

· 实践平台 ·

# 面向大规模语料库的全文检索系统研究

贺 胜 (南京师范大学文学院 江苏南京 210029)

卢亚军 (西北民族大学藏语言文化学院 甘肃兰州 730030)

**摘 要:** 随着语料库规模的不断扩大和基于语料库的应用研究逐步拓展, 对语料库的全文检索成为语料库系统中不可缺少的重要组成部分。文章对面向大规模语料库的全文检索系统的索引模式、检索算法、检索表达式的构建、自动分词、系统组成等进行了研究, 并基于大规模语料库的语言文字信息处理和应用研究的需要, 开发了中文信息处理系统——“CIPP”。目前该系统具有全文检索、自动分词、语言统计等功能, 在千万字数量级的语料库中, 其全文平均检索时间小于1秒。

**关键词:** 语料库 全文检索 自动分词

中图分类号: G356

文献标识码: A

文章编号: 1003-6938(2008)04-0093-05

## Research of Full-Text Retrieval System for Large-Scale Corpus

He Sheng (School of Chinese Language and Literature, Nanjing Normal University, Nanjing, Jiangsu, 210029)

Lu Yajun (School of Tibetan language and Culture, Northwest University for Nationalities, Lanzhou, Gansu, 730030)

**Abstract:** Recent years have seen great expansion in Corpus scale and in application of corpus technology. Full-text search has become an indispensable component for a corpus. This thesis reports research on index model, search algorithm, search expressions, automatic Chinese segmentation, and system structure in large scale corpus systems. The paper also expounds CIPP, a Chinese information processing system implemented for the purpose. The system is efficient in full-text search, automatic Chinese segmentation and statistics. Time spent on conducting full-text searches in 10-million-token corpora is less than 1 second.

**Key words:** Corpus; full-text retrieval; automatic segmentation

CLC number: G356

Document code: A

Article ID: 1003-6938(2008)04-0093-05

### 1 引言

语料库是存储于计算机中并可利用计算机进行检索、查询、分析的语言材料总体。<sup>[1]</sup>语料库检索是一种全文检索技术, 但仅用普通的全文检索技术还不能满足基于语料库的检索要求。这是因为全文信息检索一般关心的是检索的内容, 不是检索目标的语言表述形式。而面向语言研究的语料库检索, 则特别注重语言的表述形式、语言现象及其上下文, 它既需要按照字、词、字串检索, 也需要把词语或字串的语言学属

性作为检索的目标和约束条件, 并把检索结果或结果出处按照研究的目的和任务之所需进行排序、输出。除此之外, 还要具有字频、词频和特定语言形式出现频率的统计、分析等功能。随着语料库规模的逐步扩大和应用研究领域的不断拓展, 许多相关技术问题还有待逐步解决。为此, 本文针对面向大规模语料库的全文检索进行了有侧重的研究和探讨。

### 2 相关技术概述

由于中文信息处理的复杂性和特殊性, 决定了在目前的

基金项目: 本文系江苏省社会科学基金项目《语料库通用加工与应用工具开发研究》(批准号: 07YYB003) 与国家社科基金2005重点项目《藏语语料库建设研究》(批准号: 05AYY001) 研究成果之一。

收稿日期: 2007-11-23; 责任编辑: 魏志鹏

理论与技术条件下,实现一个完善的大型中文全文检索系统还存在不少困难,国内许多有关研究全文检索的文献,其系统的实现大都是使用通用数据库系统提供的全文检索功能开发的。有一些系统虽然称作全文数据库,但其实质是通过检索放在关系数据库里的结构化数据,如标题、作者、关键词、文摘等,然后链接全文以获得全文,而真正实现全文检索的不多。另外,传统的全文检索大都是基于如Sql Server、Oracle、Mysql等数据库提供全文检索,但这些系统占用资源比较大,需要服务器环境支持,不适合Windows操作系统下的PC单机环境使用。现有的大部分系统尚存在一些缺陷,离成熟的系统还有一定的距离。本文就我们开发的面向大规模语料库的全文检索系统——CIPP的索引模式、相关检索算法、检索表达式构建、词索引模式下自动分词的实现等关键技术进行论述。

