第16卷 第3期
2006年3月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol.16 No.3
Mar. 2006

全文检索系统的数据预处理研究

韩 升,刘广志
(北京交通大学 软件学院,北京 100044)

摘 要:全文检索的应用导致了信息检索领域的一场革命,是文档数据库研发的核心。在一个全文检索系统中,全文索引数据库的建立是系统的基础,其设计结构直接影响到全文检索引擎的检索算法以及系统最终的检索效率。文中主要介绍全文检索系统中索引库结构设计、文本标引技术等数据预处理技术,以及全文检索系统索引数据库的数据处理流程。最后,在此基础上研究了全文检索系统索引库索引生成算法,给出了单个文档和批处理两种情况下的索引库索引生成算法。

关键词:全文检索;预处理;文本标引;索引数据库
中图分类号:TP311.13 **文献标识码**:A **文章编号**:1005-3751(2006)03-0208-03

Study of Data-Pretreatment for Full-Text Search System

HAN Sheng, LIU Guang-zhi
(School of Software, Beijing Jiaotong University, Beijing 100044, China)

Abstract: The application of full-text search has caused a revolution of the information retrieval field. It is the core that the file database researches and develops. In a full-text search system, the setting-up of the index database of full text is a systematic foundation. Its project organization influences the final search efficiency of searching algorithm and system of the full-text search engine directly. This paper introduces such data-pretreatment technology as index database structural design, text index technology, etc. Also introduces that in the full-text retrieval system mainly, and the data processing procedure of index database of full-text retrieval system. Finally, studied the produce-algorithms of index database of full-text retrieval system on this basis, provided produce-algorithm of index database under two kinds of situations: individual file and batch processing.

Key words: full-text search; pretreatment; document indexing; index database

0 引 言

数字化革命和因特网的大发展,带来了经济、贸易、信息传播的全球化,深刻影响着社会的各个层面。巨量信息涌入因特网及各行各业之中,使其信息极为丰富,也使信息查询、检索十分困难。随着用户对信息检索服务的要求不断提高,信息检索技术也在不断地推陈出新,发展进步,而全文检索技术作为一种高效、强大的信息检索技术,近年来受到了广泛的关注,它的出现导致了信息检索领域的一场革命。

全文检索(Full-Text Retrieval),是以各类数据诸如文字、声音、图像等为处理对象,提供按照数据资料的内容而不是外在特征来实现的信息检索的手段[1]。与以前的情报检索相比,全文检索提供了全新的、强大的检索功能。全文检索可以直接根据文献资料的内容进行检索,支持多角度、多侧面的综合查询方式,有很高的查准查全率,从而能够为用户提供灵活、方便、快速的信息查询服务。近期,笔者参加了某高校的数字图书馆建设及相关全文检索系统的开发任务,对全文检索系统数据预处理技术进行了一些研究,文中对此项研究进行了总结。

1 全文检索系统的预处理技术

一个完整的全文检索系统应该包含多个功能模块,但其核心可以分为索引库与全文检索引擎两部分。图1[2]中描述了一个全文检索系统的工作模块图。人们所说的全文检索的数据预处理,也就是将数据源信息收集、加工、生成全文索引数据库的过程。

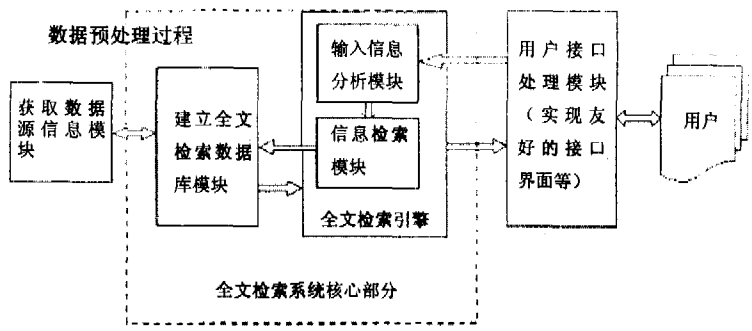

图1 全文检索系统工作模块结构图

收稿日期:2005-06-22
作者简介:韩 升(1980—),男,山西长治人,硕士研究生,研究方向为软件工程、数据仓库和数据挖掘;导师:黄厚宽,教授,研究方向为软件工程、数据仓库和数据挖掘。

全文检索索引库是一个全文检索系统的基础,它以特定的结构存储了数据资源的全文信息,从而为全文检索引擎提供可检索的数据对象,因此,它的体系结构和数据组织方式直接决定了检索引擎的检索算法和检索效率。所以全文检索索引数据库的设计与组织就显得尤为重要。

