Oracle Text Features Overview

INTRODUCTION
Oracle Text adds powerful text search and intelligent text management to the Oracle9i - the world's most popular database for e-business.

Oracle Text indexes any document or textual content to add fast, accurate retrieval of information to Internet content management applications, eBusiness catalogs, news services, job postings, etc.

Oracle Text:
- Is fully integrated with Oracle9i.
- Offers premier text search quality.
- Contains several advanced features for text management, document services, and XML.
- Has the best internationalization set of features for multilingual text search applications.

Oracle Text management capabilities allow text searches to be combined with regular database searches in a single SQL statement. The ability to find documents based on their textual content, metadata or attributes, makes the Oracle9i the single point of integration for all data management.

INTEGRATION
The basic advantage Oracle Text brings to applications involving text is integration. Integration with the database gives security, performance, scalability, and a virtually universal application development platform. Integration of text and structured queries gives a simple and flexible architecture. Integration of all text search capabilities in a single SQL function enables a flexible and powerful search system capable of finding the desired information within any collection of documents.

Oracle Text’s SQL API makes it simple and intuitive for application developers and DBA’s to create and maintain Text indexes and execute Text searches.

- **Create a Text Index**
  To create an Oracle Text index, simply use the `CREATE INDEX` command. For example:
  ```sql
  CREATE INDEX description_idx ON product_information(product_description) INDEXTYPE IS CTXSYS.CONTEXT;
  ```

- **Execute Text Searches**
  Now you can search for information about all products where “monitor” is near “high resolution” in the product description:
  ```sql
  SELECT score(1), product_id, product_name
  FROM product_information
  WHERE CONTAINS (product_description, 'monitor NEAR high resolution', 1) > 0
  ORDER BY score(1) DESC ;
  ```

March 2002
<table>
<thead>
<tr>
<th>SCORE(1)</th>
<th>PRODUCT_ID</th>
<th>PRODUCT_NAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
<td>3331</td>
<td>Monitor 21/HR</td>
</tr>
<tr>
<td>27</td>
<td>3060</td>
<td>Monitor 17/HR</td>
</tr>
<tr>
<td>14</td>
<td>1726</td>
<td>LCD Monitor 11/PM</td>
</tr>
<tr>
<td>14</td>
<td>3054</td>
<td>Plasma Monitor 10/XGA</td>
</tr>
<tr>
<td>14</td>
<td>2252</td>
<td>Monitor 21/HR/M</td>
</tr>
<tr>
<td>14</td>
<td>2243</td>
<td>Monitor 17/HR/F</td>
</tr>
</tbody>
</table>

- **Maintain a Text index**
  As with any full-text retrieval index, changes to the underlying data are applied to the index in batch. To synchronize the text index “index_one” using 2 megabytes of memory, use:
  ```sql
  exec ctx_ddl.sync_index('index_one','2M');
  ```
  The synchronize can be done periodically (every 5 minutes, every hour, etc.) or immediately. Note that the catalog indextype - designed specifically for the short pieces of text typically found in eBusiness catalogs – needs no synchronization. The catalog index is always up-to-date.

- **Simple Management**
  Oracle Text management can be done using the Oracle Enterprise Manager. Built upon open Internet standards such as Java™, CORBA, and IIOP, the Oracle Enterprise Manager product provide the first management framework designed to support Internet computing. All Oracle Enterprise Manager applications can be accessed from anywhere using a standard web browser. A reliable and scalable multi-administrator repository leverages the administrative staff by providing cooperative management. Using the Oracle Enterprise Manager product family, administrators and IT managers can increase productivity, deliver better services, and reduce the total cost of information systems.

- **Smart defaults**
  Oracle Text uses system defaults based on column type, database language settings, and database character-set settings to simplify index building.

Oracle Text is completely integrated with the Oracle database making it inherently fast and scalable. The Text index is in the database – no separate data or index storage required. Text queries are executed in the Oracle process – no separate query processing. The Oracle optimizer is aware of the Text index, so it can choose the best execution plan for any query – gives the best performance for ad hoc queries involving Text and structured criteria. Additional advantages are:

- A Text index can span many Text columns, giving the best performance for Text queries across more than one column.
- Oracle Text has enhanced performance for operations that are common in Text searching like “count hits”.
- Oracle Text utilizes scalability features such as replication.
- Oracle Text supports Partitioning.
- Oracle Text supports Parallelism.
Indexing

There are four Text index types to cover all text search needs.

- **Standard index type** (*context*) for traditional full-text retrieval over documents and web pages. This index type provides a rich set of text search capabilities for finding the content you need, without returning pages of spurious results.

