文章编号:1006-2475(2008)08-0018-03

基于音码相似度的拼音模糊查询算法

阎红灿,张淑芬,谷建涛,阎少宏

(河北理工大学理学院,河北 唐山 063009)

摘要:论述了拼音模糊检索技术在信息管理和网络信息搜索系统中的必要性,描述了基于音码相似度的语言模糊查询算法及实现同音字和近音字检索算法,在中文信息检索中有很好的应用价值。并结合实例,在获得同音字数据库基础上,提出了基于音码相似度阈值的模糊查询算法,给出了通过拼音数据库实现中文全拼和首字母简拼检索数据库字段的实现方案,从查全率和查准率两个方面对算法的检索效果进行了评价,同时分析了音码相似度阈值对查全率和查准率的影响。

关键词:拼音字典; 音码相似度; 语音模糊查询; 同音字

中图分类号:TP311 **文献标识码**:A

An Arithmetic of Speech Fuzzy Query Based on Spelling Similarity

YAN Hong-can, ZHANG Shu-fen, GU Jian-tao, YAN Shao-hong

(College of Sciences, Hebei Polytechnic University, Tangshan 063009, China)

Abstract: This paper discusses the necessary of applying speech fuzzy query technique to information management system and Web information search system, describes the speech fuzzy query arithmetic and the method of realizing homophone or similar sound words query, this technique plays all-right role in information retrieval, and with examples, on the bases of obtaining homophone words database, gives the way of achieving full spelling or the first character of Chinese words, and further more, by the rate of full query and exact query, evaluates the query effect of this arithmetic, at the same time, analyses the influence of spelling similarity clique on the rate of full query and exact query.

Key words: spelling dictionary; spelling similarity; speech fuzzy query; homophone words

0 引 言

随着信息时代的到来和 Internet 技术的发展,查询已成为人们日常生活中不可缺少的部分。对于中文信息的查询,一般都是通过对字符进行比较、判断等方法来实现的,因此易于实现精确的汉字信息查询,即使模糊查询也只是对关键词的重新排列检索,没有实现真正意义的汉字模糊查询。然而,在中文信息管理系统或网络信息搜索系统中,用户需要一种拼音的模糊查询,如查找一个名叫“李明”的人,用户即使输入“黎明”、“李敏”或“李明韩”也能检索到要搜索的数据,也就是说,只要知道某一信息的部分读音或近似读音,并不知道汉字的具体写法,通过拼音检索就能把所有基本符合这个读音的记录内容全部显示出来,这就是拼音模糊查询技术。本文介绍的拼音模糊查询技术是指通过汉字拼音的查询,实现每一个汉字的同音和近音(或者谐音)查询。

汉语单字同音现象是非常严重的。以常用 6763 个汉字为例,没有同音字的汉字只有 16 个,其它汉字都有同音字,其中最多的有 116 个同音字[1]。拼音模糊检索技术的一个关键技术就是实现同音字的检索功能。笔者借用 Windows 系统下的输入法生成器,生成了一个文本文件的拼音字典,在此基础上构造拼音数据库,给出了实现同音字检索算法[2]。此算法基于拼音检索,检索成功率可达 100%,但需多次检索拼音数据库(随着关键字数的增加,扫描数据库的次数以指数级增长),时间消耗太大。

另外,由于地方口音的不同,或者其它原因,用户

收稿日期:2007-07-26

基金项目:河北省教育厅基金资助项目(0110052)

作者简介:阎红灿(1968-),女,河北保定人,河北理工大学理学院副教授,博士,研究方向:信息系统与信息工程,数据库与 Web 数据管理; 张淑芬(1973-),女,河北唐山人,副教授,研究方向:信息系统与系统工程; 谷建涛(1979-),男,河北唐山人,助教,硕士,研究方向:计算机应用; 阎少宏(1977-),男,河北唐山人,助教,硕士,研究方向:信息工程,离散曲面。

在查询中文信息时(特别是事务名称,包括人的姓名)输入与准确关键词读音相近的信息,如果系统也能够检索会大大方便用户查询。这是拼音模糊查询的第二个技术,文献[3]给出了一个通过函数判断实现姓名模糊查询的方法,但应用范围有限,而且检索效果不好。

1　基于音码相似度的拼音模糊查询算法

实现中文信息的拼音模糊查询,首先要找出每个汉字的同音字,其中拼音码最为关键。所以创建汉字拼音字典成为关键技术的第一步;要实现近音检索,必须有拼音的相似比较。笔者在文献[2]和[3]的基础上,经过多次实验研究,将两者有机结合,构造了基于音码相似度的拼音模糊检索算法,成功实现了中文信息的拼音模糊查询,而且将扫描拼音数据库的次数变为关键字的总数(降为常数级),大大减少了扫描次数,提高了检索效率。

