中英文全文检索——

TRIP 在大型数据库系统中的应用

新华通讯社技术局 王豹臣

摘要：本文介绍了 *TRIP* 数据库管理系统主要特性及功能，实现中英文全文检索的途径，以及在大型新华数据库中的应用情况。简要分析了系统检索响应时间、占空比等技术指标性能。

TRIP 与 ORACLE、SYBASE 一样，是数据库管理系统，为瑞典 PATALOG 公司的产品。我们在建立大型综合性新华数据库系统中选用了 TRIP 系统，在 VAX 6410、DECNET 及 CLASTER 环境下经过一年多的运行，效果良好。

一、选用 TRIP 的缘由

新闻单位是信息和资料高度集中的部门，号称是信息中心，消息总汇。在建立新闻数据库的过程中，首先对浩瀚的资料进行了分析，发现其突出的特点是，数据量大、时效性强、更新快、规范性差，其大部份为文献型资料。这就要求所要建立的系统必须适应这些特点，如果系统采取惯用的受控检索方法，不仅需要大量人力和物力进行标引，而且使资料进库时间周期延长，大大降低了系统的实用性，所以系统必需具有不需标引的中英文全文检索能力。TRIP 恰好可以满足这一基本要求。

TRIP 系统是以文件倒排技术为基础的数据库管理系统，具有较强的数据库全文检索及管理功能。它既可处理文献型信息，又可处理数字型信息。如报告、合同、建议、论文、手册、标准、产品说明、市场分析等。TRIP 主要运行在 VMS 下，也可在 UNIX 下运行；既可用作单用户私人文件管理，也可用作象世界著名的 DIALOG 公司那样的大型文献数据库检索服务。用户可以从一个词、词的一部分；词组；数字、数字范围；日期、日期范围；时间、时间范围的角度进行查找，可以说当今世界上常用的检索方法它都支持。TRIP 系统对于文本字段(TEXT)内容中的每个词进行位置扫描，然后排序，最后建立以每个词的散离码为表目的倒排文件，此时文本内容的每个有检索意义的词(字)都成为可以检索的“关键词”，TRIP 系统按字处理中文的这种方法，不仅减去了人工标引工作，而且避开了汉字词义分析、自动切分这一技术难点。可以看出 TRIP 最适合用以文献型资料为主的检索系统中。

二、系统应用

1.定义数据库

TRIP 为用户提供了建库使用的多种数据类型，包括：文本(TEXT)、词组(Phrase)、数字(Number)、日期(Date)、时间(Time)和字符串(String)，其中字符串中主要用来存放图形字符，以便使 TRIP 能够处理图形信息。用 TRIP 建立的数据库其记录中各字段的长度不必定义，实际数据可长可短，另外对于数据库大小、记录的多少以及每个记录的大小均无限定。在定义数据库的过程中，需要明确有多少字段组成，每个字段属于哪种数据类型，字段是否为检索入口等。

TRIP 对数据库的容量大小没有限制，但从本系统的应用情况看，一个实际数据库容量不宜过大，最大数据量最好不超过 40MB～60MB。这样不仅数据装填快，检索快，而且维护方便。例如本系统建立的“新华社新闻稿库”，每年为一个物理库。TRIP 系统为了满足用户对多年同一数据的检索，提供了同时可以打开 30 个数据库以及采用逻辑库名方式进行查找的能力。

2.数据预处理

数据库中的数据与文件数据是不同的，任何文本数据要入数据库，均需经过预处理的过程。数据预处理通常做两方面的工作，一是提取数据库有关字段数据；二是将文本格式变为入库的格式，TRIP 系统要求的入库格式比较简单直观，每个记录头识别符为 R，结尾符为∧，字段识别符∧xF，其中"x"为字段序号。

数据处理的方式很多，不同的信息源可以采用不同的方式，可以是批处理，采用的程序方法自动提取字段数据，自动形成入库格式和启动装载倒排命令，整个过程自动进行。本系统中的"新华社新闻稿库"即采用这种方式，每天播发的新闻稿第二天一早就可供用户查询；除批处理方式外，还可以是脱机处理以及联机人工录入处理等。数据预处理是数据库系统中的重要部分，直接影响系统的实用和时效，其工作量也往往较大，应给予足够的重视。