1.1 全文检索索引数据库的结构

索引数据库一般由一个变长的主文件和一个在索引文件控制下的倒排文件组成。索引文件和倒排文件在物理上是分开的,逻辑上也可组合成为倒排索引文件。检索时,由索引文件指向倒排文件,倒排文件指向主文件[3]。索引数据库的体系结构设计与全文检索中把检索点放在词一级还是单字一级息息相关[2]。这就涉到文献内容的标引问题,采用不同的标引方式,相应的索引库的结构与数据组织方式也不同。目前在中文文本的自动标引方面,较实用的方法有词典切分法、单汉字标引法和一些特殊处理的方法。

词典切分法以词为单位进行索引。其优点是对于大规模应用,索引库可以组织得比较小,检索的处理速度也比较快,而且还容易实现同义词、反义词的概念检索。但是对于中文文献来说,汉语语言中词与词之间的分界不明确,语义和语法结构也十分复杂,这样就为汉语自然语言文本的切分带来了障碍[2]。单汉字标引法以中文单个汉字为单位进行索引。由于系统能自动识别每个汉字和其他符号分别作索引,不需进行额外的标引工作,从而大大提高了建库的效率;另外在单汉字标引检索系统中,全文中的任何信息都能被检索出来,避免了人工标引的主观武断性和标引深度不足的问题。但由于汉字自身的特点使然,使得单汉字标引检索系统有倒排文件很大;建立索引时间长;检索速度慢;误检率高;同义词、相关词无法控制等缺点。

基于这两种标引方式,全文检索索引数据库的结构可以分为以下两种主要方式:

1)以词为单位的倒排索引结构[4]。

典型的基于词的倒排索引结构(见图2)包含两部分:

① 中文词组成向量(称之为词汇表),它包含了词的基本信息和词索引在索引文件中的偏移量;

② 对于词汇表中的每一个词,都有一个它出现过的文档列表,包含了出现文档编号和在此文档中该词的词频和出现位置序列。

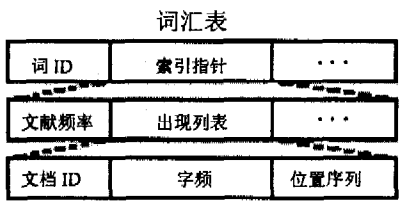

图2　词索引结构

除了词汇表外,基于词表的检索系统一般还要建立同义词表、反义词表、关联词表等多个辅助词表,用于进行同义词、反义词等的概念检索。

2)以字为单位的倒排索引结构[5]。

索引库的主要部分是每个汉字的字表信息,索引库中的字表结构见图3。其中字符 i 对应的字表记录了该字符的源文档中的所有出现位置 Pix ,出现位置通常用字符相对于文档头的偏移字节数表示。建立字表索引时,需要扫描整个源文档,对所出现的每一个有效字符,计算其在文档中的出现位置并将该位置值加入到对应的字表中。

…	……
阿	P11　P12　P13…
阿	P21　P22　P23…
…	……
网	Pi1　Pi2　Pi3 … Pim …
…	……
络	Pj1　Pj2　Pj3 … Pjn …
…	……

图3　字索引结构

当然,采用特殊标引方法或采用了特殊技术的全文检索系统的索引数据库机构与以上叙述有所不同,但大体来说,索引数据库的组织方式仍然是以上述两种索引结构为基础。

1.2 全文检索索引数据库数据添加与维护

设计好全文数据库的结构与数据组织方式后,就可以向索引数据库添加数据了[3]。全文检索索引数据库的生成包括数据准备、文本预处理和数据加载3个步骤。

1)数据准备。是指对计划加载到全文数据库中的数据进行收集、整理、归类等预先处理的过程。加载到全文数据库中的数据可以从多种途径获得。常见的数据来源有:电脑打字产生的文件;电子印刷产生的文稿;计算机网上传送的文件;电子出版物;图文处理产生的文件;专门组织人力录入建库。数据收集起来之后,要进行一些简单的分类。一般是按照数据内容进行分类,同一类内容加载到同一库中,这样便于查找。如果数据总量不大,比如不超过100万字,也可不进行分类。分类对于数据量大的情况,效果比较明显。

2)文本标引。包括:

①规范格式。当格式多种多样时,应加以整理,使文献的格式规范化。如标题与正文之间空几行;段落开始行缩进几个字;英文用什么体等等。

②批式标引。文本预处理阶段完成的批式标引效率较高。这是在建立全文数据库之前,特别是数据加载之前,对非单字索引的正文,利用文字处理软件和专用自动标引软件对数据进行的标引。

建立标引词表有几种途径:由系统建立者在浏览文本后赋词编制;由编者在计算机上对文本中的词加上特殊符号后,由专用软件对其进行搜集、合并、排序、去重而成;在批式标引基础上增加属性标引。

3)数据加载。数据准备好以后,便可以加载(拷入、输入)到数据库文件中去了。加载数据有单篇方式或批量方式。单篇方式一次加载一篇,适于平时文献随时加载的情况;批量方式一次加载多篇,适于集中大量加载的情况。