- **Catalog index type** (*ctxcat*) – the first text index designed specifically for eBusiness catalogs. This catalog index provides flexible searching and sorting at web-speed.

- **Classification index type** (*ctxrule*) for building classification or routing applications. This index is created on a table of queries, where the queries define the classification or routing criteria.

- **XPath index type** (*ctxxpath*) for improving performance on XPath searches on XML documents.

Oracle Text also provides substring and prefix indexes. Substring indexing improves performance for left-truncated or double-truncated wildcard queries. Prefix indexing improves performance for right truncated wildcard queries.

Oracle Text offers synchronization options for maintaining the text index.

Storage options

Oracle Text offers the following text storage location options:

- **In the database**: Text to be indexed may be stored in the database in any character column or LOB (Large Object), which stores character or binary data up to 4GB in size.

- **In the file system**: Text data may be stored on any file system that is accessible from the database server machine.

- **On the web**: you can create a Text index on any text that can be referenced by a URL, on the Internet or Intranet.

- **User-defined**: you can create a Text index using the output of your own PL/SQL procedure – this gives you total control over what is indexed and retrievable.

Search quality

Oracle Text has a wide range of operators and function to express any of the following:

**Exact match** - Search for a document that contains a specific word such as “Java”

**Word positioning**

- Search for a document that contains a particular phrase, such as “Java Developers Kit”

- Search for a document that contains words near to each other. “Java NEAR developer” will order results according to how close “Java” is to “developer”. Or you can search for all documents where “developer” is up to N words before/after/away from java.

- Search for a document that contains words in the same sentence or paragraph, such as “java” in the same sentence as “developer”.

**Inexact match**

- Search for a document that contains words that sound like this one.
Oracle Text: Features Overview

- Search for a document that contains words that look like this one – this search “forgives” common spelling, typing and OCR errors in the search term and the document.
- Search for a document that contains words with the same linguistic root as this one – find “camera” as well as “cameras”, “buys, buying, bought” as well as “buy”.
- Search for a document that contains words that contain a string – for example, search for a document that contains a word that starts with “govern”, or a word that has “oxy” in it somewhere. Oracle Text includes options to build extra indexing information to speed up such prefix and substring searching.
- Search for a document that contains words or phrases with the same or similar meaning – for example, search for a document that contains any word or phrase that means the same as “election” (synonym search). Oracle Text supports all thesaurus operations specified by ISO-2788 (synonyms, broader terms, narrower terms, related terms, translation terms, preferred terms, top term, etc.). Oracle Text comes with a sample thesaurus plus the tools to build, manage, and import your own thesauri.

Intelligent match

- Search for a document that is about something - Oracle Text’s unique knowledge base describes the themes of each document. You can then search by theme, and find documents that are about your search term. For example, a document may contain the word “car” without being about cars: similarly, a document may be about cars – mentioning coupes, convertible, sedan – without containing the word “car”.
- Ignore noise-words such as “the”, “of”. A default list of noise words (a stop list) is supplied for each major European language. You may alter supplied stop lists, or define your own. You may also choose to ignore a class of characters.
- Handle non-alphanumeric characters – for example, you may choose to count “-“ as a part of a word for search purposes, so that you can search on part numbers like aaa-9999-bbb. Another example is searching for “C++”.

Boolean combinations

- Combine the terms described above using AND, OR, NOT – create arbitrarily complex query expressions to get exactly and only what you are looking for.

Relevance Ranking

- Sort results according to relevance. The relevance of each result is calculated based on the number of times a word appears in the document, how common the word is in your document set, and the strength of themes. So a document that contains an uncommon word many times will rise to the top of a sorted list.
- Affect ranking by weighting terms. Retrieve documents that contain “Java” and “XML”, but if a document contains only one of those terms then list documents that contain “java” to appear before those that contain “XML”.
- Affect ranking by number of matching terms. Retrieve documents that contain “Java” and/or “XML” and/or “objects”, listing documents that have all 3 terms first, then documents that have any 2 of those terms, then documents that have any of those terms (the “accumulator” operator).
ADVANCED FEATURES

Intelligent text management
Oracle Text lets you manage and search your text data intelligently. Oracle Text “understands” text, using a very large, and custom developed knowledge base. Oracle Text users can:

- **Search for a document by theme** – find all documents that are about cars, whether or not they contain the word “car”.
- **Get the themes of a document** – find out what a document is about, not just the words it contains.
- **Get the gist of a document** – obtain a document by extracting the sentences or paragraphs that contribute most to the major themes of the document.
- **Get the theme gist of a document** – generate a gist of a document by extracting the sentences or paragraphs that contribute most to any of its themes. For example, find all the documents that are about wine: then show what else those documents are about (the document themes): summarize the top document: then summarize the wine content of the document (tell me what the document says about wine).
- **Extend the knowledge base to include terms from a particular industry**. For example, you might want to include more medical terms in the knowledge base, for more accurate search and better theme extraction in a medical application. Simply load a standard medical thesaurus into the knowledgebase.
- **Extend the knowledge base to make new associations between terms**. For example, you may want to promote a particular digital camera on a shopping web site. Simply associate the phrase “digital camera” with your promotional item, and camera shoppers will see the promotional item at the top of their search results.

Oracle Text also provides:

- **Classification**: efficiently classify a stream of documents according to a set of queries. For example, each time an e-mail arrives at a call center, send it to all interested agents. Or each time a news item arrives, post it to the relevant web sites or pages.
- **Clustering**: group similar documents together according to their themes.
- **Query refinement**: provide information for the user to refine his search
  - Show how a query was expanded: e.g. show the terms expanded from a fuzzy search or a wildcard search.
  - Show the words that are alphabetically close to your search term, with a count of hits.
  - Show the words or phrases that are semantically close to your search term (broader terms, narrower terms, related terms, etc.)

**XML search**
For XML documents, Oracle Text offers advanced within-section searching:

- Search for a document that contains search terms within an XML-tagged section.
- Nested section search.
- Search within attribute values.
Oracle Text: Features Overview

- Map multiple tags to the same name.
- Path searching, section existence (a subset of XPath).
- Ability to search for content and structure at the same time.

Document services

Oracle Text provides a number of utilities for you to view text, no matter how that text is stored

- Oracle Text supports 150+ document formats through its Inso filtering technology – includes all common document formats including XML, PDF, and MS Office.
- View the HTML version of any text, including formatted documents such as PDF, MS Office, etc.
- View the HTML version of any text, with search terms highlighted, with navigation to next/previous term in the text.
- Provide markup information: i.e. offset and length of each search term in the text, to be used for example by a third party viewer.
- Any custom filter set may be used.

MULTILINGUAL FEATURES

- Oracle Text supports all Oracle NLS character-sets. For example, ASCII, UTF-8, JA165JIS, GBK, BIG5, etc.
- Oracle Text supports search across documents in western languages (English, French, Spanish, German, etc.), as well as Japanese, Korean, Traditional, and Simplified Chinese.
- Oracle Text is completely flexible in the combination of formats (MS Office, PDF, etc.), languages and character-sets.
- Language-specific features include:
  - German and Dutch Word Decompounding – German and Dutch words are often made up of several component words. A decompounding index allows you to search on a whole word or any of its parts.
  - Alternate spelling – allows for alternate spelling. For example “Muenchen” and “München”.
  - Base letter indexing – indexes the base letter version of each word, so you don’t need to supply special characters when searching for French words for example.
  - Extensible Knowledge base – the Oracle Text knowledgebase is supplied in English and French, and is extensible in any single-byte, white space-delimited language.
- Cross-language search:
  - Can mix languages, character sets within a document collection (e.g. Chinese and English documents).
  - Can use English to query e.g. Chinese terms or vice versa.
WHAT IS NEW IN ORACLE9\textit{R2}

- Parallel indexing: users can create an index on a non-partitioned table in parallel.
- Parallel synchronization: synchronization of context indexes can now be run in parallel.
- Parallel query: users can parallelize contains queries on local partitioned context indexes.
- Create index online: users can perform DML on a table while creating and index.
- Exchange partition: improve manageability and scalability of text indexes.
- \texttt{Ora:contains()}: Linguistic processing on an XPath query.
- URIType columns can be indexed natively using Oracle Text. No special datastore is needed.
- \texttt{CTX\_CLS}: a package for building classification systems.
- \texttt{CTX\_REPORT}: DBA tools package to assist on tuning and performance monitoring.
- Web query parser: no need to write a query parser for web applications.
- Enhanced CTXCAT query language.
- Internationalization performance improvements when stem form is stored at index time.
- Extensible lexer: users can provide a different lexer.
- Performance and additional character support for Chinese.
- Reverse docid sorting.
- JDeveloper Text wizard: an add-in wizard for generation of simple text search applications. The wizard is available from the OTN (Oracle Technology Network) website.
- Clustering package for building clusters based on themes. The package is available from the OTN website.

REFERENCES

3. Oracle Text Home Page (http://otn.oracle.com/products/text/)