

# Sophisticated Mining with Oracle Text

Ultra Search adds database and internet content to the same index.

I often get asked if Oracle Text can replace full-blown document management systems, but I don't see Oracle Text as a competitor to these systems. When combined with other Oracle database features such as Internet File System (Oracle9iFS) and the new XML datatype (XMLType), Oracle Text can be a core component of document management systems. Some tools provided in the Oracle *interMedia* services, such as Annotator and Clipboard, provide functionality to maintain and organize rich data.

One of the key advantages of Oracle Text is that it allows you to implement full-text search functionality against data in your Oracle database without the introduction of additional companion/third-party tools. Its tight integration with the database has allowed me to help clients bring that functionality to their applications in ways that they manage and customize in-house.

Oracle Text has always allowed full-text indexing from multiple data sources, including those that are available via internet or intranet. A missing piece in Oracle Text had always been the ability to combine data easily from these multiple source types into one searchable index. With the introduction of Oracle Ultra Search in Oracle9i Database, however, there is now a tool that allows developers to perform that task easily. Furthermore, Ultra Search provides a configurable crawler that can, for example, combine your database table data with strategic data from the Web, e-mail, and files.

## webLOCATOR

Ultra Search on Oracle Technology Network  
[otn.oracle.com/products/ultrasearch/](http://otn.oracle.com/products/ultrasearch/)  
 Ultra Search Product Documentation  
[otn.oracle.com/docs/products/ultrasearch/](http://otn.oracle.com/docs/products/ultrasearch/)

This is an important advance in Oracle data indexing and searching because it strongly connects with some of the best goals of knowledge management. Imagine, for example, that you're a product designer, and your design must meet the needs of fashion, functionality, and ease of production. Imagine also that your company has created a way for the people on the assembly line to place into a database their ideas of how to streamline production. Finally, include in this imaginary environment the use of Ultra Search, which brings together the following in one search: your notes, information from the Web (and therefore, information about similar products), assembly line information, and possibly information in e-mails from the last round of your design for this model. This knowledge sharing, supported by Ultra Search, will lead you to an innovative and informed design based on information from the entire value chain.

## TECHNICAL OVERVIEW

First, you need to know that there are different Ultra Search versions lurking out there. Version 1.0.3 of Ultra Search is packaged with Release 9.0 of the database, while version 9.0.2 of Ultra Search is provided with Oracle9iAS. This article describes and works with Ultra Search 1.0.3.

**Ultra Search components.** Ultra Search is composed of two component sets. The server component includes:

- Data dictionary and PL/SQL packages
- Crawler and Java classes
- Product libraries
- Remote crawler

Even though these are all grouped in the server component, they really represent different tiers of configuration based on where they reside. The data dictionary (tables, triggers, etc.) and PL/SQL pack-



ages reside in the database—in the WKSYS schema. The crawler and Java classes and product libraries reside outside the database but in the same ORACLE\_HOME as the database. The remote crawler (a configuration option used to locate the crawling mechanism to a machine other than your database server) resides on a separate machine from the database.

The middle-tier Ultra Search component includes:

- Administration tool
- Java query API
- JavaServer Page query application

All of the middle-tier components reside outside the database but in the same ORACLE\_HOME as the database.

The Administration tool (a middle-tier component), shown in Figure 1, is the primary tool used to configure all of these components. You create Ultra Search instances through the Administration tool, but you must do some additional configuration with a text editor. An Ultra Search instance contains configuration information for the instance (such as a

schedule for the crawler), but more important, it holds the entire indexed and searchable result of the crawler's work. When you create an Ultra Search instance, you assign a database account to hold the tables and indexes that support that instance. So, you can have multiple Ultra Search instances in a single database. You may want to make use of multiple instances to logically divide searchable data, for example, by department or subject area, or you may wish to store all of your searchable information in one Ultra Search instance.

A simple Ultra Search configuration takes the following shape. First, install all of the Ultra Search components on the same server as the database, using the Oracle HTTP server or Oracle9iAS.

Then you open the Administration tool through a browser and configure an instance. Next, assign a schedule to the crawler for that instance along with information about what Web pages, IMAP (e-mail) servers, files, and tables are to be included in the crawl. At the scheduled time, the crawler finds information to index and feeds it to the instance that then uses the Oracle Text engine to pull it all into one searchable index.