## 2.1 索引模式

在中文全文检索系统中,索引的基本元素可以是单个汉字字符,也可以是词。因此存在两种基本的索引库结构,即基于字表的索引库和基于词表的索引库。<sup>[2]</sup>相应的全文检索方法主要分为按字检索和按词检索两种。按字检索是指对于文章中的每一个字都建立索引,检索时将查询串分解为字的组合。按词检索是指对文章中的词,即语义单位建立索引,检索时按词检索。英文等西文的词由于它是用空格自然切分开的,因而在实现上与按字处理相类似;而像中文等亚洲文种的词则需要切分词,以达到按词索引的目的。对于以字为单位的这一类检索系统,其主要以单字索引和字符串匹配为关键技术,检索结果的查全率高,但由于未进行词的切分,所以检索结果中经常会出现“非词”的问题,因此查错率较高。例如要查找“华人”,检索结果中可能会出现“中华人民共和国”,要查找“出警”,检索结果可能出现“提出警告”、“放出警犬”等。为了解决这些问题,常常需要为字符串匹配的检索表达式另外设置限制条件,但这些限制条件大多是个性的,只能排除一部分“非词”的实例。要想从根本上解决这个问题,就必须对语料作词语切分,即以词为单位建立索引。其优点是检索结果的查准率高,但其查全率又被现有分词系统的切分精度所制约。

对于语料库检索来说,一个好的检索系统应该适用于各种不同的检索需求,能够为多种不同文种的语言研究服务。从语料库的应用角度出发,CIPP除了提供字索引、词索引模式及相关检索方式外,还提供了词及词性混合索引模式及相关检索方式。字索引模式是以单个汉字为单位建立索引库,检索时将查询串分解为字的组合;词索引模式下,系统调用分词模块对文本进行自动分词处理后,以词为单位建立索引,检索时系统再次调用分词模块将查询串切分为词的组合串;

词及词性混合索引模式是将已分词及词性标注的文本以词/词性为单位建立索引,检索时以词/词性串的形式进行查询。把词或词性作为检索的关键词或限制条件,可以得到关于这些语言学属性的检索和统计结果。

## 2.2 检索算法

CIPP的检索算法属于索引检索,使用的是经过改进的倒排文件索引结构,具有快速的查询速度和较高的空间压缩比。下面以词索引模式为例,介绍该结构及相应的生成算法。

设有两段英文文本1和2:

文本1的内容为: Jack lives in Nanjing, I live in Nanjing too.

文本2的内容为: He once lived in Shanghai.

### (1) 获取关键词

由于系统是基于关键词索引和查询的,所以首先要取得这两段文本的关键词,先找出文章中的所有单词(即进行分词)。文中的“in”“once”“too”等词没有实际意义,这些不代表概念的词一般可以过滤掉。另外,用户通常希望查“live”时能把含“lives”,“lived”的文句也找出来,所以需要把“lives”,“lived”还原成“live”。文本中的标点符号通常不表示某种概念,也可以过滤掉。经过以上处理后(在检索系统中以上处理由具体的语言分析接口完成):

文本1中的所有关键词为:[jack] [live] [nanjing] [i] [live] [nanjing]

文本2中的所有关键词为:[he] [live] [shanghai]

### (2) 建立倒排索引

有了关键词后,就可以建立倒排索引了。上面的对应关系是:“文本号”对“文本中所有关键词”。倒排索引把这个关系倒过来,变成“关键词”对“拥有该关键词的所有文本号”。一般仅知道关键词在哪些文章中出现还不够,还需要知道关键词在文章中出现的次数和出现的位置。加上“出现频率”和“出现位置”信息后,我们的索引结构变为(见下表):

关键词	文本号	出现频率	出现位置
He	2	1	1
I	1	1	4
Jack	1	1	1
Live	1, 2	2, 1	2, 5, 2
nanjing	1	2	3, 6
shanghai	2	1	3

