引文格式: 陈利燕,林鸿,张新长.一种改进的 Lucene 算法及在空间数据融合中的应用[J].测绘通报,2016(10):106-109.DOI:10.13474/j.cnki.11-2246.2016.0341.

一种改进的 Lucene 算法及在空间数据融合中的应用

陈利燕[1,2],林 鸿[2],张新长[1]

(1. 中山大学地理科学与规划学院,广东 广州 510275; 2. 广州市城市规划勘测设计研究院,广东 广州 510060)

An Improved Lucene Algorithm and Its Application to Spatial Data Fusion

CHEN Liyan,LIN Hong,ZHANG Xinchang

摘要: 在"互联网+"时代,众源地理空间数据已成为重要的数据来源,但由于众源数据存在冗余和精度不高等问题,如何有效利用众源数据已成为地理信息技术研究的热点。众源 POI 数据一般通过与标准数据进行同名点匹配解决上述等问题。而同名点匹配常用的方法有编辑距离算法、最长公共子串算法、贪心字符串匹配算法,这些方法存在匹配效率低、缺少语义判断等问题。为此本文基于 Lucene 提出了一种基于语义相似度的快速匹配算法,试验结果表明,与传统的字符匹配方法相比,本文提出的方法在运算效率上有显著的提升,同时还能通过判断语义相似度提高匹配率。

关键词: 同名点匹配; 字符串匹配; Lucene 索引; 语义相似度

中图分类号: P208 **文献标识码:** B **文章编号:** 0494-0911(2016)10-0106-04

随着"互联网+"时代的来临,众源地理空间数据已成为当前空间信息应用的重要数据来源。与传统地理信息采集和更新方式相比,来自非专业大众的众源空间数据具有数据量大、现势性好、信息丰富、成本低等特点和优势[1],成为近年来国际地理信息科学领域的研究热点。在移动及 Web 环境下,众源 POI 数据与地理框架背景数据的混搭式地图应用,越来越多地出现在主流地理信息平台及 LBS 服务中。但由于众源 POI 数据存在信息冗余、缺乏质量信息或质量信息不精确等问题[1],在应用前必须利用匹配技术进行信息空间位置纠正和筛选去重操作。POI 点匹配原理主要是利用对象之间的名称或位置描述字符串相似度来判断是否为同一对象。常用字符串相似度计算方法有编辑距离算法、最长公共子串算法、贪心字符串匹配算法等,但这些匹配算法对于众源 POI 海量数据而言,逐条循环匹配效率低,且缺少语义相似度的判断。Lucene 算法作为当前流行的信息检索技术,虽然也被广泛应用于地理信息应用领域,但都是基于传统词频分析技术,对其存在的检索精确度和召回率存在的不足很少有人讨论,同时也忽视了语义的判断。为此本文提出一种改进的 Lucene 算法,以改进传统基于词频的方法对语义忽视所造成的检索不够精确的问题,同时给出一个初步判定语义相似性的算法。试验结果表明,通过这些改进,与传统的字符串匹配,本文提出的方法能较好地提高运行效率和查询准确率。

一、相关研究

文献[2]中对编辑距离(LD)[3]、最长公共子串(LCS)[4]、贪心字符串匹配(GST)[5]和改进的贪心字符串匹配[6-7]等算法原理进行了详细描述。这些算法虽然能较好地反映字符串之间的相似程度,但存在检索效率低、忽略对象之间语义关系等问题。Lucene 作为一个通用的搜索引擎开发工具包被广泛地应用于检索领域中[8-9],通过索引技术提高了检索的效率。也有学者将其应用于地理信息系统应用中,如文献[10]将其应用于地图信息搜索,文献[11—12]将其应用于地址匹配等。尽管 Lucene 已在地理信息领域被广泛应用,相关的研究也层出不穷,然而大多数研究都是基于 Lucene 内部默认实现的词频分析检索函数来考察对象之间的相似性来进行检索,很少有考虑对词项语义的 Lucene 检索研究。虽然文献[13]提出基于信息理论的词项语义相似度量方法计算词项之间的语义相似性,但该算法对于 POI 同名点在名称或空间位置字符串信息匹配上也同样存在精度不高的问题。

二、改进后 Lucene 基本工作流程

本算法在传统的 Lucene 工作流程上进行了改进(如图 1 所示),基本工作流程主要分 3 个阶段。

收稿日期: 2015-12-18
基金项目: 国家自然科学基金重点项目(41431178)
作者简介: 陈利燕(1981—),女,博士,高级工程师,研究方向为空间数据更新与融合。E-mail:jimigao@163.com

