引用格式：叶鹏,张雪英,杜咪.顾及字符特征的中文地名词典查询方法[J].地球信息科学学报,2018,20(7):880-886. [Ye P, Zhang X Y, Du M. Query method of Chinese gazetteer based on the character features[J]. Journal of Geo-information Science, 2018,20(7):880-886.] **DOI:** 10.12082/dqxxkx.2018.170530

顾及字符特征的中文地名词典查询方法

叶　鹏[1,2],张雪英[1,2*],杜　咪[1,2]

1. 南京师范大学 虚拟地理环境教育部重点实验室, 南京 210023; 2. 江苏省地理信息资源开发与利用协同创新中心,南京 210023

Query Method of Chinese Gazetteer Based on the Character Features

YE Peng[1,2], ZHANG Xueying[1,2*], DU Mi[1,2]

1. Key Laboratory of Virtual Geographic Environment, Nanjing Normal University, Ministry of Education, Nanjing 210023, China;
2. Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China

Abstract: With the rapid development of mobile Internet and the wide application of location-based service technology in various industries, the public's demand for the application of place information is growing rapidly. The gazetteer query, which can provide the support for place names knowledge, is an important basic link in the location information service. At present, because of the significant increase of the data volume of the place names, the query performance of gazetteers is facing a severe challenge. Most of the existing gazetteers directly use general retrieval methods, ignoring the characteristics of the characters and the description rules of the place names themselves. In order to solve these problems, a Chinese gazetteer query method (CGQM) is proposed based on the character features of place names. The CGQM uses the character features of the names with the same character characteristics, character's number and character's position, and query the gazetteer according to the main line of "candidate place name query, place name filtering, place name similarity ranking". Firstly, the single character index of the gazetteer is constructed, and based on this index, the place names containing the same characters in the gazetteer are queried to form a candidate dataset. Secondly, the place names are filtered from the candidate dataset, which has large differences in the number of characters with the search place names. The aim of this step is to enhance the accuracy of the candidate dataset and to ensure the efficiency of the later sorting process. Thirdly, the candidate place names are sorted based on the algorithm of character position similarity. Taking the national Chinese gazetteer as an example, an experiment was implemented with CGQM and a full text query method (Lucene) on 5 test datasets. The purpose of the experiment was to verify that the CGQM

收稿日期：2017-11-12；**修回日期**：2018-05-04.

基金项目：国家自然科学基金项目(41671393、41631177)；国家重点研发计划(2017YFB0503602)；江苏省高校自然资助项目(15KJA420002)；公安部科技强警基础工作专项项目(2016GABJC43、2017GABJC23)；警用地理信息技术公安部重点实验室开放课题(2016LPGIT01)。 [**Foundation items:** National Natural Science Foundation of China, No.41671393, 41631177; National Key Research and Development Program of China, No.2017YFB0503602; University Natural Funding Project of Jiangsu Province, No.15KJA420002; Special Project of the Basic Work of Science and Technology Police in the Ministry of Public Security, No.2016GABJC43, 2017GABJC23; Open Project of Key Laboratory of Police Department of Police, Geographic Information Technology, Ministry of Public Security, No.2016LPGIT01.]

作者简介：叶　鹏(1991-),男,博士生,主要从事时空大数据挖掘、遥感影像处理和地理信息系统研究。E-mail: yep730@163.com

***通讯作者**：张雪英(1970-),女,博士,教授,主要从事时空大数据挖掘、空间位置服务和地理信息系统等方面研究。E-mail: zhangsnowy@163.com

method could accurately and efficiently query the gazetteer. The experimental performance evaluation indexes include the operation efficiency, the precision rate, the recall rate and the F value. The results of experiment prove that CGQM can achieve much more better query performance than the Lucene based method. In the future research on gazetteer query, we will also consider many other factors, such as glyph, semantics, etc., and learn from the distributed and multithreading techniques in the retrieval system at the same time. These methods will promote the accuracy and efficiency of gazetteer query and expand the public service of place information.

