拼音检索方法在 Web 系统中的研究与实现①

刘风华[1], 陈燕红[2], 郑卫斌[3]

[1](新疆工程学院 计算机工程系, 乌鲁木齐 830091)
[2](新疆农业大学 计算机与信息工程学院, 乌鲁木齐 830052)
[3](西安交通大学 电子与信息工程学院, 西安 710049)

摘 要: 在使用标准化代码的系统中, 为解决新增记录精确匹配代码的问题, 系统中采用了拼音检索方法, 拼音检索主要应用在就业信息采集页面中, 对采集的数据进行模糊匹配, 并将匹配到的数据加载到页面中由用户自行选择到最符合要求的数据, 为了提高匹配精度, 在系统中采用双重模糊查询方法, 解决了系统中有大量待查数据时查询效率与查询精度的问题, 该系统投入使用后收到了良好的效果.

关键词: 拼音检索; 模糊查询; SQL; 正则表达式

Research and Implementation of Phonetic Retrieval Methods in Web Systems

LIU Feng-Hua[1], CHEN Yan-Hong[2], ZHENG Wei-Bin[3]

[1](Department of Computer, Xinjiang Institute of Engineering, Urnmuqi 830091, China)
[2](School of Computer and Information, Xinjiang Agricuitural University, Urnmuqi 830091, China)
[3](School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: At present, most of the management information system is the B/S mode, so users can access and manipulate data in the network terminal. Standardized coding provides a convenient for the data statistics and query operations in database management. A novel fuzzy matching method was proposed, which uses Pingyin to fuzzily match the retrieved data from database. The method was used in the employment information gathering page, and the data was filtered twice. It would improve the web searching efficency and accuracy when there were massive data yet to be checked. The system was accomplished and deployed in our school, and had made good achievements.

Key words: Pinyin search; fuzzy query; structured query language; regular expressions

随着现代信息技术特别是网络技术的迅速发展和逐步深入的应用, 通信网络越来越发达, Internet 规模越来越大, Web 已经成为全球传播与共享科研、教育、商业和社会信息等最重要和最具潜力的巨大信息源. 各行各业为更好的管理和使用这些信息开发各种管理信息系统, 目前, 大多数管理信息系统都是 B/S 模式, 这样用户可以在网络终端进行访问和操作, 因此标准化编码成为了系统数据规范的必要条件, 而且在数据库管理中, 有了统一的编码, 为数据统计、查询等操作提供了方便.

1 标准化代码系统中数据录入问题

编制统一的代码为数据的管理提供了方便, 但是当有新的信息需要录入到系统中时, 录入的信息必须非常准确才能找到正确的匹配代码, 以高校毕业生管理系统为例, 在单位代码库中有 986 条数据, 生源所在地代码中有 3380 条数据, 如果通过人工查找效率会非常低且易出现错误. 于是, 在信息录入中, 如何能找到对应的准确的信息是一个重要问题. 一般的解决方法是:

(1) 当记录较少时, 可以采用下拉式列表的方式将

① 基金项目:国家科技支撑项目(2008BAH37B04);中央高校基本科研业务费专项资金
收稿时间:2012-06-18;收到修改稿时间:2012-07-11

可能要用的数据列出来, 然后由用户自己选择.

(2) 当记录量比较大时, 给用户一个代码对应表, 自己查找后对应填写.

以上两种方法均不满足在数据量较多时用户方便快速准确录入信息的原则.

2 需要解决的问题和解决方法

因为代码库中数据量大, 如何才能高精度的匹配到对应代码是关键问题, 为了解决该问题, 决定采用拼音检索信息的方式来完成, 这样做有以下三个原因:

(1) 避免不准确信息出现, 如"新疆乌鲁木齐市沙依巴克区"在库中对应的编号是"650103", 如果学生按习惯称呼输入"乌市沙区", 系统则无法完成由名称到代码的转换(最后要求收集的全部是代码).

(2) 避免同音错字, 在使用汉字进行查询时, 大部分学生是使用拼音输入法, 这样经常出现同音不同字的错误, 如果输入的汉字有误, 则无法查询到想要的数据.

(3) 减少字符输入, 学生普遍喜欢使用各种拼音输入法, 录入汉字时, 也是先输入拼音再选择到需要的汉字, 而采用拼音检索时就大大简化了输入的字符数.