1.1 同音字查询算法

利用操作系统提供的输入法生成器将微软全拼的码表文件 winpy. mb 逆转换成码表原文件 winpy. txt,码表原文件的头部是说明部分,正表部分将近 56000 行的内容中,每行由汉字及对应全拼音组成。码表原文件有如下特点:有的行描述单个汉字,有的行描述词组;有简体字,也有繁体字;描述单个汉字的行中,凡是多音字都由以空格隔开的两个拼音描述,且常用读音置于后面。如图 1 所示。

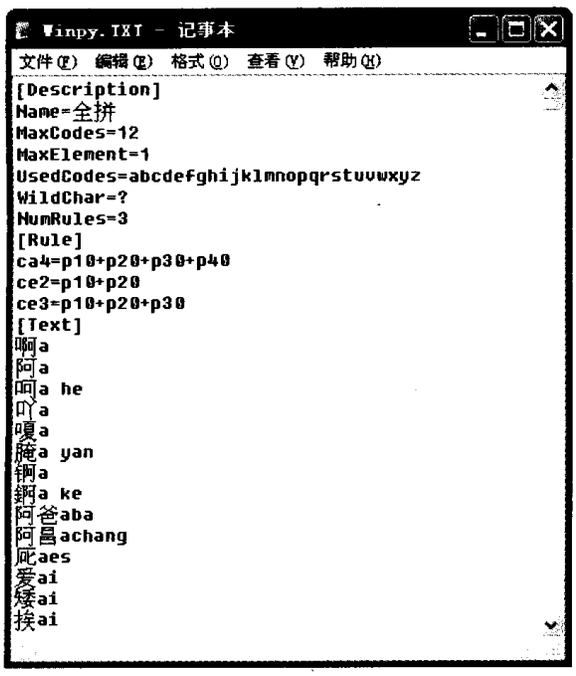

图 1　码表原文件

利用 VFP 的数据维护和导入、导出功能构造拼音数据表,经过二次转换生成 Access 数据库 PYDB. MDB[2]。拼音数据库共收录汉字单字 27954 个,如果只想保留常用单字,可以根据汉字内码和区位码提取国标一、二级简化字。原理如下:

GB2312-80 信息交换码表共收集汉字和图形符号 7445 个,将这些汉字和图形符号排列在 94 × 94 列的二维表中,表中的行称为区,列称为位,区和位构成汉字的区位码。常用汉字位于 16 至 55 区,按汉字的拼音顺序排列。计算机处理汉字以内码的形式表示,内码与区位码的关系:汉字内码 = 汉字区位码 + A0A0H。

有了拼音数据库,构造基于音码相似度的同音字检索算法如下:

第一步:根据用户输入关键字 Query,检索拼音数据库,形成拼音字符串 Query_String。如输入查询信息为“李克勤”,在拼音数据库中分别查找单字的拼音并连接成字符串“likeqin”。

第二步:对待查询字段 Field_String 计算音码相似度。

假定 Query 字符串长度为 L,chrs(i)为待查询字段 Field_String 中每个汉字的拼音字符串,position(i)为 chrs(i)字符串在 Query_String 字符串中的起始位置,position(i) =0 表示 chrs(i)不是 Query_String 的子串。相似度计算描述如下:

```
Count = 0
Similar = 0
For i = 1 to L
If position(i) > 0 then
Count = count + 1
Endif
Next I
Similar = count/L
```

例如:Query = "李克勤",Query_String = "likeqin",L = 3

Field_String = "李科秦",chars(1) = "li",chrs(2) = "ke",chrs(3) = "qin",

Position(1) = 1,position(2) = 3,position(3) = 5,

则 count = 1 + 1 + 1 = 3,similar = 3/3 = 1

第三步,定义相似度阈值 F。

根据多次编程实验,如果用户要求精确同音字检索,可以定义阈值 F 为 1;如果需要模糊信息,如检索关键词为“李克勤”,要求“李科”的信息也能检索出来,则定义阈值 F 为 0.5 比较合适。

第四步,扫描信息数据库,对每个待查询数据计算音码相似度,满足阈值的记录为检索数据。

1.2 近音字查询算法

近音字指读音相近的汉字,如“兰”和“来”等。有时在信息检索时需要这种检索功能。只需将同音字检索算法的第二步改变一下,就能完成用户需要的近音字检索能力。如将 chrs(i)定义为查询字段 Field_String 中每个汉字的拼音字符串中前 2 个字符,则“李米”和“黎明”的音码相似度为 1。用户可根据实际需求定义 chrs(i)字符串的长度来控制相似性,如只检测声母(第 1 个字符)。可见,chrs(i)的长度越小,则近音检索模糊度越大。

下面是笔者用 VB 实现近音字子串定位的主要语句。

```
temp = Left(Trim(chrs(i)), 2)
position(i) = InStr(Query_String, temp)
```