Key words: Chinese place name; gazetteer query; Chinese gazetteer index for single Chinese characters; the similarity of place name; place name character features

***Corresponding author:** ZHANG Xueying, E-mail: zhangsnowy@163.com

摘要：地名词典查询是地名校正、地名匹配等地名服务应用的重要基础，但是地名数量的快速增长使得词典查询性能面临严峻挑战。针对大规模数据环境中传统词典查询方法准确率不高且效率较低等问题，提出了一种顾及字符特征的中文地名词典查询方法（CGQM）。首先，查询具有相同字符特征的地名形成候选地名集合，同时构建单字索引提升查询效率；其次，依据字符数量特征比较查询地名与候选地名的差异，进一步过滤候选地名集合；最后，基于字符位置特征优化查询结果排序策略，使得结果排序更为合理。实验以全国地名词典为例，构建5组测试集进行CGQM方法与Lucene检索方法的对比分析。研究结果表明，CGQM方法对于增强地名词典查询功能、提升查询效率具有实际意义。

关键词：中文地名；地名词典查询；地名词典单字索引；地名相似度；地名字符特征

1 引言

近年来，随着移动互联网的快速发展和基于位置的服务技术在各行业的广泛应用，社会公众对地名信息的应用需求日显增长[1-2]。在国家“十二五”、“十三五”和相关行业发展规划的推动下，我国地名信息化建设在数据规模和地名服务等方面均取得了长足发展[3]。地名词典查询是地名服务中的一个重要基础环节，为地名文本校正、地名语义消歧、位置信息匹配等提供地名词语知识的支持。由于地名数据积累规模的日益增大[4-6]，实现地名词典的快速、准确查询成为地名服务面临的重要技术挑战。

早期的词典查询主要是基于传统Hash方法进行[7]，将词典结构分为词典正文、词索引表、首字散列表等三级。通过首字散列表的Hash定位和词索引表获取查询词的位置范围，进而依据二分法在词典正文中定位。由于在查询过程中采用全词匹配，效率较为低下。基于Trie树的词典查询机制[8-9]由首字散列表和Trie索引树结点两部分组成，词典查询时按照树链顺序逐字匹配，减少无谓的字符串比较。但是，Trie树结构对于内存消耗巨大，同时索引构造与维护也较为复杂。双字Hash机制词典查询方法的提出，开始将字符特征融合在查询过程中[10-11]。对于2个字词以下的短词用Trie索引树机制实现，3字及以上的长词部分用线性表组织。能够避免部分的深度搜索，一定程度上提高了查询性能。由于词典文件可以作为一种全文数据，全文检索方法也被越来越多应用到词典查询中[12]。全文检索具有灵活高效的特点，但是基于关键字/词的检索机制可能会返回大量无关结果。事实上，中文地名具有字符长度较短、数据量巨大、描述形式多样等特点。现有的地名词典查询大多直接采用或借鉴通用检索方法，忽略了地名本身的字符特征和描述规律。因此，如何在中文地名词典查询中有效利用其字符结构特性，成为实质性提升查询性能的突破口。

2 基本思路

同一地点的地名表述存在多种方式[13]，而且查询地名输入错误的情况也比较普遍。因此，地名词典查询不仅要求对于输入的查询请求具有较好的容错性，而且能够高效返回完全准确或者最为接近的查询结果。本文利用地名中的相同字符、字符数量、字符位置等语言特征，按照“候选地名查询-字符数量过滤-相似程度排序”的技术路线（图1），设计一种高效的中文地名词典查询方法（简称CGQM）。首先，从地名词典中查询拥有相同汉字的地名形成候选集合，同时构建地名单字索引以提升查询效率；其次，将候选地名集合中与查询地名字符数量差异过大的地名进行筛除，加