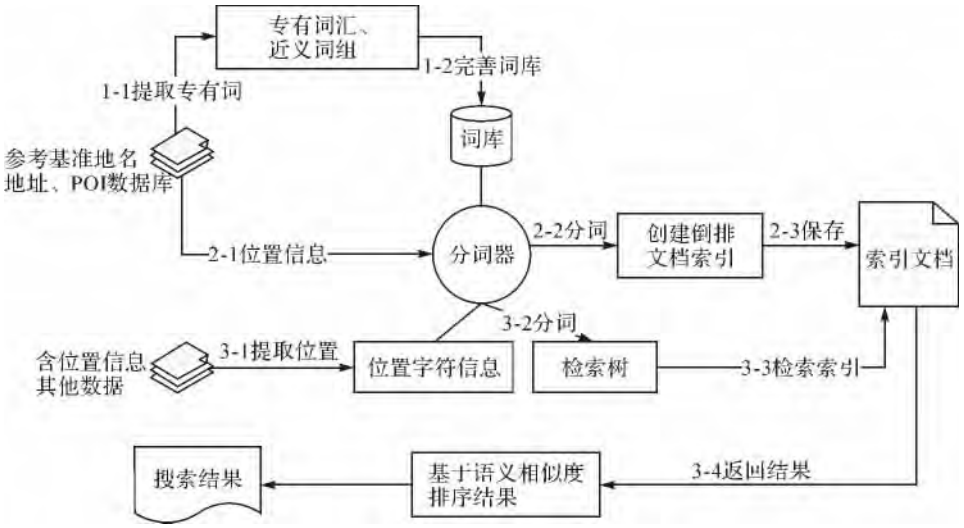

图 1　算法基本工作流程

第 1 阶段，根据区域内地名特征，对词库进行完善，包括: 1-1 对研究区范围的专有地名、街路巷等词汇进行提取; 1-2 将专业词汇及相关的近义词添加至词库。

第 2 阶段，将参考基准数据库建立索引文件，包括: 2-1 将基准数据库的名称和位置信息提取; 2-2 利用词库通过分词器将提取的字符串信息进行分割; 2-3 将分词后的信息按照规则保存至索引文件。

第 3 阶段，众源数据与参考基准数据的匹配，包括: 3-1 从众源数据中提取出包含名称和位置等可用于匹配的信息; 3-2 对可用于匹配的信息利用分词器进行词法分析和语义处理得到系列检索词树; 3-3 依据搜索词树通过读取索引搜索结果集; 3-4 对搜索结果集采用基于语义相似度评分机制进行排序，并返回最终结果。

算法的改进主要体现在图 1 所示的“词库”的完善和“基于的语义相似度评分机制”。“词库”主要包括专有名词和同义词(近义词) 库的补充和完善。词库完成后利用分词器可实现字符串按语义进行分词，如“中山大学”将被分割成“中山”“大学”和“中大”。“基于语义相似度评分机制”主要是利用分词后的词项之间的语义相似度评分来判断是否为同名点对象。

三、基于语义的词库的完善

“词”是字符串中最小的可以独立运用的单位，如“广州大学”可以分词为“广州”和“大学”。由于中文本身的复杂性及地址信息描述规则的不确定性，使中文分词成为分词技术中的难点。如“广州大学”和“广大”都表示的是同一个对象，如何让计算机能够基于语义判断地理对象之间的相似度是提高匹配准确率的关键技术。为此本文借助“充分大的”盘古词库，按照一定的策略将待分析的位置信息与词库的词条进行匹配后切分成一系列有意义的词。本文主要有两个方面的改进: ①根据试验区内的专有地名、街路巷等信息完善扩充和完善“词库”，位置信息中含有许多专有词汇，因此“词库”不完整会导致分词不准确而降低检索的查准率。如“广州市越秀区白米巷”因词库中缺少“越秀区”和“白米巷”会被分割成“广州/广州市/越秀/区/白米/巷/”; ②组建了地名、地址和 POI 点的近义词库。利用文献[14]的规则，将试验区内的位置、地名和机构等简称进行特征分析，利用人工交互的方法，建立近义词库，从而使得语义相同或相近的“词”可以进行有效匹配。地理位置信息的同名点判断通过“词”与“词”之间的语义相似度的相关函数确定。