以表中live行为例说明该结构:live在文本1中出现2次,在文本2中出现1次,它出现的位置为“2, 5, 2”。这表示什么呢?我们需要结合文本号和出现频率来分析, live在文本1中出现

了2次,那么位置中的“2,5”就表示live是在文本1中出现的第二和第五个关键词;live在文本2中出现1次,剩下的“2”就表示live是文本2中第二个关键词。

系统将上面3列分别作为词典文件、频率文件、位置文件保存。其中词典文件不仅保存每个关键词,还保存它在频率文件和位置文件中的指针,通过指针可以找到该关键词的频率信息和位置信息。词典中的关键词是按顺序排列的,因此可以用优化的二分查找算法快速定位关键词,并通过其附属的指针信息,找到该关键词出现次数和所有的位置信息。

为了有效地减小索引文件大小,需要对索引数据进行压缩处理。首先,对词典文件中的关键词进行压缩,关键词压缩为<前缀长度,后缀>。例如:如果当前词为“阿拉伯语”,上一个词为“阿拉伯”,那么“阿拉伯语”可压缩为<3,语>;其次是对数字的压缩,数字只保存与上一个值的差值(这样可以减小数字的长度,进而减少保存该数字需要的字节数)。例如当前文章号是17289(如不压缩,则要用3个字节保存),上一文章号是17282,可压缩保存为7来表示(只用1个字节保存)。

### 2.3 检索表达式

在海量的检索信息中,一般符合要求的结果信息甚少,但往往还要费劲地构思合适的检索表达式和对检索结果数据进行人工筛选。这也是目前文本信息检索工具普遍存在的问题——缺少智能化功能和难以用自然语言表达检索需求。检索需求一般要编写成合适的检索表达式,但目前大多数用户不善于或不喜欢费点心思和时间去构造准确、复杂的检索表达式。国内外搜索引擎研究学者的研究表明,大多数用户很少使用系统的高级搜索服务,且检索式的复杂程度,如像布尔逻辑符号的使用,对检索结果的影响并不明显。<sup>[3]</sup>因此,我们在设计系统的检索表达式时,强调和注重了简单型检索功能和尽量使检索表达式的构建自然语言化。

一个检索系统灵活性的高低和检索功能的强弱主要体现在检索表达式的处理上,基于语言学研究的需要,本系统提供与、或、非三种逻辑形式的表达式处理。用符号“^”表示非,“[]”表示与,“,”表示或。例如对于查询表达式:“[很]高兴[的,地]”,表示在查询的关键字串“高兴”的前面不出现“很”,在“高兴”的后面只出现“的”或“地”。另外,本系统支持两个关键字串的查询,可以设定两个关键字串前后间隔字符的个数或范围;支持复杂检索表达式(比如不相邻关键字查询,指定距离查询);支持对标点符号的查询(比如查询“?”可以检索语料库中所有疑问句);用户可定制查询结果的显示方式(如左右长度,排序等)等。

### 2.4 中文自动分词

中文全文检索的关键技术之一是分词的问题。目前,国

内已研制出的分词系统已有数十种之多,但由于此类系统大多都是针对专门领域或专门用途研制的,所以其识别精度在某些领域稍高,而在另外一些领域就降低了。针对这一现状,我们专门研制了一种可适应中文不同领域语料的面向大规模语料库处理的自动分词模块。

由于语言是一个开放集,其词语总是处在不断的变化之中,所以很难有一个十分完善的词典来描述它。所以,词典的完备性和可扩展性是我们自动分词时必须认真考虑和对待的一个问题。现有的大部分分词系统词典收录的都是些出现频率较高的常用词汇,而对一些专业领域的专有词语和术语则因受条件限制而收录的较少,由此就造成在处理相关领域的语料时,就会有大量未登录词出现的现象,从而造成分词精度的下降。对此,我们采用系统词典+用户词典的双词典可扩展性机制策略,较好地解决了这个难题。用户可以对系统词典、用户词典进行增、删、查阅等编辑操作,从而使系统的词典具有较强的通用性和扩展性。