如我校毕业生在填报相关信息时, 需要填写专业代码、学校所在地代码、生源地代码等多个信息, 以填报院校所在地代码为例, 新疆工程学院位于新疆乌鲁木齐市沙依巴克区, 在填写"院校所在地代码"中, 点击对应的按钮, 系统会弹出一个输入拼音的对话框, 学生可以按照输入法智能 ABC 的习惯只输入拼音首字母如"xjwlmq", 然后点击"检索"按钮, 网页中会显示所有拼音中包含 xjwlmq 字符, 并将模糊匹配的结果显示在列表中, 如图 1 中显示的是学生在拼音框中输入"xjwlmq"并点击"检索"后返回的结果. 学生只需要将要选的内容双击便可填写到文本框中, 这样选择的信息就可以准确的与库中的代码对应, 从而可以收集到准确的代码信息.

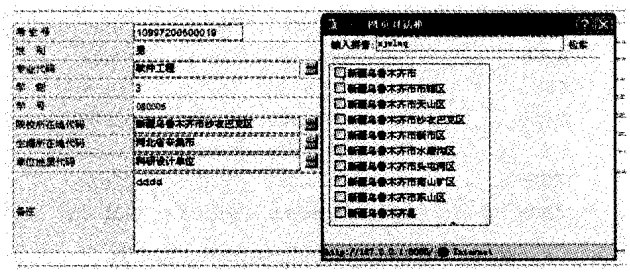

图 1 拼音检索示例

3 拼音检索方法在系统中的实现

3.1 生成拼音检索库

要实现拼音检索[1], 检索系统必须知道每个汉字的拼音, 这就需要有汉字拼音对照表. 利用操作系统的输入法生成器将微软全拼的码表文件 winpy.mb 逆转换成码表源文件 winpy.txt, 通过编程把 winpy.txt 中汉字与拼音的对应关系提取出来并存入数据库即得到了汉字拼音对照表, 用户只要输入包含拼音字符中的部分字符, 就可以查找到对应数据.

code_no	code_meaning	meaning_pinyin
10109	昌吉回族自治州人事局	changji2hui2zu2zi4zhi4zhoulren2shi4ju2
10110	哈密地区人事局	halmi4di4qulren2shi4ju2
10111	吐鲁番地区人事局	tu3lu3fan1di4qulren2shi4ju2
10112	巴音郭楞蒙古自治州人事局	balyinlguolleng2meng3gu3zi4zhi4zhoulren2shi4ju2
10113	阿克苏地区人事局	alke4sudi4qulren2shi4ju2
10114	喀什地区人事局	kalshen2di4qulren2shi4ju2
10115	克孜勒苏柯尔克孜自治州人事局	ke4zille4sulkeler3ke4zi1zi4zhi4zhoulren2shi4ju2
10118	和田地区人事局	he2tian2di4qulren2shi4ju2
20000	区直单位小计	quzhi2danlwei4xiao3ji4
20101	新疆煤炭工业管理局	xinlijianglmei2tan4gonglye4guan3li3ju2
20102	新疆机械电子工业行业管理办公室	xinlijiangljilxie4dian4zi5gonglye4xing2ye4guan3li3ban4gonglshi4
20103	自治区交通厅	zi4zhi4quljiaoltongltingl
20104	新疆轻工业行业管理办公室	xinljianglqinglgonglye4xing2ye4guan3li3ban4gonglshi4
20105	新疆石油化工行业管理办公室	xinlijianglshi2you2hua4gonglxing2ye4guan3li3ban4gonglshi4
20106	新疆药业集团公司	xinlijianglyao4ye4ji2tuan2gonglsil
20107	新疆纺织工业行业管理办公室	xinlijianglfang3zhilgonglye4xing2ye4guan3li3ban4gonglshi4
20108	新疆建材行业管理办公室	xinlijiangljian4cai2xing2ye4guan3li3ban4gonglshi4
20109	新疆机械设备成套局	xinlijiangljilxie4she4bei4cheng2tao4ju2
20110	新疆国际经济合作公司	xinlijianglguo2ji4jingljilhe2zuo4gonglsil
20111	南疆煤电公司	nan2jianglmei2dian4gonglsil

图 2 数据表中代码、中文与拼音对照表

3.2 模糊查询方法

(1) SQL 中的模糊查询

SQL[2](Structured Query Language)是目前使用的最广泛的结构化查询语言, 在 SQL 中提供了 LIKE 子句进行模糊查询, 简单的模糊查询语句格式如:

SELECT <列> FROM <表> WHERE <字段名> LIKE <条件>

其中LIKE 后的条件, 可以是四种通配符的任意组合形式实现数据在某范围或不完整字符串的模糊查询.

(2) 正则表达式

正则表达式[3]是用于模式匹配的专用语言, 针对不同的模式正则表达式可以灵活地构造不同的专用语言来实现模式匹配, 正则表达式[4,5]是一个用来描述或者匹配一系列符合某个模式的字符串的单个字符串, 通过正则表达式将字符串中符合某一模式的文本进行查找和替换.