四、Lucene 相似度函数的改进

设参考基准空间数据库为 D，对任意参考基准记录 d_k 位置信息文档，经语义分词后将每个词表示为如下 m 维向量形式

$$V(d_k) = [w_{k,1}\ \ w_{k,2}\ \ \cdots\ \ w_{k,m}]$$

对于待匹配的任一记录空间位置信息 q 个经分词后形成由搜索词项 t_m 的权重组成查询向量

$$V(q) = [s_1\ \ s_2\ \ \cdots\ \ s_n]$$

传统 Lucene 搜索模型是基于 VSM，即向量空间模型[18]，将参考基准向量及查询向量放到一个 N 维空间中，每个词 t(term) 是一维。两个向量之间的夹角越小，相关性越大。计算夹角的余弦值作为相关

性的度量，夹角越小，余弦值越大，分值越高，相关性越大（如图 2 所示）。相似度最终可表示为[15]

$$\mathrm{sim}(q,d)=\frac{v(q)\cdot v(d)}{|v(q)||v(d)|} \quad (1)$$

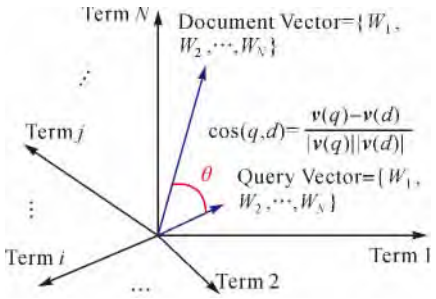

图 2 字符相似度空间多维向量

在传统 Lucene 搜索模型中，权重 w 由 TF-IDF 即“词频-逆向文档频率”计算公式来求得，即 $w=\mathrm{tf}(t)\cdot\mathrm{idf}(t)$，其中 tf（term frequency）表示一个词项 t 在一个文档 d 中出现的频率，idf（invert document frequency）表示词项 t 的逆文档频率，即文档集合中包含词项 t 的文档数目 $\mathrm{df}(t)$ 的倒数。将 TF-IDF 代入式（1）

$$\mathrm{sim}(q,d)=\left(\frac{\sum_{k=0}^{m}\max\{\mathrm{sim}(w_k,S_i)\mid i\in(1,2,\cdots,n)\}}{m}+\frac{\sum_{k=0}^{n}\max\{\mathrm{sim}(w_i,S_k)\mid i\in(1,2,\cdots,m)\}}{n}\times\frac{1}{2}\right) \quad (3)$$

式中，$\mathrm{sim}(w_k,s_i)$ 表示查询项搜索词 S_i 与参考基准文档 d 的 w_k 的语义相似度；$\max\{\mathrm{sim}(w_k,s_i)\}$ 表示与 w_k 语义相似度的最大值。

与式（2）相比，本文简化了相似度计算的公式，使得相似度计算值更容易理解并比较，同时检索结果不仅包含了与被检索词项相同的文档，而且还包含了与被检索词项相似的文档，从而更为准确地体现了检索的含义。

五、试验与分析

1）试验数据：广州电信公司提供的越秀区的网点数据共 1509 条（XLS 格式），从 2014 年广州市基础测绘成果中提取越秀区门牌数据共 40 490 条（shape file 格式）。

2）试验环境与试验平台：试验在 Window7 操作环境下进行，本文采用 Visual Studio 2010 结合 ArcGIS Engine10 开发了基于字符串相似度匹配算法的点要素空间数据融合原型系统。

3）试验结果评估：在众源数据融合使用中，主要是根据匹配融合的效率、匹配的精度（匹配总量和误匹配量）来评估算法的优劣性。

后经过归一化处理得到字符串相似度的计算公式为

$$\mathrm{sim}(q,d)=\frac{|q\cap d|}{|q|}\cdot\frac{1}{\sqrt{w(q)\times\sum_{t\in q}\left(1+\log\frac{N}{\mathrm{df}(t)+1}\right)^2}}\cdot\sum_{t\in q}\left(\mathrm{tf}(t)\cdot w(d)\cdot w(t)\cdot w(f)\left(1+\log\frac{N}{\mathrm{df}(t)+1}\right)^2\right)\cdot\frac{1}{\sqrt{N}} \quad (2)$$

式中，$|q|$ 为查询 q 的长度；$|q\cap d|$ 为同时出现在参考基准空间数据库 D 和待匹配检索 q 中的词项的数目；$\mathrm{tf}(t)$ 表示搜索词项 t 在参考基准文档 d 中出现的频率；$\mathrm{df}(t)$ 表示 t 在参考基准空间数据库 D 中出现的频率；$w(d)$、$w(t)$、$w(f)$ 分别表示参考基准文档、搜索词项和查询字段的权重；N 表示 d 中词项数。