我们的分词模块采用基于多步处理策略的中文自动分词及词性标注一体化方法。即首先对文本语句进行规整处理及提取待处理的语句。为了避免自动分词过程中出现切分盲点,保证句子中任何可能的切分都不会被后续处理所遗漏,我们对待处理的语句先进行全切分。在生成切分路径的过程中,采用基于有向图的歧义发现算法,找出所有可能歧义。切分歧义中最常见的一类歧义是交集型歧义,但由于大多数交集型歧义是属于伪歧义。所谓伪歧义是指包含交集型歧义但实际文本中仅有或几乎只有一种切分可能的字段。我们通过查交叉伪歧义表的方式直接确定切分形式,不再参与后续的分词处理过程。对于组合型歧义,由于大多数组合型歧义是从合的,因此对少量从分的组合型歧义采用规则进行处理。另外,某些词具有特殊的词性,不会与其它词产生交叉切分歧义,因而可以确定下来,如成语、惯用语、叹词、语气词等。汉语的数词是典型的单字词,数词串就是单字词串,如果句子中出现数词串,有必要将其合并,从而减少句子的切分结果数量,有利于后续的处理。汉语的重叠现象也需要单独处理,因为许多重叠词的存在会把属于同一词或同一节点的汉字分割开来,给后续的处理造成困难。通过这些手段,可以大大减少全切分生成的路径数量,然后通过分词及词性标注一体化处理,得出最佳切分及标注路径,最后利用规则进行切分及词性标注校正。

## 3 系统的分析与设计

一般来说,全文检索需要具备建立索引和提供查询的基本功能,此外还需要有简便的用户接口。在功能上,全文检索

系统的核心具有建立索引、处理查询、返回结果集、增加索引、优化索引结构等功能, 外围则由各种不同的应用功能组成; 在结构上, 全文检索系统的核心是索引引擎、查询引擎、文本分析引擎和对外接口等等, 加上各种外围应用系统等共同构成全文检索系统。<sup>[4]</sup>一个全文检索系统的性能优越程度, 根本上是由全文检索引擎来决定的。因此, 提升全文检索引擎的效率就是提升全文检索系统性能之根本。另一个方面, 一个优异的全文检索引擎, 在做到效率优化的同时, 还应该具有开放的体系结构, 以方便程序员对系统进行不断的优化改造。比如, 在当今多语言处理的环境下, 有时需要给全文检索系统添加处理某种语言或者文本格式的功能, 所以系统的开放性和可扩充性就显得十分重要。

一个优秀的系统设计必须具有合理的功能模块划分、清晰的接口定义、高度的抽象性和良好的可扩展性。按照这些要求, 我们将系统划分为3个层次、4个主要功能模块。

### 3.1 系统的层次

系统的3个层次分别为: 词法语法分析层, 系统核心接口层和存储层。词法语法分析层包括语言分析器和查询分析器, 负责对索引文档源和用户的查询命令进行词法语法分析, 通过扩展该部分可以适应不同语言的文档源和不同形式的查询命令。词法语法分析层对系统核心接口层屏蔽了不同的语言特性和不同形式的查询命令, 使得系统核心层可以提供一致简单的接口; 系统核心层则是系统中最重要的组成部分, 其中索引接口负责创建索引、访问索引和索引性能的优化, 而查询接口则负责分析查询命令, 并通过调用索引接口得到检索结果集; 存储层负责文档对象和索引对象的存储和维护多线程下访问的同步。

### 3.2 系统的主要模块

#### 3.2.1 语言分析器

语言分析器定义的是一个抽象的语言分析接口, 它负责把一个输入流中的字符串转换成一系列标记的集合, 这些标记将是建立索引的基本单位(项)。本系统提供了一个能够处理中英文的语言分析器, 对中文能以字、词以及词和词性混合三种模式为索引基本单位(项), 并可提供简单的中英文的停用词过滤器。系统可以通过扩展语言分析器来方便地实现对其它亚洲语言如藏文、日文的支持。