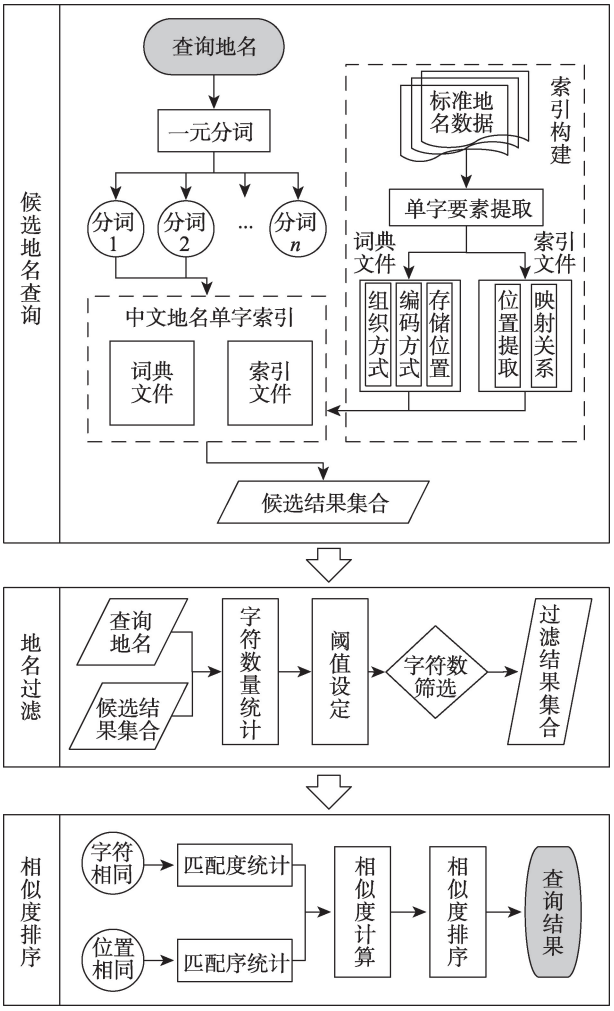

图1 中文地名词典查询的技术框架

Fig. 1 The technical framework of the Chinese gazetteer query method

强查询结果精确程度的同时保证后期排序过程效率;最后,对地名过滤结果依据字符位置相似度高低排序,将排序靠前的地名作为查询结果以进一步完善查询的准确性。

3 基于相同字符的候选地名查询

3.1 单字索引构建

传统的词典索引结构多数以词组为对象提取索引词,由于受到词组描述粒度及数量的影响,对于查询条件的容错性存在一定局限。汉字作为汉语构成的最小单元,其对于词典中词项的关联性较词组更为丰富[14]。当查询地名中存在部分信息失真时,基于单个汉字的索引形式能够依据剩余的部分准确字符与目标地名知识建立关联关系。在大规模数据环境下,索引词的增多极易产生数据冗余,也容易导致检索效率下降。中文地名中常用汉字规模较为固定,词典数据规模的扩大不会导致索引项的线性增长。在索引查找过程中,能够有效避免无关词项的深度搜索。因此,以地名中包含的单字作为索引词建立中文地名索引,对于地名查询具有更强的适应性。

中文地名单字索引由词典文件和索引文件两部分组成。词典文件用于存储地名词典中全部的地名数据,按照无换行无间隔的方式依次排列,形成一条连续的字符串;索引文件是存储索引记录的物理文件,用于存储索引记录和词典文件中地名词项之间的对应关系。一条索引记录中包含3部分信息:地名个数,字符编码以及词典位置。假设词典文件中共有n个不重复汉字W_i,$i\in[1,n]$,C_i表示汉字W_i的UTF-8编码,N_i为词典文件中包含汉字W_i的地名个数,每个地名的起始位置与结束位置分别表示为S_{nm}、E_{nm},那么地名在词典文件中的存储位置序列表示为$<S_{n1},E_{n1},S_{n2},E_{n2},\cdots,S_{nm},E_{nm}>$。以地名“中岗子”为例,将“中岗子”存储到词典文件中,记录下S_{nm}(“中”在字符串中位置1001)与E_{nm}(“子”在字符串中位置1003)。之后在索引文件中生成“中”、“岗”、“子”3条索引记录,其中“中”字索引为[11079][0xE4B8AD][1001,1003,1015,1017,..,83475,83478],记录字符编码(0xE4B8AD)、词典文件中所有包含“中”字地名的个数(11079)及其存储位置,既有“中岗子”所在位置(1001,1003),还有“中夹滩”、“姜尾林中”等其它含“中”地名所在位置,如(1015,1017)(83475,83478)等(图2)。