逆文档频率权重的加入，使得完全相同的两个位置信息描述字符串比较最终的相似度也不是 100%，因此为了便于相似程度的比较，将相似度计算函数作如下的简化和改进

1. 试验 1

为了比较不同数据量匹配融合的时间消耗，从数据源中提取部分街路巷的数据，匹配对比分为 3 组，各组的数据量（电信网点数据×基础地理门牌数据）分别为 500×5092 条（组 1）；1000×9881 条（组 2）、1509×40 490 条（组 3），各种算法在不同匹配数据量上的时间消耗见表 1，时间增长率如图 3 所示。

表 1 各组匹配数据时间消耗 s

方法	组 1	组 2	组 3
编辑距离	113 340	252 693	1 562 875
最长公共子串	80 823	254 971	1 582 463
贪心字符串匹配	76 501	245 720	1 540 836
改进贪心字符串匹配	76 501	245 720	1 540 836
传统 lucene 算法	8930	8996	39 359
本文算法	8960	9020	39 469

从图 3 可以很明显地看出，在数量量不大的情况下，本文所提算法的优势并不明显；但随着数据量的增大，传统字符匹配算法时间消耗增长率呈线性增长，而本文提出的算法继承了传统 Lucene 算法在检索时间上的优势，在时间消耗上增长并不明显，由

此也可以看出本算法在大数据融合使用中具有更好的应用前景。

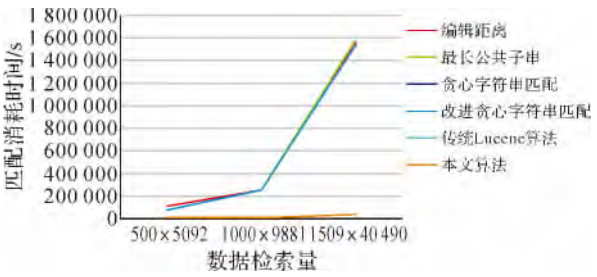

图 3　算法时间消耗对比

2. 试验 2

为了进一步比较各类算法的优劣性，对组 1（500×5 092 条）匹配结果进行了详细分析，结果见表 2。传统的 Lucene 算法和编辑距离算法在误匹配量上最大，本文算法的误匹配数量最少，查准率最高。通过对误匹配的数据记录进行分析，发现造成误匹配的原因主要有两个方面: 一是基准数据不完整，由于各类算法的匹配策略是将相似度判断值最大的对象作为结果进行匹配，且基础测绘成果中的门牌数据不完善，确有部分电信网点的地址信息不被包含，由此而造成了误匹配; 二是算法的不完善，因为各算法相似度计算方法不完善造成的错误匹配。

由于传统的 Lucene 算法在相似度函数中加入了逆文档频率权重的判断，使得在匹配过程中字符信息完全相同的对象之间的得分不一定是最高，而造成了误匹配。针对上述问题，本文对相似度函数进行了改进，同时加入了语义相似度的判断，减少了因算法不完善而引起的误匹配量，提高了匹配的准确率。

表 2　各类算法匹配结果

算法	匹配量	误匹配数量	漏匹配	查全率/(%)	查准率/(%)
编辑距离	500	59	0	100	88
最长公共子串	500	51	0	100	90
贪心字符串匹配	500	55	0	100	89
改进贪心字符串匹配	500	55	0	100	89
传统 Lucene 算法	500	59	0	100	88
本文算法	500	43	0	100	91

六、结　论

同名对象匹配作为矢量空间数据融合的重要过程，其匹配的效率和准确率决定了数据融合使用的质量。基于传统的字符串相似度匹配技术存在着效率低和缺少语义判断等问题，在匹配效果上不尽如人意。本文基于语义相似度判断对 Lucene 算法进行改进，有效提高了匹配的效果和精度。通过试验，得出的结论如下:

1）随着同名匹配数据量的增加，本文的方法在效率上优势明显，且保持了较高准确率，在众源海量空间数据融合上具有良好应用前景。

2）本文通过对近义词（同义词）库的补充和完善，实现了基于语义相似度的比较，有效解决了因数据来源不一致导致的同点不同名的现象，极大地提高了匹配的准确度。

本文的不足之处在于同名字符信息的语义分词依赖词库的完整性，由于词库的局限性，对特有名词或新词的匹配上有些不尽如人意的地方，需要对词库进行不断的更新和完善。

参考文献:

[1] 王明，李清泉，胡庆武，等. 面向众源开放街道地图空间数据的质量评价方法 [J]. 武汉大学学报（信息科学版），2013，38(12)：1490-1494.

[2] 牛永洁，张成. 多种字符串相似度算法的比较研究 [J]. 计算机与数字工程，2012，40(3)：14-17.

[3] 刁兴春，谭明超，曹建军. 一种融合多种编辑距离的字符串相似度计算方法 [J]. 计算机应用研究，2010，27(12)：4523-4525.