#### 3.2.2 查询分析器

查询分析器在接收到用户的查询命令后, 对查询语句进行词法和语法分析, 并生成同等语义的内部查询结构。一个查询语句可以包括若干个子句, 而一个子句可以是一个索引标记或另一个完整的查询语句。查询分析器并不对用引号括起来的短语做任何处理, 而是直接交给相应的语言分析器处理, 语言分析器把短语解析为一个或若干个标记。

#### 3.2.3 索引接口

每个索引包括若干个不同的索引段(或称为子索引), 而每个索引段包含下列信息: 字典集, 按字母顺序存放了所有的词语(索引标记Term), 以及该词语在词语频率集中的偏移量; 词语频率集, 存放了字典集中每个词语所出现的文档和频率以及该词语出现在位置集中的偏移量的三元组对; 词语出现位置集, 包含每个词语所出现的位置信息。

#### 3.2.4 索引的读写

索引的读写是全文检索系统中主要的性能瓶颈, 因此有效地对索引读写进行优化是全文检索系统至关重要的部分。系统使用了一种两级索引(索引和索引段)的方法, 尽量避免或减少对索引文件的修改。同时系统还把索引定义为一个抽象的目录结构并提供了两种不同的实现: 内存索引和磁盘索引。它们分别以相同的逻辑结构存储在内存和磁盘上, 在存储条件容许时使用内存索引将极大的提高索引操作的效率和性能。

## 4 系统特性介绍

本系统在研究、开发的过程中借鉴了国外开源的全文检索引擎工具包——Lucene, 吸收了其良好的系统组成结构。<sup>[5]</sup>在此基础上, 我们精心设计、实现了一个高度灵活的中文全文检索系统框架, 能够通过扩展来适应大规模的文本语料库的应用。在大规模全文检索应用中, 可以结合内存索引结构和分布式查询器(文件结构)两种技术实现海量数据的快速检索。

本系统的特性与创新之处:

(1) 索引文件格式独立于应用平台。系统定义了一套以8位字节为基础的索引文件格式, 使得在不同平台的应用所建立的索引文件能够共享;

(2) 在传统全文检索引擎的倒排索引的基础上实现了分块索引。能够针对新的文件建立小文件索引, 提升索引速度, 然后通过与原有索引的合并达到优化的目的;

(3) 优异的面向对象的系统架构, 使得系统易于扩展, 可以方便地扩充新的功能;

(4) 设计了独立于语言和文件格式的文本分析接口。索引器通过接受字节流完成索引文件的创立, 用户扩展新的语言和文件格式, 只需要对文本分析的接口进行扩展即可, 目前本系统只支持中文TXT格式文本。今后可通过扩展, 使其支持藏文、维文等少数民族语言和其它如HTML、RTF、PDF、DOC等文本格式;

(5) 系统提供了字索引、词索引和词及词性混合索引三种索引模式, 适用面和应用面广, 是一个通用的可适合于现代汉语任何领域文本的全文检索系统。其中字索引、词及词性混



合索引可支持古汉语的全文检索;

(6)本系统的查询方式灵活,功能强大,易用性较高。索引数据结构技术先进,而且较好地解决了检索结果大批量输出的瓶颈问题。

## 5 结语

本文对面向大规模语料库的全文检索系统的索引模式、检索算法、检索表达式的构建、自动分词、系统组成等进行了研究。并基于大规模语料库处理及应用的需要,编程实现了基于大规模语料库处理及应用的CIPP中文全文检索、自动分词、统计软件(注:此软件可到CIPP—中文信息处理平台网站:www.cipp.cn下载)。该系统具有全文检索、自动分词、语言统计等功能,其全文检索的速度在千万字数量级的大规模语料库中的平均检索时间小于1秒。

(上接第58页)