3.用户界面

对于一个系统来讲，用户界面十分重要，现在很多系统都在追求用户界面的友好性，努力使界面清晰、操作简单、易学易用。TRIP 系统为用户提供了两种友好的检索方式，一种是 CCL 查询语言，一种是格式查询，CCL 语言为欧洲网的公共命令语言，功能比较强，虽然与国际通用的 SQL 语言不同，但操作简单，非专业人员易于掌握，可以直接作为用户界面。CCL 语言基本命令 29 条，其中 13 条为用户操作命令，其余为系统管理人员使用，用户最常用命令不过 3～4 条，采用 CCL 命令查找时，屏幕分为 4 个窗口：命令窗口，用作输入 CCL 命令；数据库窗，显示已打开的数据库名；过程窗，显示已执行的命令及其结果；输出窗，显示查询结果。各种数据在不同窗口显示，用户一目了然。CCL 语言还可以写成程序方式存放在 TRIP 的过程模块中，然后令其执行，以实现程序方式的检索。其命令格式为：

trip／user＝用户名／Psss／口令／strat＝"run 过程名"

本系统就是利用这种方法与异型机系统实现了联机检索。

TRIP 设计的打印界面也很灵活、方便简单。用户采用 PRINT 命令后跟不同参数，即可实现不同形式和内容的打印。特别值得说明的是利用 TRIP 命令很容易将查询结果变为 VMS 下的文本文件，供用户作进一步的处理。另外 TRIP 的屏幕设计颇似良好的微机方式，数据库建立、检索、数据录入、系统管理等功能都通过下拉式菜单方式实现。

4.定题检索

定题检索是数据库系统应具有的重要功能之一，TRIP 为用户提供了 SDI 服务，在数据库中每个记录都有一时间标记，表示该记录是何时产生的，以及何时对它做过一次修改，利用下述命令即可实现定题检索目的。

DEFINE SCOPE SDI（启动 SDI 服务）

FIND 定题检索词（定题检索）

UPDATE SDI（结束 SDI 服务）

最后一条命令给出之后，TRIP 系统将修改用户记录，指示出刚被检索的记录在下一次的 SDI 服务中不再被检索。本系统应用了这种功能，效果很好，不仅能达到对某种专题资料的检索跟踪，而且检索响应时间大大缩小，很受用户欢迎。

5.主题词表管理

TRIP 系统具有词表管理的能力，这是一般其它数据库所不具有的。TRIP 将词表建成数据库，每个记录都含有主词，即受控词(CT)以及与该词有关的上位词(BT)、下位词(NT)、相关词(RT)、用代词(UF)等，每个词为一个字段。词表库与实用库利用 CCL 命令连接后即可实现词表方式检索，以提高系统的查全率。如"按劳分配"为受控词，"按劳付酬"为用代词，通过词表方式检索，不仅检索出有"按劳分配"的全部记录，而且可以查出含"按劳付酬"的记录。TRIP 系统允许建多个词表库，可以分别与实用库相连接，以提高系统能力和方便用户查询。本系统已将新华社编制的我国"新闻叙词表"建成了词表库，并投入使用。

三、几项性能指标的应用分析

1.检索响应时间

对于一个数据库系统而言，影响检索响应时间的因素很多，除系统运行环境外，最重要的是数据库管理系统具有的检索响应时间特性，从本系统的实际应用看，TRIP 的检索响应时间是令人满意的。经测试 8 个用户

（下转第 43 页）

(上接第 40 页)

采取相同的检索式。查询同一库,同时进行全文检索,检索响应时间平均 5 秒以内。检索响应时间的快慢与检索策略关系很大,针对 TRIP 而言,它是以每个汉字为单位进行切分的,并对每个字作了索引,根据这种特点如果选用检索词为一个单"字",检索响应时间最快(一秒钟),检索词为"双字"词时,可以理解为是两个单"字"逻辑与,响应时间与字频度有关,频度越小响应时间越快。可以看出,检索词含字过多时,响应时间会变慢,对于 TRIP 系统检索词最好在 5～6 个字以内。另外 TRIP 还提供词组(Phrase)字段全文检索,这就使用户有了快速查询的另一手段。

2.占空比

占空比是衡量一个数据库管理系统的重要性能指标之一。占空比是指实际数据量与数据装载倒排后占用空间二者之间的比例关系,显然占空比越小越好,从本系统实际情况看,TRIP 的这项性能指标是比较好的。如每个库由三个文件组成,即顺序文件、倒排文件和词汇文件,不同类型和文种的数据其占空比也有所不同。例如"新华社中文新闻稿库"占空比约 1:2.4,即 1MB 数据所需空间 2.4MB;"英文新闻稿库"约 1:1.8。另外随着库容量的增大,占空比将会随之减小。这是因为词汇文件加大到一定程序后其增长速度明显减慢所致。

3.查全率和查准率

对于一个数据库而言,即要有高的查全率,又要有好的查准率。TRIP 系统具有很高的查全率,而查准率相比之下较低,如果采用一定的检索策略,选择好检索词,其查准率会明显提高,检索响应时间也会加快。另外 TRIP 提供了二次查询功能,以取得满意效果。