全文检索技术和 CGRS 软件

浙江省经济信息中心 毛楚祥

摘要：本文通过国家经济文件文献信息系统的开发，总结了全文信息检索需要解决的技术问题及其解决途径，并介绍了为此而开发的通用全文本型信息检索系统实用软件 *CCRS* 的功能。

一、前言

信息检索系统是利用计算机加工、存储文献信息为用户提供相关文献的检索服务。七十年代以来，我国已建立了一些信息检索系统，在已建立的系统中绝大部分存储的文献是二次文献，用户检索到是一些指标性的信息，例如书目或论文的标题等，要想取得实质性详细数据则只能根据这些提示性的结果从原文中获得。按照国内外经验，一个经受控标引的二次文献至少半年以后才能和读者见面，时效性较差。而在电子信息服务和办公自动化资料的存储和检索的应用中，用户已远远不满足从计算机中仅获取延迟半年指示性的二次文献信息，它们要求及时、准确、直接从计算机中检索到文献的全文。八十年代以来全文检索技术和全文数据库检索系统发展十分迅速。例如 DIALOG 系统 1983 年在 228 个数据库中，全文数据库七个占数据库总量的 3%，至 1991 年在 345 个数据库中全文数据库有 86 个，占总数的 25%。1988 年以来，DIALOG 数据库中全文数据库的数量如下：

年	总数据据库量	全文数据库	全文库占的比例
1988	279	45	16%
1989	289	54	18.7%
1990	312	61	22%
1991	345	86	25%

由于全文数据库检索系统是基于全文标引的检索系统，它把文献中出现的每一个词或短语都看作为一个检索入口，允许用户使用自然语言检索全文的任何章、节、段句，同时全文检索不需要人工整理、标引成二次文献，对全文的标引往往采用计算机自动标引的方法，很显然它提供文献速度快、专指性强、内容齐全、用户不须再通过它去从原文中找详细资料。这也是全文检索为什么受到人们普遍欢迎的原因。

二、全文检索技术

从全文检索系统的含义可以看出全文检索主要涉及三方面的问题：全文数据库、全文检索、自然语言查询。对此，全文检索系统在理论上和技术上有许多新的问题值得研究。这主要有：全文数据库的存储组织、全文标引技术、用户界面(自然语言查询)、检索算法、全文检索系统性能的优化。

1.全文数据库存储组织

传统的数据库管理系统是为处理格式化而设计的。全文数据库长度变化很大，不适合用传统的数据库管理系统来存储和管理。根据检索方法的不同全文数据组织也很不相同。目前常用的以下几种：顺序组织、倒排文件方法、标识文件方法、物理聚集方法和多属性散列方法。顺序组织空间节约、查询效率低、倒排索引方法检索效率最高，但所需要索引空间大，约占数据文件的 50%～300%。标识文件方法其检索速度介于顺序组织和倒排索引组织之间，节约空间，插入处理也方便。

2.全文标引技术

为了能对全文进行检索，必须在文献中提供各种类型的检索词。抽取检索词可以由人工来完成。也可以由计算机来做。人工抽取检索词对文献进行标引工作量太大，难以满足实际需要。同时，信息检索的一个发展趋势

© 中国科学院软件研究所 http://www.c-s-a.org.cn

是逐步增加采用关键字这种自然语言的基本单位作为检索依据。经过科研人员多年的努力，书面汉语自动标引技术已经取得长足的进展，有多个通过计算机自动从全文中抽取关键字的系统已实际使用。在使用的汉语自动抽词方法主要有以下几种：词典切分词组法、部件词典法、词频统计逐步求精词典法、后缀链接表法、主题词典法、组配组词法、统计信息切分法、词素自动分词、大词典分词法等。以上这些方法都没有很好解决词的外延难以确定、多义性能于排除这两个问题。看来一时不大可能解决，要彻底解决非得要等人工智能技术和语义分词技术的发展。

“单汉字自动标引”是以单汉字为处理单位，通常事先将文本中所有汉字去除虚字，如“的、在、地、了、……”等留下的其它汉字均作为标引词用。采用单汉字模式建立起来的检索系统，不同于以词为单位进行标引和检索的“常规”检索系统，在系统性能，检索方法上均有自己的特点。它易于实现截断检索，且检索模式组配灵活，可以任意调整检索专声指度水平，有利于“字面成族”检索，同时也避免人工标引的主观性，节约大量人工标引的劳动力。存在的问题是占用索引空间往往大大超过数据空间。