### 5.2 加强对教师队伍的信息素质教育

21世纪,知识经济初见端倪,随着信息技术的发展,世界范围内掀起围绕信息技术在教育中应用的教育改革热潮,新的教育信息技术不断地被运用于现实的教学活动中,对教育思想、观点、模式、内容和方法产生着深刻的影响。世界上越来越多的国家和地区认识到教育信息技术对于国家发展的巨大作用,在制定国家发展战略时把教育信息化作为重要因素加以考虑。在推进教育信息化进程中,世界各国都把教师的信息技术培训列为重要内容。

教师信息行为的发展最终应实现信息社会要求的教育理念。教师不仅是知识的传授者,也是学生学习活动的合作者和指导者。面对网络,教师和学生获取信息的手段上是平等的。此时,教师的职责显得更为重要,要能够帮助学生了解如何构建自己的认识结构,获得信息判断、信息筛选等方面的能力。教师的信息行为还将促使其教学形式、教学内容等发生变化。比如,可以进行虚拟教学实验,利用网络的特点开展虚拟教学的组织、实验实习等活动以及可以设计、编排教学计划与内容。教学质量的提高在某种程度上是教师信息行为导致的结果,而网络为教师的信息获取提供了一个广阔的空间。<sup>[9]</sup>

学校以及相关的教育机构应建立和完善对教师信息素质的培训,帮助教师认识和学习与教学相关的网络工具,组织教师进行网上学习。调查中发现,40岁以上的教师的信息化水平低于40岁以下教师的水平。所以,在对老师进行相关培训的时候,要有区别、有针对性地对不同年龄段的教师进行不同的信息化教育。同时,鼓励教师提高信息应用水平,并把学习的信息技能转化为实践,应用于他们平常的教学当中,这样才能在信息化的教学环境下更好地教学。

参考文献:

- [1] 黄昌宁,李娟子.语料库语言学[M].北京:商务印书馆,2002:101-104.
- [2] 曹元大,贺海军.全文检索索引技术的研究与实现[J].计算机工程,2002,(6):54-57.
- [3] 林建,张帆.搜索引擎的关键词检索策略[J].中国信息导报,2006,(8):55.
- [4] 贺胜.基于Lucene的中文全文检索系统[J].中国高校科技与产业化,2007,(6):142-143.
- [5] Lucene1.4 API[EB/OL].[2007-11-20].<http://Jakarta.apache.org/lucene/docs/index.html>.

作者简介:贺胜(1971-),男,博士,南京师范大学文学院讲师,研究方向:中文信息处理;卢亚军(1956-),男,西北民族大学语言文化学院教授,研究方向:藏语言文学,语料库语言学。

## 6 结语

本文通过问卷调查和文献调查法,对中学教师网络信息行为进行了调查,并对调查的结果进行了分析,总结归纳了教师网络信息行为的特点以及存在的问题。在网络环境下,教师在网络上进行与教学相关的网络信息行为,并把它作为辅助教学的现代化手段之一,这是提高教师信息素质和实现教育信息化的重要内容。随着社会的进步和发展,网络技术将会在教育领域担任更重的角色,开创现代教学的新纪元。

参考文献:

- [1] 王良成.网络环境下大学生信息需要与利用行为调查研究[J].情报科学,2002,(2).
- [2] 李书宁.网络用户信息行为研究[J].图书馆学研究,2004,(7).
- [3] 谢漫.近五年我国用户信息行为研究综述[J].广东青年干部学院学报,2006,(5).
- [4] 岳剑波.信息管理基础[M].北京:清华大学出版社,1999.
- [5] [EB/OL].[2008-02-20].<http://tech.sina.com.cn/other/2005-07-06/2053656122.shtml>.
- [6] 何晓丽.IT环境中的教师信息行为分析[J].宁夏大学学报,2003,(1).
- [7] Dietmar Wolfram. Search characteristics in different types of Web-based IR environments: Are they the same? [J]. Information Processing and Management, 2008, (3).
- [8] [EB/OL].[2007-7-20].<http://www.acm.org/sigir/forum/F99/Silverstein.pdf>.

作者简介:谭玮琼(1984-),女,中山大学资讯管理系研究生,研究方向:网络信息组织与传播。