[4] 张毅超，车攻，马骏. 求最长公共子串问题的算法分析 [J]. 计算机仿真，2007(12)：97-100，116.

[5] 于海英. 字符串相似度度量中 LCS 和 GST 算法比较 [J]. 电子科技，2011，24(3)：101-103，124.

[6] WISE M J. Running Karp-Rabin Matching and Greedy String Tiling [C] // The Third International Conference on Intelligent Systems for Molecular Biology. Cambridge，England: [s.n.]，1993：393-401.

[7] 牛永洁. RKR-GST 算法在.NET 中的分析与实现 [J]. 信息技术，2012(3)：171-174.

[8] 张校乾，金玉玲，侯丽波. 一种基于 Lucene 检索引擎的全文数据库的研究与实现 [J]. 现代图书情报技术，2005 (2)：40-43，48.

[9] 张俊，李鲁群，周熔. 基于 Lucene 的搜索引擎的研究与应用 [J]. 计算机技术与发展，2013，23(6)：230-232.

（下转第 124 页）

图 5 工商综合监管平台界面

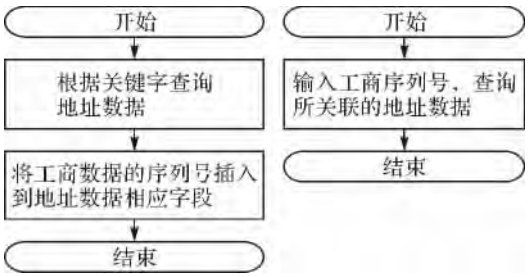

图 6 法人数据匹配流程

部门已有的综合监管平台，以 3 部门业务协同为典范，解决了之前各部门地名地址数据坐标、标准不统一的难题。通过这次尝试探索了多部门间的地名地址业务协同的技术和机制，基本实现了地名地址数据的自生长和实时更新。但是，地理信息技术在业务协同中的应用不可能一蹴而就，今后将结合物联网、移动技术、时空信息云平台等新技术，串联起其他部门，最终实现基于“一张图”的应用，真正解决信息孤岛的难题。

参考文献:

[1] 朱建伟，王泽民.地理编码原理及其本地化解决方案[J].北京测绘，2004(2)：24-27.

[2] 徐开明. 地理信息公共服务平台建设与现代测绘服务模式[J].地理信息世界，2016，4(3)：41-48.

[3] 刘宏伟，严妍.快速响应码的识别和解码[J]. 计算机工程与设计，2005，26(6)：1560-1562.

[4] 刘刚，魏锋. 基于 LDP 算法的手写数字串切分[J]. 北京邮电大学学报，2003，26(1)：14-18.

[5] 任金昌，赵荣椿，张炜. 一种快速有效的印刷体文字识别算法[J]. 中国图象图形学报，2001，6(10)：1011-1015.

[6] 邹崇尧，朱贵方，赵双明.基于搜索引擎技术的地名地址定制查询研究[J].测绘通报，2014(8)：92-94.

[7] 张雪英，闾国年，李伯秋，等.基于规则的中文地址要素解析方法[J].地球信息科学学报，2010，12(1)：9-16.

[8] 马照亭，李志刚，孙伟，等.一种基于地址分词的自动地理编码算法[J].测绘通报，2011(2)：59-62.

[9] 王丙义.信息分类与编码[M].北京：国防工业出版社，2003.

(上接第 109 页)

[10] 梁明，罗荣，胡最. 基于 Lucene 和 PostGIS 的地图搜索研究[J]. 测绘通报，2014(11)：42-45.

[11] 柴洁. 基于 IKAnalyzer 和 Lucene 的地理编码中文搜索引擎的研究与实现[J]. 城市勘测，2014(6)：45-50.

[12] 陈德权. GIS 地名搜索系统的关键技术设计与实现[J]. 测绘与空间地理信息，2013，36(8)：58-60.

[13] 黄承慧，印鉴，陆寄远. 一种改进的 Lucene 语义相似度检索算法[J]. 中山大学学报(自然科学版)，2011，50(2)：11-15.

[14] 郝娟，杨静. 采用上下文特征匹配的中文机构名简称识别[J]. 小型微型计算机系统，2015，36(7)：1432-1437.

[15] 任树怀. LUCENE 搜索算法剖析及优化研究[J]. 图书馆杂志，2014，133(12)：17-23.

[16] 陈换新，孙群，肖强，等. 空间数据融合技术在空间数据生产及更新中的应用[J]. 武汉大学学报(信息科学版)，2014，31(1)：117-122.