对地名单字索引解析算法的时间复杂度进行分析,设索引文件中共有n个索引项,查询地名中共有m个不同的单字,则与之相关的m个索引项中共有r个位置映射记录。在依据索引进行地名查询时,对以上n个索引项进行一次扫描即可获得全部的查询结果。其时间复杂度$T=T_1(n)+T_2(r)$,即$T=O(\max(f(n),g(r)))$。$f(n)$与$g(r)$都为单循环遍历查找,时间复杂度都为$O(N)$。因此单字索引解析计算的时间复杂度为$O(N)$。

3.2 候选地名查询

候选地名查询的目的是从地名词典中查询到与查询地名包含相同汉字的地名。首先对输入查询地名进行中文分词,采用一元分词形式将所有中文字符按照单字形式输出,如查询“中岗子”拆分为

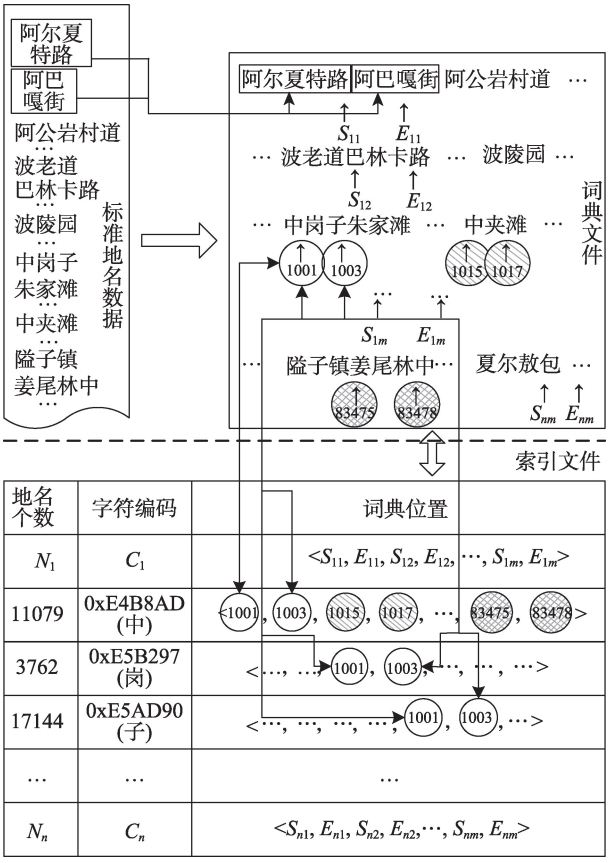

图2　中文地名单字索引组织方式

Fig. 2　Chinese gazetteer index based on single characters

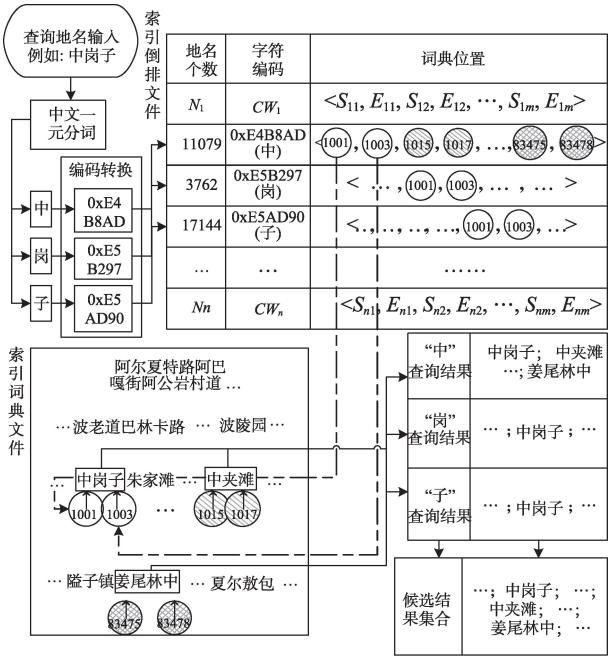

图3　中文地名单字索引查询方式

Fig. 3　The query mode of Chinese gazetteer index based on single characters

“中/岗/子”。其次，以分词结果分别作为查询关键字，在索引文件中查询其对应的索引记录。之后对索引记录中的信息进行逆向解析，根据索引中位置信息查询词典文件中对应的地名数据，并将全部查询结果返回形成候选结果集合（图3）。