2 拼音模糊查询技术的应用

在一些特殊应用领域,如药品管理、餐饮管理等,用户往往需要快捷地输入检索,如果输入大量汉字信息,显得极为繁琐,于是汉语拼音识别[5]、分词研究成为关注热点,汉字拼音检索或汉字首字母检索成为备受青睐的检索方式。有了拼音数据库,汉字全拼和简拼检索实现就变得非常简单了。

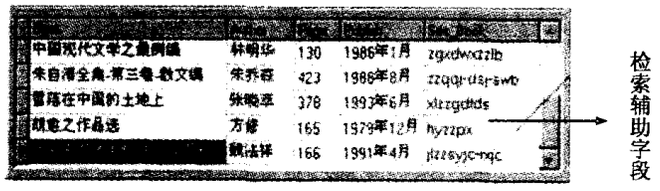

图 2 通过辅助字段实现汉字首字符检索

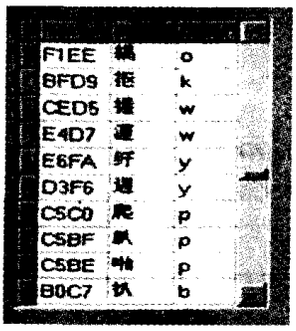

图 3 添加首字符的拼音数据库

在信息数据库中增加查询关键字的辅助字段,如图 2 所示。在信息入库时,将中文信息作为检索关键字,通过检索拼音数据库得到该信息的拼音,将该字段的拼音字符串或拼音首字符写入数据库[4],用户使用时,可直接键入汉语拼音进行检索,大大减少了用户输入量。

系统设计时,也可以提前在拼音数据库中增加一个单字首字符字段,如图 3 所示。将每个汉字的首字符取出放在该字段中,用户录入中文信息时,同时将每个汉字的首字符检索出来。

3 算法检索效果评价与结论

数据查询技术的检索效果主要通过两个评价指标来判定:查全率(Recall)和查准率(Precision)。查全率是指系统在进行某一检索时,检出的相关记录数与系统数据库中相关信息记录数总量的比率,它反映该系统数据库中实有的相关信息量在多大程度上被检索出来。

查全率 = [检出相关记录数/数据库内相关信息记录总数] × 100%

查准率是指系统在进行某一检索时,检出的符合要求的相关信息记录数与检出的数据记录总数的比率,它反映每次从该系统数据库中实际检出的全部数据记录中有多少是精确的。

查准率 = [检出相关信息记录数/检出数据记录总数] × 100%

本算法已成功地应用于公安民用信息系统和科技企业数据管理系统等需要快速实现模糊检索的数据库系统中,检索过程中检索效果同时受相似度阈值 F 的影响,通过各种数据测试,测试结果如表 1 所示。

表 1 系统测试效果表

测试项目(字段)	相似度阈值 F	查准率	查全率
姓名	0.9	99.2	90.8
企业名称	0.5	90.4	99.6

从表中 1 可以看出,相似度阈值 F 定义值越大,查准率越高,相反相似度阈值 F 定义值越小,系统查全率提高,降低了查准率。所以在实际应用中可以根据实际需求调节相似度阈值 F。一般情况下 F 在 0.4 ~ 1 之间。

本文的应用实例中,算法描述采用的是 VB 语言,其原理也可在其它语言平台下实现。

参考文献:

[1] GB12200.1-90,汉语信息处理词汇 01 部分:基本术语[S].

[2] 阎红灿,等. 语音模糊查询在信息管理系统中的实现[J]. 计算机系统应用, 2006(3):52-55.

[3] 孙威. 汉字姓名模糊检索的实现[J]. 警察技术,2001(5):27-28.

[4] 王克宇,莫祥银,王伟. 用汉语拼音检索数据库中的中文信息[J]. 南京师范大学学报(工程技术版), 2004,4(3):76-78.

[5] 陈立万. 基于语音识别系统中 DTW 算法改进技术研究[J]. 微计算机信息, 2006,22(5):267-269.

****

(上接第 17 页)的概念,在语言的符号集中引入了新的等价关系,提出了正则语言的代数判定定理。应用该定理可以判断某一给定语言的正则性和非正则性,并且证明过程更加简练。

参考文献:

[1] 叶瑞芬,沈百英. 正则语言的特性性质[J]. 软件学报, 1995,6(7):416-419.

[2] 叶瑞芬,沈百英. 关于正则语言的泵引理[J]. 华东理工大学学报,1994,20(5):654-656.

[3] Michael Sipser. Introduction to the Theory of Computation [M]. Beijing: China Machine Press, 2000:2-50.

[4] John E Hopcroft, Rajeev Motwani, Jeffrey D Ullman. Introduction to Automata Theory, Languages, and Computation (Second Edition) [M]. Beijing: China Machine Press, 2004:1-112.

[5] Downey R G, Fellows M R. Parameterized Complexity[M]. Springer-Verlag, New York, 1997.