3.检索技术及系统的优化

全文数据库的存储组织和对它的检索是建立全文检索的两个主要技术。

全文检索应具有一般布尔检索功能，如检索项之间AND、OR、NOT、(、)运算、位置检索、截词检索功能。位置检索是对指出词在文本中的位置要求进行查询，截词检索是字符屏蔽操作的特殊形式。例如词的前方一致、后方一致、中间一致运算。

在一个没有词汇控制的全文检索系统中为了优化全文检索性能，提高查全率和查准率，建立后控词表是一种有效的措施。所谓后控词表是将同义词和相关词收集起来编成的词表。这种词表不供标引使用只供检索之用。用户为了查全某一主题的资料，不了解相应的同义词或相关词，只要输入一个自己已知的检索词，系统通过后控词表，自动地将有关的同义词、相关词用“OR”连接起来进行查词从而提高查全率。

4.用户界面技术

在电子信息服务和办公自动化的应用中，全文检索系统的用户已不是专职的检索人员和训练有素的操作人员，而是对情报检索和计算机了解甚少的办公人员。因此系统具有友好的界面是十分重要的。除要求有一般的应用软件友好界面设计的技术外，作为一个全文检索系统软件在界面设计时还要特别注意两点：(1)自然语言的接口。前面说到检索技术时要求应具有布尔检索的功能。对于一个不懂计算机的用户要求用“AND”、“OR”、“NOT”三种操作准确地描述查询要求，是极其不方便的。必须在检索系统中配备自然语言接口，用户使用它可方便地与检索系统交谈，信息检索系统的自然语言接口必须具有三个方面的功能：①自然语言的理解能力；②在表层内部表示的基础上借助知识规则推理出深层的内部表示；③构造全文数据库查询语句的能力。

(2)减少汉字输入。中文全文检索要求用户输入中文检索词，不少的用户不会用计算机汉字输入方法，从而对全文检索就望而却步。能否做到尽可能少，甚至不要用户键入汉字实现汉字查询呢？浙江经济信息中心研制通用全文检索软件 CGRS 采用“属性字典法”和“输出代替输入法”很好地解决了这个问题。

三、通用全文检索软件 CGRS

通用全文检索软件 CGRS 是国家“七五”科技攻关项目“国家经济文件文献信息系统”的软件科研成果，浙江经济信息中心的科技人员经过五年的努力使该软件在使用中不断地完善和发展，已有大、中、微机上都能运行的全文检索软件。它有在 IBM 计算机上使用的版本 H-CGRS / MVS、H-CGRS / VM、H-CGRS / VSE，在 NOVELL 网络下使用的 M-CGRS / NOVELL 版本和 M-CGRS / DOS 版本；它们形成了主微机一体化的产品。

CGRS 软件主要功能有：

1.允许定义一层至三层的文献数据库

每层数据库把正文分成若干章，每章分若干节(条)组成。检索时可以定位到文献的章或节位置。如法规数据库可以查询到符合要求的法律条文。

2.检索功能

对文献进行组配检索、字检索。在组配检索式中使用布尔算子、园括号，也可以用检索式进行选代检索；检

© 中国科学院软件研究所 http://www.c-s-a.org.cn

索词可进行前方一致和后方一致匹配检索。在M-CGRS字检索中利用前后台技术较好地解决检索速度和存储空间的矛盾。

3.自动抽取关键字

采用"字典抽词逐步求精"的办法对文献进行自动抽词,每次可以同时对五个字段抽取关键字。关键字允许中、西文混合。

4.同义词检索

为了优化检索性能,系统把关键字中的同义、相关词组合在一起建立近义词词典,检索时用户使用它进行扩检,提高全文检索的查全率。

5.文献格式检查功能

对入库文献的格式进行检查,保证入库文献的正确性。检查包括字段值的数据类型、取值范围、出现的次数等。

6.用户管理

使用文献库的用户必须事先注册。对用户的给定操作权限、所使用的库、口令等进行管理;同时对用户查用过的检索词进行记录。

7.查询统计实用功能

统计用户查询词的使用次数、数据空间占用数、资料输出数,用户查询费用记帐等。

8.数据库的维护

数据库中文献的删除、修改;对文献及工作文件的备份和恢复。

CGRS软件已在多个单位的电子信息检索服务、办公自动化中应用,存储了上亿级的数据获得了很大的成功。目前CGRS正在利用迅速发展的多媒体技术使它管理的类型更广泛,用户使用更方便。

© 中国科学院软件研究所 http://www.c-s-a.org.cn