3.3 模糊查询中遇到的问题

系统中需要频繁调用数据, 每一次查询都涉及到表中的大量数据, 因此在模糊匹配时遇到了问题:

(1) 数据量太大, 当信息加载到查询页面时响应速度过慢, 甚至造成死机;

(2) 模糊匹配精度不高, 查询结果中模糊匹配到

的数据集较大, 用户无法快速选到自己需要的信息.

在前面介绍的两种模糊查询方法中, 对于任何一种查询方法来说, 当数据库中有大量待查询数据时, 效率都比较低, 单独使用一种方法不能很好的解决查询精度和查询效率的问题.

3.4 系统中模糊查询算法

在拼音检索框中输入要查询的信息的拼音, 将输入的内容定义为字符串 String1, 以输入 xjwlmq 为例, 对应表中的值定义为字符串 String2, 其中 String2 是包含了 String1 中部分字符可能出现的结果集, 查询的目的是让 String2 结果尽量的精确到用户需要的数据, 达到快速准确选择的目的. 具体算法如下:

① 把字符串 String1 分解为若干个有效字符 S1,S2...Sn; 如将 xjwlmq 分解为 x、j、w、l、m、q.

② 筛选出 String1 分解后的单个字符包含在被查找的字符串 String2 中的数据; 此时 String2 是一个庞大的数据集, 只要包含了 x、j、w、l、m、q 中任意一个字符的数据都被包含在了 String2 中.

③ 在筛选过的数据中, 将按照 S1,S2 排序的 String2 筛选出来, 定义为模糊集合 A,再在模糊集 A 中筛选按照 S1,S2,S3 排序的数据,定义为模糊集合 B, 再在 B 集合中依次查找按照 S1,S2,S3,S4 的, 以此类推, 最后得到按照 S1,S2...Sn 排序的数据集合; 结合前面的例子, 首先查找出按照第 1 个字符是 x, 第 2 个字符是 j 的模糊集合 A,再在 A 集合中查找按照第 1 个字符是 x, 第 2 个字符是 j, 第 3 个字符是 w 的模糊集合 B, 以此类推, 最后查找到按照 x、j、w、l、m、q 字符顺序的数据集合.

④ 将按照 S1,S2...Sn 排序的数据集合返回给用户; 数据库将包含了字符序列为 x、j、w、l、m、q 的数据对应到的记录返回给客户端.

⑤ 用户在最后匹配的数据集中找到需要的准确数据, 然后记录下该数据对应的代码写回数据库中.

在系统中, 筛选出 String1 分解后的单个字符在字符串 String2 中的数据, 此过程使用 SQL 中的 LIKE 模糊查询, 由于 SQL 中的"%"通配符代表了任意多个字符, 因此数据集中时包含任意单个字符的数据都被筛选出来, 但是由于数据库服务器反应快速, 所以可以进行初步筛选. 为了精确数据集, 需要将初次查询结果再按字母出现顺序进行模糊匹配, 此过程使用正则表达式匹配, 由应用服务器端完成, 这样两次模糊查询分别分担在数据库服务器端和应用服务器端, 适当的分担了服务器的负载, 提高了查询效率.

3.5 系统中模糊查询的实现

系统中需要频繁调用数据, 每一次查询都涉及到表中的大量数据, 而模糊查询结果要最大程度的接近用户需要的值, 这样才能方便用户选择. 在前面介绍的两种模糊查询方法中, 对于任何一种查询方法来说, 当数据库中有大量待查询数据时, 效率都比较低, 所以单独使用一种方法不能很好的解决查询精度和查询效率的问题, 系统中为了解决这个问题, 采用了两次模糊查询方法, 既当客户端将查询请求发送给应用服务器, 应用服务器接受请求后提交 SQL 查询给数据库服务器, 数据库服务器接受到请求后进行初次模糊查询, 然后将查询结果返回给应用服务器, 应用服务器端再用正则表达式对数据进行二次模糊查询, 这样得到的数据的匹配度就比较高了. 而且两次查询分别在数据库服务器端和应用服务器端进行, 分担了服务器的压力, 提高了查询速度, 具体实现过程如图 3 所示.