#### ADMINISTERING ULTRA SEARCH INSTANCES

Use the Administration tool to create and configure Ultra Search instances. The tool is available for use after you've performed the installation steps provided in the documentation. After you create an

Ultra Search instance, you'll be required to specify it when first entering the Administration tool. All Ultra Search Administration tool work from that point until you change the instance name is performed against that instance. You can see in Figure 1 that the instance in this example is named WK01OUS, and it has been created in an account named WK01.

#### CONFIGURING THE ULTRA SEARCH CRAWLER

Figure 2 shows the Ultra Search crawler configuration page, which is used to locate and pull information for the Oracle Text index used in an Ultra Search instance. You'll want to review all of the parameters on this screen, but most of the default values are OK to start with. I changed a few for this example. I've set Crawling depth to 1. I've set the value for Temporary directory (the location of files used and then removed in the crawling process) to c:\oracle\admin\orcl\ous\wk01\temp. There is no required or recommended location for the crawler's temporary files, but I've chosen to put them under the Oracle server instance-specific directory instead of the ORACLE\_HOME\ultrashow directory, since the files are Oracle server instance- and Ultra Search instance-specific, not Oracle-installation (ORACLE\_HOME) specific. I've set the Max size (megabytes) field to 50. You can set a small size for this parameter, but if the information set being indexed is large, a small size value will cause Oracle Text index fragmentation to occur, leading to slower searching performance and larger storage space requirements in the database. The final value that I changed is the logfile directory. The logfile directory is the location of the files that the crawler creates with information about its crawling process. It creates a new logfile every time it crawls, and it does not remove previous logfiles, so you should review the logfile directory regularly to make sure its contents are not consuming too much space. I've set my logfile directory name using a similar model to the one I used with the temporary directory: c:\oracle\admin\orcl\ous\wk01\log.

figureONE

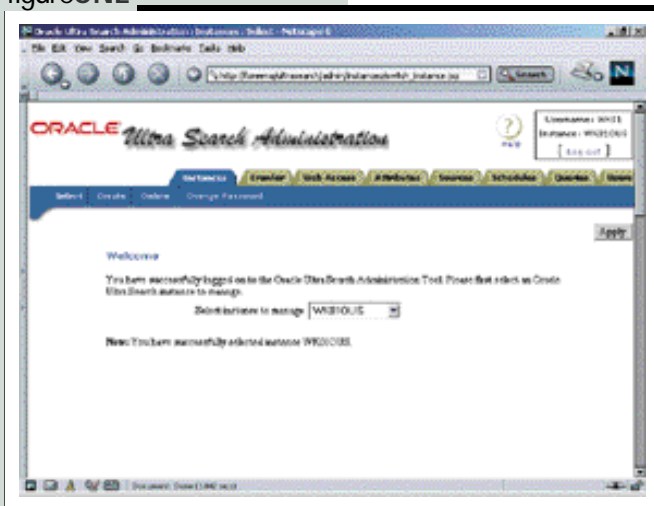


FIGURE 1: Selecting the Ultra Search instance name.

figureTWO

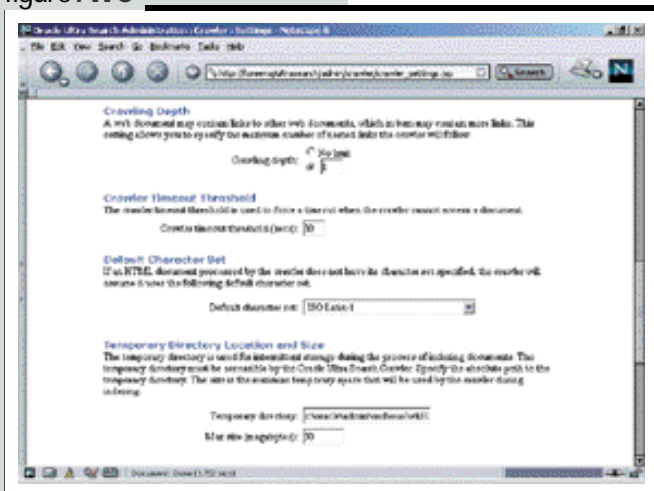


FIGURE 2: Setting the Ultra Search crawler configuration.

