

一种按拼音首字母对数据库中的汉字进行检索的方法研究

刘 辉

(西安理工大学 西安 710082)

摘 要 根据 SQL Server 的排序规则,在不改变查询表结构的条件下,探讨并研究了在 SQL Server 的数据库服务器端直接按汉字的拼音首字母对汉字进行快速检索的方法。

关键词 SQL Server,拼音首字母,快速检索

中图法分类号 TP312 **文献标识码** A

Method of Chinese Characters Retrieval According to Pinyin Initials in Database

LIU Hui

(Xi'an University of Technology, Xi'an 710082, China)

Abstract According to the collation of SQL Server without changing the structure of the query table conditions, this paper discussed and studied fast retrieval methods according to the first letter of alphabet characters in a SQL Server database server.

Keywords SQL server, Pinyin initials, Quick search

在存储有中文信息的信息系统中,经常需要对汉字进行检索,这个操作重要且频繁。若在输入汉字信息时能以汉字的拼音首字母代替^[1],则可以避免用户输入汉字时所进行的输入法切换,很大程度地提高使用效率。本文就 SQL SERVER 中在不改变数据库表结构、不增加附加信息的基础上,按顺序输入汉字拼音首字母的全部或部分,实现快速查询的方法进行了探讨。

实现按拼音首字母对汉字进行检索的方法,通常有以下几种。

方法一:改变查询表结构,在表中增加一个辅助字段,预先进行词条翻译并存储汉字的拼音首字母^[2]。这种方法的缺点是前期的翻译和录入工作耗时且极大地增加了对存储空间的需求,尤其是在表中已经存在有大量数据时增加这个辅助字段几乎得不偿失。

方法二:根据汉字的读音设置 26 个监视点,每个监视点是一个英文字母及对应读音的第一个汉字^[3],在查找时把拼音首字母转换成与之相对应的第一个汉字与最后一个汉字所在的区间,如字母 B 所对应的汉字区间在“八”和“簿”之间。由于拼音字母只有 26 个,因此检索速度可得到极大地提升。

本文旨在探讨研究方法二,在它的基础上给出了一个实用的运行于数据库服务器端的存储过程。

1 用 SQL Server 的排序规则取得汉字拼音首字母表

在 SQL Server 中,字符串的物理存储由排序规则控制。排序规则指定表示每个字符的位模式以及存储和比较字符所使用的规则^[4]。

SQL Server 的排序规则用“collate Chinese_PRC_CS_AS

_KS_WS”指定,其中前缀“Chinese_PRC”是针对大陆简体字 UNICODE 的排序规则,后半部份的含义如下。

_BIN:二进制排序。

_CI(CS):是否区分大小写,CI 不区分,CS 区分。

_AI(AS):是否区分重音,AI 不区分,AS 区分。

_KI(KS):是否区分假名类型,KI 不区分,KS 区分。

_WI(WS):是否区分宽度,WI 不区分,WS 区分。

由于目前 UNICODE 字符集共收录汉字 20902 个,简体 GBK 码汉字 UNICODE 值从 19968 开始,因此不用字典,使用以下 T-SQL 命令可以得到所有汉字并按拼音顺序排序。

```
create table #temp(code int,CNWord char(2))
declare @row int,@i int
select @row=1,@i=19968
while @row<=20902
begin
    insert #temp values(@i,nchar(@i))
    select @row=@row+1,@i=@i+1
end
select * from #temp order by CNWord
collate Chinese_PRC_CI_AS,code
```

按字母查找时不区分大小写,只区分重音。经过分析临时表 #temp 的内容,可以发现读音为“A”的第一个字为“吖”,最后一个字也即读音为“B”的第一个字为“八”,读音为“K”的第一个字为“咔”,最后一个字也即读音为“L”的第一个字为“垃”,以此类推。使用以下 T-SQL 命令可以产生一个汉字拼音首字母表。注意字母“I”、“U”、“V”没有对应汉字的拼音首字母,则使用其后面字符所对应汉字的拼音首字母。

```

CREATE TABLE pycx
(
    letter char(1) primary key,
    hz1 nchar(1) collate Chinese_PRC_CI_AS,
    hz2 nchar(1) collate Chinese_PRC_CI_AS,
)
go
insert pycx
select 'A',N'吖',N'鰲' union all
select 'B',N'八',N'簿' union all
select 'C',N'嚓',N'錯' union all
select 'D',N'咻',N'鵓' union all
select 'E',N'婀',N'械' union all
select 'F',N'发',N'鰻' union all
select 'G',N'咎',N'髌' union all
select 'H',N'姣',N'杏' union all
select 'J',N'开',N'揀' union all
select 'K',N'呿',N'髻' union all
select 'L',N'垃',N'纒' union all
select 'M',N'妈',N'螯' union all
select 'N',N'拏',N'糯' union all
select 'O',N'噢',N'漚' union all
select 'P',N'娉',N'曝' union all
select 'Q',N'七',N'褰' union all
select 'R',N'咄',N'鵓' union all
select 'S',N'仨',N'蟀' union all
select 'T',N'他',N'籐' union all
select 'W',N'缶',N'鵲' union all
select 'X',N'夕',N'鏢' union all
select 'Y',N'丫',N'韻' union all
select 'Z',N'市',N'咭'

