 Can I Customize the Software?

In short, yes. For more specifics please see Section 8.

7.11 How Secure is the Product?

This software uses a proprietary web server written in Java (1.4.2). As it is very low-powered the security risks are relatively low. Administrative session data is all passed in plaintext and so might be open to being intercepted. Administrative sessions are initialized based on password alone. Session validation is checked based on the IP address and login time of the current session. If there is no current session (ie. previous administrator logged out) the password must be supplied.

7.12 What should I do if I need help?

First, read this document thoroughly. If your question is not answered, please read the included technical documentation. If your question is still not answered, you will have to consult the supplied source code.

8 License

9 Credits

This program was written by:

Matthew Berntsen, Bucknell University Class of 2005 (*http://mattberntsen.net*)
Evan Kaiser, Bucknell University Class of 2005 (*http://www.evankaiser.com*)

Don Frehulfer, Bucknell University Class of 2005 (*http://typical_dfrehulf.blogspot.com*)

This program was commissioned by:

Professor Peter Drexel, Plymouth State University (*http://turing.cs.plymouth.edu/ ae1t*)

Supervision and Advice by:

Professor Xiannong Meng, Bucknell University (*http://www.eg.bucknell.edu/ xmeng*)