#### SPECIFYING SEED URLS

Seed URLs are the starting locations that the crawler uses when Web crawl-

figureTHREE



FIGURE 3:  
Defining seed URLs.

figureFOUR

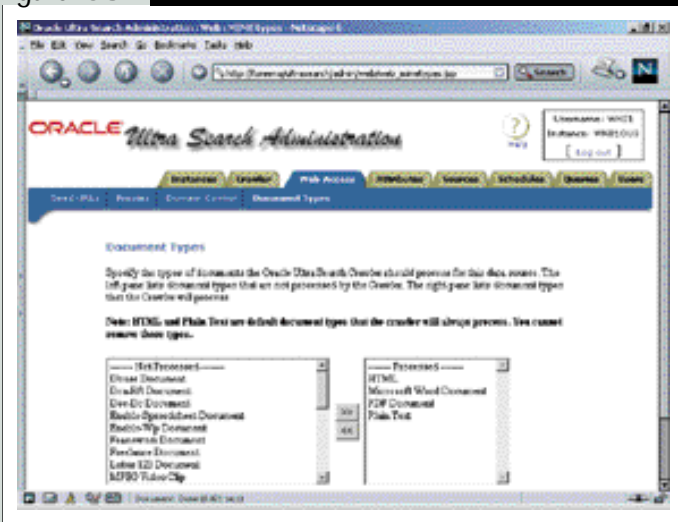


FIGURE 4:  
Choosing document types to include in the search.

figureFIVE

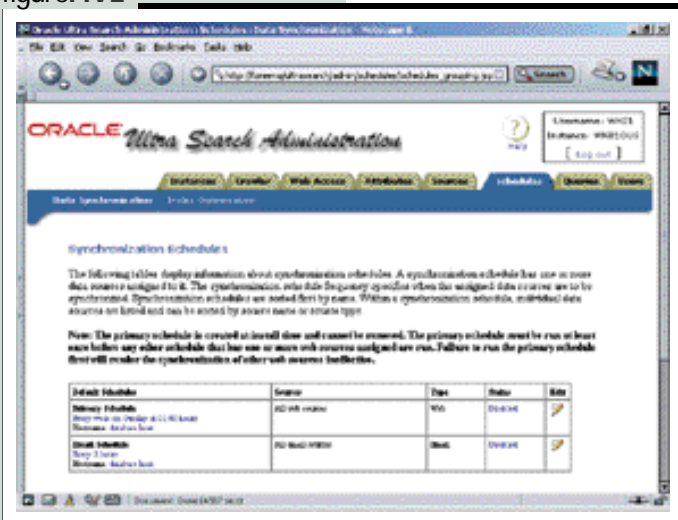


FIGURE 5:  
Setting the Ultra Search Scheduler.

ing. They are configured under the Web Access tab in the Administration tool (sub-tab Seed URLs). In Figure 3, you can see that I've defined one seed URL: <http://www.foodtv.com>. By default, the crawler will consider only the Plain Text and HTML type documents it finds. You can configure it to consider other types of documents by choosing the sub-tab Document Types. As shown in Figure 4, I've configured the crawler to consider Microsoft Word documents and PDF documents in addition to Plain Text and HTML.

### USING THE ULTRA SEARCH SCHEDULER

The scheduler is responsible for starting the crawler at defined times. There are several schedules for each Ultra Search instance. By default, the schedules are disabled, as shown in Figure 5. To start the primary crawler, click on the Disabled link; on the next page, click on the Enable Schedule button. Check the Next Attempt At date and time. If you want to run the crawler now instead of waiting for that time, click the Execute Immediately button. The crawler may take a while to return. You can click the Refresh Status button to see the current status of this crawler session. To stop the crawler session, click the Stop Schedule button. Whatever information has been indexed to that point will remain in the Ultra Search instance. Once the crawler has finished, you may wish to click the Disable button so the crawler doesn't continue to start itself.

### QUERYING IN ULTRA SEARCH

Applications making use of Ultra Search use the Ultra Search Query API. The API is written in Java and is compatible with a large spectrum of Web application servers that support Java-based technology.

Ultra Search includes a sample query application that allows you to search the information just indexed. The application is a nice way to review the result of the crawler's work and can be used as a template for developing your own application. The application comes with the username and password WK\_TEST /WK\_TEST hard-coded. So, you'll receive the error message, "java.sql

figureSIX

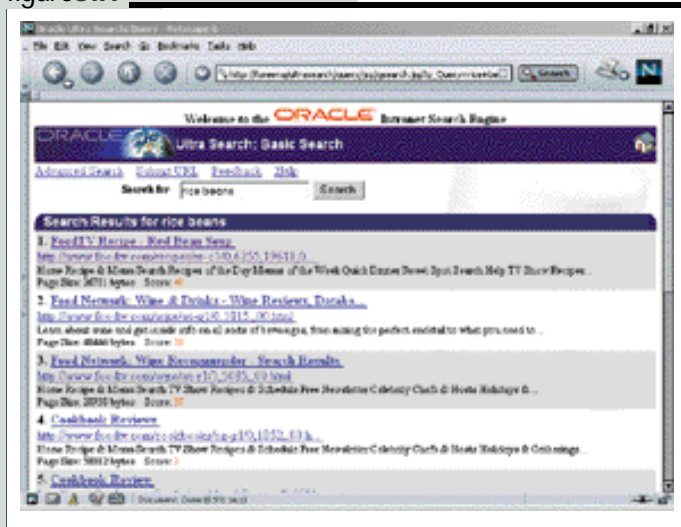


FIGURE 6: Searching an application with Ultra Search.