```

2 按拼音首字母实现检索算法

人的姓名、通讯地址、书名、课程名等是数据库中常用的中文字段,以具有两个列(课号 CNo、课名 CName)的课程表 Course 为例,如果需要对含有汉字信息的课程名检索,查找第一个字发“W”音的课程,则表示课程名的第一个汉字应位于“缶”和“鵲”之间,根据字符“W”查找对应汉字的 SQL 语句如下:

```

declare @hz1 nchar(1),@hz2 nchar(1)
select @hz1=hz1,@hz2=hz2 from pycx where letter='W'
select * from Course where substring(CName,1,1) between @hz1
and @hz2

```

利用这种思路,当对于给定的一段文字的拼音首字母@letters,在以@bm为名的表中按以@zdm为名的列进行查找时,实施查询的存储过程如下:

```

create proc up_SearchOnPinYin
@bm varchar(50),@zdm varchar(50),@letters varchar(100)
as
declare @LEN int,@i int,@let nchar(1),@str varchar(max),@s
varchar(max),@hz1 nchar(1),@hz2 nchar(1),declare @b varchar
(100)

```

```

set @len=len(@letters) --计算拼音首字母的个数
--把查询表中的初始数据放入中间结果表 temp 中
select @b=convert(varchar(100),getdate(),8)
set @str='select * into [temp'+@b+'] from '+@bm+' where
len('+@zdm+')>=cast(@len as varchar(max))'
exec(@str)
set @str='select * into [t'+@b+'] from [temp'+@b+']'
exec(@str)
set @i=1
while @i<=@len
begin
    set @str='delete from [t'+@b+']'
    exec(@str)
    set @let=substring(@letters,@i,1)
    select @hz1=hz1,@hz2=hz2 from pycx where letter=@let
    --先处理非汉字字符
    set @s='substring('+@zdm+',',cast(@i as varchar(max))
    +',1)'
    set @str='insert into [t'+@b+'] select * from [temp'+@b
    +'] where unicode('+@s+')<19968 and '+'@s+'='"+@let
    +"'"
    exec(@str)
    --再处理汉字字符
    set @str='insert into [t'+@b+'] select * from [temp'+@b
    +'] where unicode('+@s+')>=19968 and '+'@s+' between "'
    +@hz1+"'" and "'"+@hz2+"'"
    set @str=@str+' delete from [temp'+@b+']'
    set @str=@str+' insert into [temp'+@b+'] select * from [t'
    +@b+']'
    exec(@str)
    set @i=@i+1
end
set @str='select * from [t'+@b+']'
set @str=@str+'drop table [temp'+@b+']'
set @str=@str+'drop table [t'+@b+']'
exec(@str)

```

结束语 利用本方法可方便地按拼音首字母完成查询,如需要对 Course 表的 CName 列查询 C 语言相关课程,可用如下命令:

```
exec up_SearchOnPinYin 'course','cname','cyy'
```

本算法可对任意表的任意列用拼音首字母实现查询,并且包含繁体汉字和不常用的地名汉字。本算法已经在我校的学籍管理系统中得到了应用。

参 考 文 献

- [1] 张嘉惠,崔超艳.一种汉语拼音首字母快速查询数据库的方法[J].电脑知识与技术,2010,6(15):3874
- [2] 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):76-77
- [3] 毛先成,胡敬文,刘志强.基于 WebGIS 的中文属性模糊匹配服务引擎的实现[J].地理空间信息,2008,6(4):1-2
- [4] Date C J. 深度探索关系数据库[M].熊建国,译.北京:电子工业出版社,2007