4　基于字符数量的地名过滤

候选结果集合规模的降低有利于减小后期相似度计算的运行时间，进而提升查询方法的整体效率。需要选取特征对候选结果进行筛选，尽可能多的过滤掉与查询目标不相符的候选项。常见的地名不准确描述方式包括替换字符、缺失字符、增加字符及交换字符等形式（表1）。虽然错误方式形式各异，但是查询地名与目标地名在字符长度上具有较高相似性。因此，可以根据查询地名的字符数量，对候选结果集合中的地名项进行过滤。首先，记录查询地名 P 的字符数量为 n，候选结果集合 W 中地名 W_x 的字符数量为 m，设定阈值范围为 k，则当满足 $abs(m-n)\leqslant k$ 时，将 W_x 保存到过滤结果集合 C 中。

表1　查询地名中常见的不准确描述方式

Tab. 1　Common inaccurate description form in the place name search

类型		查询地名	目标地名
替换字符		候家宅子村	侯家宅子村
		晌滩村	响滩村
缺失字符		采石南路南	采石南路南口
		合肥南	合肥南站
增加字符	多余空格	南京 市	南京市
	特殊符号	凉水–井湾	凉水井湾
	偏旁分离	夕卜坡	外坡
交换字符		北新桥路口南	北新桥南路口
		塔什库尔干塔吉克县	塔库什尔干塔吉克县

5　基于字符位置的地名相似度排序

字符匹配法是较为典型的中文词汇（或字符串）相似度判别方法之一[15]。假设有 A 和 B 两个字符串，N 表示 A 与 B 之间的相同字符数（匹配度），C_1 表示 N 与 A 的总字符数之比，C_2 表示 N 与 B 的总字符数之比。N、C_1、C_2 共同构成 A 与 B 的匹配度，以此判断 A 和 B 之间的文本相似度。字符匹配法只考虑词汇之间的字符相同程度，却忽视了匹配字符位于

字符串中的位置信息[16]。中文里绝大多数汉字都是表意单元,词语搭配比较灵活多样。然而,地名作为一种专有名词,其中各字符间顺序通常不可调换。因此,对于中文地名间相似度的判定,需要在相似度评价时增加对字符串间词序位置关系(匹配序)的计算。基于此,本文提出一种基于字符位置的地名相似度计算方法(式(1))。

$$sim(P,\ W)=\alpha\times\frac{1}{2}\left(\frac{c}{m}+\frac{c}{n}\right)+\beta\times\min\left(\frac{m}{n},\frac{n}{m}\right)\times\frac{1}{2}\left(\frac{\sum_{i=1}^{c}L_1(i)}{\sum_{t=1}^{m}t}+\frac{\sum_{i=1}^{c}L_2(i)}{\sum_{k=1}^{n}k}\right)\quad(1)$$

式中:P与W分别表示2个地名字符串;m与n分别表示P与W的字符总数;c表示P与W的字符匹配度;$L_1(i)$与$L_2(i)$分别表示匹配字符i在P与W中的匹配序;α与β分别表示匹配度与匹配序评价结果的权重,并且α与β的和为1。通常情况下α与β的取值依据黄金分割定律,分别取0.6与0.4。匹配序按照从左到右的顺序,从起始位置1开始以递增的方式计算。以P="师范大学",W="南京师范大学"为例,P与W的匹配字符为"师"、"范"、"大"、"学"。其在P中的匹配序为1(师)、2(范)、3(大)、4(学),在W中的匹配序位3(师)、4(范)、5(大)、6(学)。按照本文的相似度计算方法,P与W的相似度定义为:

$$sim(P,\ W)=0.6\times\left(\frac{4}{4}+\frac{4}{6}\right)\times\frac{1}{2}+0.4\times\min\left(\frac{4}{6},\frac{6}{4}\right)\times\frac{1}{2}\times\left(\frac{1+2+3+4}{1+2+3+4}+\frac{3+4+5+6}{1+2+3+4+5+6}\right)\approx0.75\quad(2)$$