.SQLException: ORA-01017: invalid username/password; logon denied” unless you’ve created your Ultra Search instance in the account WK\_TEST.

To configure the query application to use your username and password (in this example, WK01), you need to edit the application’s .jsp files. The .jsp files are located in ORACLE\_HOME\ultrasearch\sample\jsp. Back up that directory before making any changes, in case you need to restore the original values. Then, open each file in ORACLE\_HOME\ultrasearch\sample\jsp and replace the lines

```
qt.setUser("wk_test");
qt.setPassword("wk_test");
```

with the username and password of the Ultra Search instance you want to query. In this example, it’s:

## getMORE

### Ultra Search Documentation

Ultra Search documentation is located in several places. A copy is installed with the online Oracle documentation (when the database is installed) in the directory ORACLE\_HOME\doc\ultra.901. Another copy is installed in the Ultra Search directory ORACLE\_HOME\ultrasearch\doc\help. There is a readme file for Ultra Search at ORACLE\_HOME\ultrasearch\doc\README.html. Check the README.html file for last-minute release-specific information. In addition, you can find the documentation and other Ultra Search information on the Oracle Technology Network (OTN) at [otn.oracle.com/products/ultrasearch](http://otn.oracle.com/products/ultrasearch). Some of the local documents point to documents on the OTN site.

```
qt.setUser("wk01");
qt.setPassword("wk01");
```

The URL for the sample query application is <http://<HOSTNAME>/ultrasearch/query/jsp/gsearch.jsp>. Figure 6 shows the result of an Ultra Search search in the sample application. You can see that there are links to matching pages sorted in descending order of the Oracle Text CONTEXT score (for example, Score: 40, Score: 38, and so on).

### TIPS

If you’re using the Oracle HTTP server, the Ultra Search installation is mostly taken care of for you. Still, you should review all of the documentation to make sure everything is set correctly. Here are a couple of tips (some of them undocumented) to keep in mind when installing Ultra Search.

1. Set SHARED\_POOL\_SIZE to at least 50MB and LARGE\_POOL\_SIZE to at least 5MB.
2. Adjust the database connection string so that the Administration tool knows in which database you’re going to keep Ultra Search instances. In ORACLE\_HOME\ultrasearch\jsp\admin\config\database.properties, change the line `connection.url=jdbc:oracle:thin:@localhost:1521:<oracle_sid>` to the correct information for your Oracle database. In this example,

it’s: `connection.url=jdbc:oracle:thin:@localhost:1521:orcl`.

3. Start your work with Ultra Search by setting the crawling depth to be shallow (you do this in the Administration tool under the Crawler tab). Once you get more experience, you can deepen the crawling depth. If you start with a shallow depth, you won’t wait as long for your experimental crawls to be completed and you’ll be utilizing fewer local and external network resources.
4. Use the buttons, tabs, and links to navigate the Administration tool instead of the Back and Forward buttons of your browser.

### ROOM FOR IMPROVEMENT

Ultra Search 1.0.3 adds great capabilities to Oracle Database functionality, but there are some things you should be aware of when considering using the tool in development:

- There’s no way to specify a login string for URLs you want to crawl that require them. Basic authentication is planned for a future version.
- Only one Oracle System Identifier (SID) can be specified within the database.properties file. That means that you can create Ultra Search instances only in one Oracle database, unless you’re willing to edit the database.properties file each time you want to switch to a different database.
- There’s no easy or published way to adjust storage parameters for the database objects in an Ultra Search instance.

### CONCLUSION

Ultra Search provides a great step forward in text indexing within an Oracle database. You can use this enhanced indexing to support organizational knowledge sharing, leading to better-informed decision-making and innovation. ■

---

*Douglas Scherer (dscherer@coreparadigm.com) is president of Core Paradigm, a management consulting practice in New York. He is a frequent presenter at conferences, an author of books and articles on management and Oracle technology, and a member of the adjunct instructional faculty at Columbia University’s Executive Information Technology Management program.*