数据库建立以后，需要经常对数据库的内容进行索引、更新、追加和清理，以保证数据库的实用性、有效性和时新性。对全文数据库的维护通常包括：全文数据库的结构定义内容；全文数据库的数据内容；全文系统中所用词表(字表)；存储空间的利用统计及调整。

1.3 全文检索索引数据库的索引生成算法

以词索引全文检索系统为例，系统建立索引的算法流程如下：

1)对文档进行自动分词，对结果排序，合并相同词的信息。

2)定位词在词表中的位置，如果是以前未出现过的词，就在词汇表的末尾分配一个固定大小的基本空间，对于低频词来说太大的基本空间将造成浪费，所以需要分配合适大小的基本空间。

3)如果这个词以前出现过，将文档的读写指针定位到这个词的索引区的末尾。

4)写入每个词的索引信息到索引区。

5)对于文档中的每个词，重复步骤(2)～(4)，直到把所有词的索引信息写入索引区中。

对于批处理方式加载数据的系统，可以将词索引信息先写到一个临时文件当中，当将所有文件的信息处理完毕后，再统一将临时文件中的索引信息写入词汇表索引区去，也可以动态地将内存区域作为临时文件进行处理，相关算法如下(临时文件的结构见图 4)：

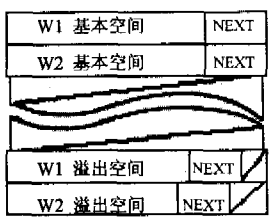

图 4 临时文件结构图

(1)对文档进行自动分词，对结果排序，合并相同词的信息。

(2)定位词在词表中的位置，得到词索引区在临时文件中的偏移量，如果是以前未出现过的词，就在临时文件的末尾分配一个固定大小的基本空间，对于低频词来说太大的基本空间将造成浪费，所以需要分配合适大小的基本空间。

(3)如果这个词以前出现过，将文档的读写指针定位到临时文件中这个词的索引区的末尾。

(4)写入每个词的索引信息到临时文件。如果此时分配给该词的空间用完，则在临时文件末尾给其分配新的溢出空间，出现次数越多的词分配的溢出空间也越大。索引写完后，将该词上一索引区的向前指针更新为新分配空间在临时文档中的偏移量。

(5)对于文档中的每个词，重复步骤(2)～(4)；对于每篇文档重复步骤(1)～(5)。

(6)所有文档处理完后，对于每个词，将分散在临时文档中的索引信息合并在一起，然后按照图 2 的格式写入最终的倒排文档。

以字为单位的全文检索系统的索引生成算法与词索引全文检索系统基本相同，不同之处在于字索引全文检索系统索引生成时不需要进行分词操作。

2 结束语

随着计算机技术及数据库技术，尤其是网络技术的飞速发展，人们对信息检索技术的研究也越来越向纵深发展。全文检索技术的出现，导致了信息检索领域的一场革命。它不仅可以实现情报检索的绝大部分功能，而且还能直接根据数据资料的内容进行检索，实现了多角度、多侧面地综合利用信息资源。文中系统地介绍了全文检索系统索引数据库的结构，文本标引等全文检索数据预处理技术，并给出了全文检索索引数据库索引生成算法，供大家参考研究。

参考文献：

[1] 牟有静，侯丽梅. 浅谈数字图书馆与全文检索技术[J]. 情报学报，2002，21(5)：535－537.

[2] 李 梅，王庆林. 中文全文检索技术的研究及实现[J]. 情报学报，2003，22(1)：10－17.

[3] 王兰成，蒋 丹，刘庆辉. 全文数据库建库原理与应用技术[J]. 情报学报，1999，18(4)：319－326.

[4] 陈 玮，陈玉鹏. 一种高效的全文检索索引技术[J]. 计算机应用研究，2004(7)：35－37.

[5] 曾元鉴，李孝明. 一个中文全文检索系统的设计与实现[J]. 计算机与数字工程，2004，32(3)：12－15.

(上接第 207 页)

金矿，还有待于开发，远没有探测完。当前对数字信号处理问题的关注是产业界流行的趋势。DSP 的市场范围十分广大，因此应该尽量避免在市场中出现硬碰硬的竞争。即使这样，也仍然可以获得可观的效益；值得为之奋斗。

参考文献：

[1] 张在峰，马义德. DSP——数字化时代的基因芯片[J]. 信息技术，2003(2)：53－56.

[2] 张小鸣，马正华. DSP 信号处理器的特点和应用方向[J]. 江苏石油化工学院学报，2001(9)：53－55.

[3] 杨之峰. DSP 在仪器仪表领域的应用[J]. 电子质量，2003(7)：32－33.

[4] 周宏图. DSP 应用——全数码助听器[J]. 福建师范大学学报，2002(12)：45－49.

[5] 吴炳欣. 数字信号处理器的市场竞争及技术发展趋势[J]. 世界电子元器件，2001(4)：9－12.