 Oracle一直致力于全文检索技术的研究，当Oracle9i Rlease2发布之时，Oracle数据库的全文检索技术已经非常完美，Oracle Text使Oracle9i具备了强大的文本检索能力和智能化的文本管理能力。Oracle Text是Oracle9i采用的新名称，在Oracle8/8i中它被称作Oracle interMedia Text，在Oracle8以前它的名称是Oracle ConText Cartridge。使用Oracle9i和Oracle Text，可以方便而有效地利用标准的SQL工具来构建基于文本的新的开发工具或对现有应用程序进行扩展。应用程序开发人员可以在任何使用文本的Oracle数据库应用程序中充分利用Oracle Text搜索，应用范围可以是现有应用程序中可搜索的注释字段，也可是实现涉及多种文档格式和复杂搜索标准的大型文档管理系统。Oracle Text支持Oracle数据库所支持的大多数语言的基本全文搜索功能。本文将介绍如何使用Oracle9i的全文检索技术来为自己的应用提供一个优秀的解决方案。 1 Oracle Text的体系架构 下图是Oracle Text的体系架构。 图1 Oracle Text的体系架构 以上面的体系架构图为基础，Oracle Text 索引文档时所使用的主要逻辑步骤如下：（1）数据存储逻辑搜索表的所有行，并读取列中的数据。通常，这只是列数据，但有些数据存储使用列数据作为文档数据的指针。例如，URL_DATASTORE 将列数据作为 URL 使用。（2）过滤器提取文档数据并将其转换为文本表示方式。存储二进制文档 (如 Word 或 Acrobat 文件) 时需要这样做。过滤器的输出不必是纯文本格式 -- 它可以是 XML 或

HTML 之类的文本格式。（3）分段器提取过滤器的输出信息，并将其转换为纯文本。包括 XML 和 HTML 在内的不同文本格式有不同的分段器。转换为纯文本涉及检测重要文档段标记、移去不可见的信息和文本重新格式化。（4）词法分析器提取分段器中的纯文本，并将其拆分为不连续的标记。既存在空白字符分隔语言使用的词法分析器，也存在分段复杂的亚洲语言使用的专门词法分析器。（5）索引引擎提取词法分析器中的所有标记、文档段在分段器中的偏移量以及被称为非索引字的低信息含量字列表，并构建反向索引。倒排索引存储标记和含有这些标记的文档。2 简单的示例 这里先给出一个简单示例说利用Oracle Text实现全文检索的方法与步骤，在后面在进行具体的说明。Orcale9i提供了Oracle Text Manager可以简化许多工作，所有在Oracle Text Manager中完成的工作，都可以在通过PL/SQL来实现。要使用Oracle Text，必须具有CTXAPP角色或者是CTXSYS用户。Oracle Text为系统管理员提供CTXSYS用户，为应用程序开发人员提供CTXAPP角色。 CTXSYS用户可执行以下任务：启动Oracle Text服务器，执行CTXAPP角色的所有任务。具有CTXAPP角色的用户可执行以下任务： 创建索引，管理 Oracle Text 数据字典，包括创建和删除首选项，进行Oracle Text 查询，使用 Oracle Text PL/SQL程序包。 使用Oracle Text的步骤：（1）创建表来保存某些文档。该示例使用一个主关键字列来标识每个文档，使用一个小的VARCHAR2列来保存每个文档。CREATE TABLE docs (id NUMBER PRIMARY KEY, text VACHAR2(80)).（2）将两个示例文档置入该表：INSERT INTO docs VALUES (1，the first doc)；INSERT INTO docs

VALUES (2，the second doc)；COMMIT；（3）使用Oracle Text Manager来创建和修改首选项，首选项将与索引相关联。（4）使用Oracle Text Manager创建文本索引。另外，可以输入以下使用默认首选项的 SQL 语句：CREATE INDEX doc_index ON docs(text) INDEXTYPE IS CTXSYS.CONTEXT.（5）使用 CONTAINS 函数，发出基于内容的文档查询。例如：SELECT id FROM docs WHERE CONTAINS (text, first) > 0. 这将在文本列包含单词 first (即文档1) 的 docs 中查找所有行。语句中的>0部分是有效的Oracle SQL所必需的，Oracle SQL不支持函数的布尔返回值。 以上只是一个简单的示例，旨在给出使用Oracle Text建立全文索引的完整步骤，归纳起来如下：（1）建表并装载文本（包含带有需要检索的文本字段）（2）配置索引（3）建立索引（4）发出查询（5）索引维护：同步与优化（将在后面介绍） 3 文本装载 要实现文本的全文检索首先必须把正确的文本加载到数据库表中，默认的建立索引行为要求将文档装载在文本列中，尽管可以用其它方式 (包括文件系统和 URL 形式)存储文档 (在"数据存储"选项进行设置)。默认情况下，系统应该将文档装载在文本列中。文本列可以是VARCHAR2、CLOB、BLOB、CHAR或BFILE。注意，只有在将Oracle7系统移植到Oracle8的情况下才支持用LONG和LONG RAW 这两个相反的列类型存储文本。不能为列类型NCLOB、DATE和NUMBER建立索引。 关于文档格式，因为系统能为包括HTML、PDF、Microsoft Word和纯文本在内的大多数文档格式建立索引，可以将其中的任何文档类型装载到文本列中(在"过滤器"选项中设置)。有关所支持的文档格式的详细信息，可以参阅Oracle Text Users Guide and

Reference 中的附录"Supported Filter Formats"。 装载方法主要有以下几种：（1）SQL INSERT 语句 （2）ctxload 可执行文件 （3）SQL*Loader （4）从 BFILE 中装载 LOB 的 DBMS_LOB.LOADFROMFILE() PL/SQL 过程 （5）Oracle Call Interface4 为文本建立索引 文本装入文本列后，就可以创建Oracle Text索引。文档以许多不同方案、格式和语言存储。因此，每个 Oracle Text 索引有许多需要设置的选项，以针对特定情况配置索引。创建索引时，Oracle Text可使用若干个默认值，但在大多数情况下要求用户通过指定首选项来配置索引。 每个索引的许多选项组成功能组，称为"类"，每个类集中体现配置的某一方面，可以认为这些类就是与文档数据库有关的一些问题。例如：数据存储、过滤器、词法分析器、相关词表、存储等。[1] [2] 下一页 100Test 下载频道开通，各类考试题目直接下载。详细请访问 www.100test.com