6 实验评估分析

以480万条全国地名数据构建实验词典,从中抽取1700条地名作为查询地名。为模拟实际查询情况,对查询地名人为增加错误。错误类型涵盖了表1归纳的各类描述方式,并依据与原有标准地名对比的准确度将其划分为5个等级(表2)。准确度定义如式(3)所示。

$$A(p,w)=\frac{c}{n}\quad(3)$$

式中:c表示查询地名p中与目标地名w相比准确的字符数量;n表示查询地名p字符数量。开源全文搜索引擎Lucene在文本分类、信息检索等方面有大量研究与应用[17]。词典作为非结构化文本文件,能够应用Lucene索引机制。因此,本文选取Lucene检索方法与CGQM进行对比实验。查询性能评价指标包括运行效率、准确率(P)、召回率(R)、F值。其中,运行效率是指单条地名查询所耗费的时间。P、R与F度量值的具体计算公式如式(4)-(6)所示。式中,n_{ij}是指目标地名i和查询结果j之间相同的数量,n_i是指目标地名i的数量,n_j是指模型查询结果j的数量,$F(i,j)$是指i和j之间的F度量值。本次实验中设置的地名过滤阈值k,为查询地名与目标地名中较长地名字符数量的30%。同时以相似度数值大于60%的候选地名作为查询结果,结果依据相似度数值大小进行排序。实验测试机器配置为Intel Core i7-7700HQ 主频2.8 GHz处理器,内存16 GB,Windows 10操作系统,开发语言为Java。

$$P(i,j)=\frac{n_{ij}}{n_j}\quad(4)$$

$$R(i,j)=\frac{n_{ij}}{n_i}\quad(5)$$

$$F(i,j)=\frac{2\times P(i,j)\times R(i,j)}{P(i,j)+R(i,j)}\quad(6)$$

实验结果表明,CGQM的性能明显优于Lucene方法。CGQM在大规模数据环境下可以保持较高的运行效率,同时能够在查询地名不准确的情况下较为准确的查询到目标地名(表3)。Lucene以词元为单位,通过对查询地名分词再与索引进行精确匹

表2 实验测试集划分明细及示例

Tab. 2 Samples of test datasets

等级	测试集单条地名准确度	测试集地名数量	测试集示例	对应目标地名
测试集1	[90%, 100%]	133	南京明文化村阳山碑村	南京明文化村 阳山碑村
测试集2	[80%, 90%)	377	候家石良村,勒图音敖包	侯家石良村,勒图音敖包
测试集3	[70%, 80%)	389	大新冊村,豆家吕村	大新册村,豆家营村
测试集4	[60%, 70%)	665	椿木槽,达强	椿木槽,达强弄
测试集5	[50%, 60%)	136	横山,痳冲	横山,麻冲

表3　实验结果评价指标统计

Tab. 3　Statistics of experimental results

测试集	地名数量/个	CGQM方法				Lucene方法			
		P/%	R/%	F	平均效率/ms	P/%	R/%	F	平均效率/ms
1	133	95.49	100.00	98.08	409	93.23	98.50	95.79	576
2	377	91.78	94.43	93.09	335	90.45	91.51	90.98	537
3	389	82.26	88.95	85.47	437	79.18	84.83	81.91	548
4	665	72.03	80.00	75.81	388	69.02	76.09	72.38	513
5	136	53.97	73.53	62.25	186	50.74	68.38	58.25	562