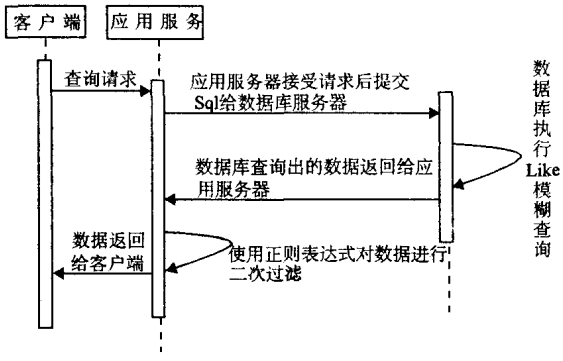

图 3　B/S 模式下的查询过程示意图

xm	xbdm	mzdm	zydm	syszddm	dwdm
张会松	1	01	11010400	610928	40102
孙天宝	1	01	11010400	620602	40102
张金泽	1	01	11010400	620600	10109
张建	1	01	11010400	652900	10109
马贤勇	1	01	11010400	371726	40105
何婵	2	01	11010400	620000	40105
曹宁	1	01	11010400	652701	40105
易文光	1	01	11010400	420222	10101
周连胜	1	01	11030200	650100	10101
张琼	2	01	11030200	654300	10101
李文强	1	01	62030300	652301	10101
田亚丽	2	01	62030300	652701	40105
何丽娜	2	01	62030300	330185	60110
朱邦柱	1	01	08060500	431201	70349
朱思华	1	01	08060500	653130	30103
刘坤	1	01	02000000	654002	30105
龙世喜	1	06	08190200	652301	30326
李龙飞	1	01	08190200	654000	70604
戴善亮	1	01	08190200	650100	30109
魏海波	1	01	08190200	652323	10102

图 4　验证系统中收集的代码信息

以查询包含字符串"xinjiang"为例，在此次查询中，数据库表中共有数据6937条，使用SQL模糊查询到的数据是 811 条，再用正则表达式进行匹配后查询到的匹配数据是 136 条，验证得到此查询方法的效率得到了大大提高. 通过这样的查询方式最后用户可以选择到最想要的结果，而此结果也能够与数据库表中的代码匹配，从而能够高效的完成信息采集. 图 4 为按照要求记录到数据库中的数据部分内容截图.

4 结语

重点描述了在使用标准化代码系统中数据精确录入的问题，为提高拼音检索查询速度和精度使用了双重模糊匹配方法，为验证该方法的可用性和数据的准确性，使用程序跟踪的方式验证了查询精度，并对该模块做了数据准确性的测试，实验结果表明，在信息采集页面中使用拼音检索的方式查询数据，方便了用户操作，提高了数据采集的准确性，双重模糊匹配方法也兼顾了查询效率与查询精度的问题，达到了系统预期的效果.

参考文献

1 孙威.汉字姓名模糊检索的实现.警察技术,2001,5:27–28.
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.南京师范大学学报.工程技术,2004,4(3):76–78.
3 刘松业.正则表达式的 Web 数据提取研究.电脑编程技巧与维护,2008,(16):89–91.
4 余石泉,周肆清.正则表达式在编程题自动阅卷中的应用.计算机技术与发展,2007,17(7):224–246.
5 杨成科.基于正则表达式的模糊查询和数据匹配验证.电脑知识与技术,2008,10:411–412.
6 陈天河.Java 数据库高级编程宝典.北京:电子工业出版社,2005.
7 思维科技,叶达峰.Eclipse 编程技术与实例.北京:人民邮电出版社,2006.
8 邓子云,张赐.JSP 网络编程从基础到实践.北京:电子工业出版社,2005.
9 刘金晓,马素霞,齐林海.Web 应用系统中权限控制的研究与实现.计算机工程与设计,2008,29(10):2550–2553.
10 陈继南,姜莹,孔祥荣.基于角色的 Web 信息系统权限管理方法.武汉理工大学学报:信息与管理工程版,2008,30(2):265.

(上接第 191 页)

要经过模数转换，得到具体的数值，最后再通过 Web 服务将最终的结果展示给终端用户.

4 结论

本文给出了基于 S3C2440 处理器和 Linux 操作系统，通过接入互联网，采用 Web Service 将视频与传感器信息集中展示供用户查看的智能监控终端设计方案. 该设计具有成本低廉，易于大规模应用的特点，同时具有功能强，系统实时性高等优点. ARM 架构芯片从成本控制设计上则更加适合多媒体便携式产品的大规模使用. 由此，可用较低的成本搭建具有较强功能的安全监控网络.

参考文献

1 王凯全,邵辉.事故理论与分析技术.北京:化学工业出版社.
2 蒋林涛.互联网与物联网.电信工程技术与标准化,2010,2:1–5.
3 沈苏彬.物联网技术架构.中兴通讯技术,2011(1).
4 沈苏彬,范曲立,宗平,毛燕琴,黄维.物联网的体系结构与相关技术研究.南京邮电大学学报(自然科学版),2009,29(6):1–11.
5 李红娟,吴雪莉.基于 ARM 和 RFID 技术的嵌入式系统研究.吉林化工学院学报,2008(2).
6 Samsung Electronics Co.Ltd.S3C2440 Datasheet.Korea:2.