基于拼音首字母查询的去重优化设计

王　东[1]，熊世桓[2]

（1.2. 贵州师范学院，贵州 贵阳　550018）

摘要：分析了现有信息系统中检索数据时直接输入检索词的不足，提出了基于汉字拼音首字母的查询方法，以汉字拼音首字母串替代检索词，并针对重码问题进行了分析研究，得出一种可行的编码方案。该方法对检索词是姓名、地名、商品名等专用名的查询具有很好的实用价值，能显著提高查询效率。

关键词：汉字拼音；中文检索；编码

中图分类号：TP311.13　　**文献标识码**：A　　**文章编号**：1674－7798(2010)06－0037－03

The optimized design for reducing the re－code based on the Pinyin first letter inquiry

WANG Dong[1], XIONG Shi－huan[2]

(1.2. Guizhou Normal College, Guiyang, Guizhou, 550018 China)

Abstract: The insufficiency in retrieving the data by inputting inquiry keywords directly in the existing information system has been analyzed. An inquiry method based on the first letter of Chinese character Pinyin has proposed, which substitutes the method of inputting Chinese character directly by the first letter of Chinese character Pinyin, and conducts the analysis research aimed at its repetition－code. We obtaine one kind of feasible code scheme. This method has the very good practical value for inquiring keywords. For example: the name, the geographic name, business commodity name and so on. It can enhance the inquiry efficiency obviously.

Key words: Chinese character Pinyin; Chinese retrieval; code

查询数据是信息系统中一项重要频繁的操作功能，尤其对包含大量数据的系统，查询时尽量减少用户查询的输入量，选择简单、快捷的输入方法，提高检索效率是系统设计必须考虑的问题。

一般情况下，输入查询关键字的一般过程可以被描述为：选择输入法→输入查询关键字→查询，这是传统的查询方式。当查询数据的实时性要求较高时，便显露出它的缺陷：一是需要在输入法选择之间来回切换；二是输入法的击键次数都较多。虽然现有的汉字输入法都提供了词组输入，但对人名、地名及一些比较专业的名称，输入时基本上还是单字输入；三是查询速度严重依赖于用户所采用的汉字输入法和用户对输入法的熟练程度。

针对以上问题，一种常见而有效的方法是使用汉字拼音首字母来取代直接输入汉字。然而，该方法又导致了重码灾难，用户需要在过多的重码中进行二次选择，降低了方法的实用性。如何设计有效的识别码，减少重码率是值得研究的重要问题。本文中，我们提出了一种简单而实用的方法。

1　汉字编码

汉字在计算机中表示为内码，对不同的操作系统，汉字内码可能有所不同。GB2312－80是1980年由中国国家标准局颁布的《通用汉字字符集及其交换标准码》，它共收入汉字、字母、符号等7 445个，其中汉字6 763个，包括一级汉字3 755个，二级汉字3 008个。

在机器内码中，汉字的排列顺序为：一级汉字3 755个按拼音字母顺序排列；二级汉字3 008个按部

收稿日期：2010－03－20

作者简介：1. 王　东(1978－)，男，汉族，贵州师范学院讲师，主要研究方向：中文作信息处理；
2. 熊世桓(1973－)，男，贵州师范学院副教授，主要研究方向：网络通信及软件工程。

首顺序排列。对于一级汉字,通过判断区位码位于一级汉字内码的范围即可确定它的拼音首字母。例如:区位码大于等于1601,小于1637的汉字拼音首字母为"A"。对二级汉字,由于其顺序是按汉字笔画结构排列的,确定汉字的拼音首字符不能采用以上方法。可通过系统内置输入法提供的API函数反查出汉字的拼音,从而得到各个汉字的拼音首字母。将所有二级汉字顺序获得的拼音首字母连接成二级汉字拼音首字母串存储起来,之后根据二级汉字的区位码,可以计算其拼音首字母在拼音首字母串中的位置。

2 重码分析与处理

在一些面向专用名检索的应用中,输入检索词中各字的拼音首字母替代直接输入各个汉字具有简单、易学的优点,不需要用户掌握特定的输入法。但重码问题又降低了该方法的实用性。因此,去重研究是确保上述方法实用化的必要条件。

降低重码率直接而有效的方法是增加识别码。一般情况下,识别码设计有如下的原则。第一,识别码生成规则应简明、易掌握;第二,识别码生成规则要尽量和原有编码规则保持同一性;第三,识别码不易过长。本文根据汉字构成的特点,提出了一种简单而实用的去重方法。为了叙述的方便,先给出如下定义。

定义1: 汉字部件,由笔画组成的具有组配汉字功能的构字单位。

定义2: 汉字成字部件,汉字的组成单元中,拆分后能独立成字的部件。如'岸'字,成字部件有:山、厂、干、岸。

定义3: 汉字首位成字部件,按汉字的书写顺序进行拆分后,得到的第一个成字部件。如警字的首位成字部件有:口、句和苟。

定义4: 汉字首位最大成字部件。按汉字的书写顺序对汉字进行拆分后,得到的第一个最大(笔画最多)成字部件。如警字的首位最大成字部件为苟。

在获取汉字的首位最大成字部件时,拆分汉字还应满足以下规则:

1)笔画交叉重叠的,不拆分。如'串'字、'东'字。

2)拆开后的各部分均为非成字部件或均不再构成其他汉字的,不拆分。如'非'字。

3)规定没有成字部件的汉字的首位最大成字部件就是它本身。

具体操作时,将汉字与首位成字部件的对应表存储在数据库中,如表1。

表1 汉字与首位成字部件的对应表

汉字	首位成字部件	首位最大成字部件
岸	山	山
警	口、句、苟	苟
靓	月、青	青
局	尸	尸
……	……	……

由汉字串到编码串(拼音首字母串)转换的过程看作是一种映射过程,ζ为映射的规则,S表示汉字串,C表示编码。

汉字串到编码的转换过程表示为:

$$\zeta(S) -> C \qquad (1)$$

重码被描述成给定字串$S_1, S_2, \cdots, S_n, n >= 2$,对任意的$S_i \neq S_j (1 < i, j <= n$且$i \neq j)$,$\zeta(S_i) = \zeta(S_j)$。为了减少产生重码的几率,将原汉字串进行扩展加长(文中称为扩展汉字串),再将扩展汉字串转换成编码。

用τ表示扩展加长规则,定义为取原汉字串首尾两字的首位最大成字部件组成的汉字串。

S的扩展汉字串可表示为:$S + \tau(S)$。

为了区分生成的编码中哪部分是原字串产生的,哪部分是由加长字串产生的,编码中用空格来分隔这两部分,即:

S的扩展汉字串为:$S + " " + \tau(S)$

原汉字串S到编码的转换(1)式$\zeta(S) -> C$变为:

$$\zeta(S + " " + \tau(S)) -> C \qquad (2)$$

我们用10 123个人名库为实验数据,对上述方法进行了分析和验证。在26个字母中,'i','u','v'不能成为拼音首字母,拼音声母为'z','s','c','zh','sh','ch'的拼音首字母归为'z','s','c',有23个字母成为编码单元。考虑2字长姓名和3字长姓名,理论上,对2字长的姓名共有232种,即529种,三字词可以产生233种,即12167种。进行重码处理后,二字词有234种编码,即279 841种,三字有种235编码,即6436 343种,编码空间显著增大。对实测数据平均重码数的统计结果如下表。

表2 实测数据统计表

		总数	平均重码数
二字	①	5623	18.5
	②	5623	1.06
三字	①	4500	5.1
	②	4500	1.01

注:二字表示二个汉字组成的姓名,三字表示三个汉字组成的姓名。①表示直接以每个字的拼音首字母为编码,②表示按(2)式得到的编码。

可以看出,二字词的平均重码数由18.5降低至1.06,三字词的平均重码数由5.1降低至1.01,当数据量不大时,编码和人名可看成是一对一的关系。以上从理论和实际应用两方面验证了该方法的有效性。

当然,为解决重码将原字串进行扩展再生成编码,有效降低了重码率,同时也增加了编码长度。但在实际的应用系统中,检索时采用的是动态过滤的方法,随着编码中每个字母的输入,重码的数量将急剧下降,因此在实际应用过程中,并不需要将编码全输入完毕就能检索出查询结果。

3 算法描述

①返回给定汉字串的扩展汉字串

函数:ExteedString(ByVal WordString As String) As String

```
Dim i As Integer
Dim str As String
'从表 BJTB 中获得每个汉字首位最大部件
For i = 1 To WordString. Length
Dv = db. runselectSQL ("select bj from BJTB where word = '" + Mid(WordString, i, 1) + "'")
Str + = dv(0)(0)
Next
Return WordString + " " + str
```

②返回给定字符串的首字母

函数:IndexCode(ByVal WordString As String) As String

```
Dim i As Integer
'获取给定字符串的拼音首字母串
For i = 1 To WordString. Length
Dim str As String = GetOneIndex(Mid(WordString, i, 1))
str + = str
Next str
```

③返回单个字符的首字母

函数: GetOneIndex (ByVal OneIndexTxt As String) As String

```
如果是标点、数字、字母,返回原字符
If Asc(OneIndexTxt) > = 0 And Asc(OneIndexTxt) < 256 Then GetOneIndex = OneIndexTxt
Else
Dim str As String
str = GetX(CInt(Format((Asc(OneIndexTxt) + 65536) \ 256 - 160, "00") & Format((Asc(OneIndexTxt) + 65536) Mod 256 - 160, "00")))
Return str
End If
```

④根据区位得到首字母

函数:GetX(ByVal GBCode As Integer) As String

```
'通过区位码,判断一级汉字
If GBCode > = 1601 And GBCode < 1637 Then
GetX = "A"
If GBCode > = 1637 And GBCode < 1833 Then
GetX = "B"
      ……
'通过区位码,计算在拼音首字母串中的位置,判断二级汉字
If GBCode > = 5601 And GBCode < = 8794 Then
Dim CodeData As string
CodeData = "cjwgnspgcenegypbtwxzd…"
GetX = Mid(CodeData, (Left(CStr(GBCode), 2) - 56) * 94 + (Right(CStr(GBCode), 2)), 1)
End If
```

4 结束语

本文针对直接用汉字拼音首字母替代检索词重码多的问题,给出一种简单、易行的汉字部件拆分方法,通过扩展检索词长度,加长编码已达到消解重码的目的。实测数据证明方法是高效、实用的。上述方法已在"贵州师范学院报名系统"中进行了运用,取得了较好的效果。

参考文献:

[1]中国标准出版社. 字符集和信息编码国家标准汇编[M]. 中国标准出版社,2004.

[2]孙宏凯, 王彦勋. 中文数据排序与快速检索方法研究[J]. 微计算机信息 , 2007,(03) .

[3] 中国语言文字网. http://www. china - language. gov. cn.

[4] 张炜, 唐慧强. 将汉字转化为拼音的研究与实现[J]. 计算机应用 , 2003,(S1).