配。①Lucene在进行查询时，借助分词器与分词词典对查询地名进行拆分更为复杂；②也会因分词词典中缺少必要词项，使用分词时遇到歧义而产生了错误切分影响查询准确性。CGQM对查询地名分割不依赖任何语义知识，同时操作简单。特别是，当查询地名准确度在80%以上时，CGQM能够较为准确查询到目标地名。受到测试集准确度的影响，各测试集间准确率存在差异。随着查询地名准确度的降低查询准确率也不断降低，但更多是由于查询条件中准确信息缺失导致的语义改变。

对CGQM方法的各个中间环节进行具体分析，部分实验数据查询过程如表4所示。查询过程体现出，CGQM方法利用的地名中相同字符、字符数量、字符位置等语言特征，能够较为有效的逐级排除无

表4　部分实验数据查询过程明细

Tab. 4　Details of the query process of the part of experimental data

查询地名	所属测试集	初步结果集合（部分示例）	过滤结果集合（部分示例）	查询结果排序（部分示例）	目标地名
努木其音乌	测试集2	力努；努松；桥努；…；株木塘；木底塘；木山冲；…；佳木斯我的家生态健康社区；米欠扎木阿吉坎儿孜买里斯；树木岭民营工业园基地三门；…（共50 430个）	哈达音努如；努木乃淖日；努和廷沙图；…；额尔格勒音努如；沙巴日努很超浩；居努斯阔克铁木；…（共22 101个）	努木其音乌兰	努木其音乌兰
兩山村	测试集4	雨道；雨潭；山岗；…；社山后；开化山；山马岭；…；石家庄华南新村；浚县王升屯新村；平成日式度假村；…（共313 867个）	雨花冲；梧桐雨；雨冲子；雨水冲；…；青山程家；落雁山村；东畈横山；…（共265 970个）	村山村；山村；陈山村；东山村；三山村；阳山村；檀山村；嶂山村；横山村；兴山村	雨山村

关候选项。然而实验中仍有部分地名查询不准确，分析其原因主要在于汉语中具有相同词形结构的地名数量众多。以查询地名“兩山村”为例，查询结果排序前十位的结果都不是目标地名“雨山村”，但是都为统一的“X山村”词形结构。这不利于相似度评价结果的区分，进而影响了最终查询结果排序的准确性。

7　结论

本文以挖掘中文地名的字符特征为突破口，提出了一种较为有效的中文地名词典查询方法。该方法基于相同字符特征查找候选地名，对查询地名具有良好的容错性，并提出地名词典单字索引结构提升了查询效率。利用字符数量进行候选地名过滤，结合字符位置特征进行相似度排序，使得查询结果更加准确与人性化。在今后研究中地名查询还应综合考虑字形、语义等其他多种因素，同时借鉴检索系统中分布式、多线程等技术手段。以此促进地名查询准确率与效率的进一步提升，推动地名信息公共服务的拓展。

参考文献(References)：

[1] 张文元,周世宇,谈国新.基于Lucene的地名数据库快速检索系统[J].计算机应用研究,2017,34(6):1756-1761. [Zhang W Y, Zhou S Y, Tan G X. Place name database quick searching system based on Lucene[J]. Application Research of Computers, 2017,34(6):1756-1761.]

[2] Delmastro F, Arnaboldi V, Conti M. People-centric com-

puting and communications in smart cities[J]. IEEE Communications Magazine, 2016,54(7):122-128.

[3] 李东阳,方俊杰,许大璐.GIS技术支持下的多部门地名地址业务协同研究与实现[J].测绘通报,2016,62(10):121-124. [Li D Y, Fang J J, Xu D L. Collaborative research and implementation of multi- sector address business based on GIS technology[J]. Bulletin of Surveying and Mapping, 2016,62(10):121-124.]

[4] 张雪英,闾国年,杜咪,等.大数据驱动的地名信息获取与应用[J].现代测绘,2017,40(2):1-5. [Zhang X Y, Lv G N, Du M, et al. Acquisition and application on geographical names information based on large data driving[J]. Modern Surveying and Mapping, 2017,40(2):1-5.]

[5] 许普乐,王杨,黄亚坤,等.大数据环境下基于贝叶斯推理的中文地名地址匹配方法[J].计算机科学,2017,44(9):266-271. [Xu P L, Wang Y, Huang Y K, et al. Chinese place-name address matching method based on large data analysis and bayesian decision[J]. Computer Science, 2017,44(9):266-271.]

[6] 董洁钰,马梦宇,陈荦,等.基于对象关系型数据库的多级地名地址服务研究[J].地理信息世界,2017,24(4):92-95,100. [Dong J Y, Ma M Y, Chen Y, et al. Research on multi- level geographical names and addresses service based on object relational database[J]. Geomatics World, 2017,24(4):92-95,100.]

[7] 王秀坤,李政,简幼良,等.基于Hash方法的机器翻译词典的组织与构造[J].大连理工大学学报,1996(3):108-111. [Wang X K, Li Z, Jian Y L, et al. Machine translation dictionary based on Hash method[J]. Journal of Dalian University of Technology, 1996(3):108-111.]

[8] 孙茂松,左正平,黄昌宁.汉语自动分词词典机制的实验研究[J].中文信息学报,2000(1):1-6. [Sun M S, Zuo Z P, Huang C N. An experimental study on dictionary mechanism for Chinese word segmentation[J]. Journal of Chinese Information Processing, 2000(1):1-6.]

[9] 梁南元.书面汉语自动分词系统—CDWS[J].中文信息学报,1987(2):44-52. [Liang N Y. The mordern printed Chinese distinguishing word system[J]. Journal of Chinese Information Processing, 1987(2):44-52.]

[10] 李庆虎,陈玉健,孙家广.一种中文分词词典新机制——双字哈希机制[J].中文信息学报,2003(4):13-18. [Li Q H, Chen Y J, Sun J G. A new dictionary mechanism for Chinese word segmentation[J]. Journal of Chinese Information Processing, 2003(4):13-18.]

[11] 李江波,周强,陈祖舜.汉语词典的快速查询算法研究[J].中文信息学报,2006(5):31-39. [Li J B, Zhou Q, Chen Z S. A study on fast algorithm for Chinese dictionary lookup [J]. Journal of Chinese Information Processing, 2006(5):31-39.]

[12] 吴鹏飞,马凤娟,李文革,等.开源全文检索引擎Lucene本地化实践研究[J].现代图书情报技术,2009(4):19-22. [Wu P F, Ma F J, Li W G, et al. Localization of the open source full-text retrival engine based on Lucene[J]. Data Analysis and Knowledge Discovery, 2009(4):19-22.]

[13] 李淑霞.地名本体及其在地理空间数据组织中的应用研究[D].郑州:解放军信息工程大学,2009. [Li S X. Research on ontology of place and its applications in geospatial data organization[D]. Zhengzhou: PLA Information Engineering University, 2009.]

[14] 胡盈盈.单汉字标引与检索技术综析[J].情报理论与实践,1999,36(2):74-77. [Hu Y Y. Analysis of indexing and retrieval techniques for single Chinese characters[J]. Information Studies: Theory & Application, 1999,36(2):74-77.]

[15] 宋明亮.汉语词汇字面相似性原理与后控制词表动态维护研究[J].情报学报,1996,15(4):22-32. [Song M L. The principle of literal similarity of Chinese words and the dynamic maintenance of post controlled vocabulary[J]. Journal of the China Society for Scientific and Technical Information, 1996,15(4):22-32.]

[16] 张雪英,闾国年.基于字面相似度的地理信息分类体系自动转换方法[J].遥感学报,2008,23(3):433-441. [Zhang X Y, Lv G N. Approach to Automatic Conversion of Geographic Information Classification Schemes[J]. Journal of Remote Sensing, 2008,23(3):433-441.]

[17] Hirsch L, Hirsch R, Saeedi M. Evolving Lucene search queries for text classification[C]. Proceeding of the 9th Annual Conference on Genetic and Evolutionary Computation. New York: ACM Press, 2007:1604-1611.

[18] Milosavljevic B, Boberic D, Surla D. Retrieval of bibliographic records using Apache Lucene[J]. Electronic Library, 2010,28(4):525-539.