基于字符余弦相似度的地址数据治理方法

张 帆[①] 邓 慧[①] 李庆丰[①] 张治国[②] 梁会营[①*] 夏慧敏[①]

[文章编号] 1672-8270(2019)10-0094-04 [中图分类号] R-058 [文献标识码] A

[摘要] **目的**：探讨基于字符余弦相似度的地址数据治理方法，为医院病案室、传染病报病及科研统计分析提供患者的地址清洗数据。**方法**：使用字符的余弦相似度评估患者地址与标准数据集的相似条目，选取相似对最高的前10条地址后，通过弹性距离评估将匹配最好的第1个地址作为映射地址，若无合适地址则以“不详”进行地址标记，供患者下次就诊时更正。**结果**：经过人工复核，每200名患者手工填写的住址以95%置信区间(95%CI)可以正确修复170～186个地址；修复错误的地址多为患者填写的“某街道”，而标准数据集中尚无该街道名称，对其关键词进行过滤可以进一步提高地址信息修复水平。经由热力图对比，地址修复后能够提供更清晰集中的位置信息。**结论**：通过采用基于字符余弦相似度的地址数据治理方法，拓展一种修复基础数据和进行数据映射的有效方法，可为医院相关部门提供准确的患者基础信息数据资料。

[关键词] 地址；数据治理；余弦相似度；弹性距离；信息增益

DOI: 10.3969/J.ISSN.1672-8270.2019.10.026

Data governance method about address based on the similarity of character cosine/ZHANG Fan, DENG Hui, LI Qing-feng, et al//China Medical Equipment,2019,16(10):94-97.

[Abstract] Objective: To discuss the data governance method about address based on the similarity of character cosine so as to provide the cleaning data about addresses of patients for medical records room of hospital, reporting of infectious disease and statistical analysis of scientific research.**Methods:** The character-based cosine similarity was used to assessthe similar items between the addressesof patients and the standard data set. Selected the top 10 pair addresses withthe highest similarity, and the first address with the best matching was used as mapping address through the assessment for elastic distance. If there was no suitable address, the address was marked with "unknown" so as to becorrected bypatientswhen they come to the hospital in next time.**Results:**After manual review, there were 170-186 addresses could be repaired in 95% confidence interval (95% CI) of 200 addresses that were written by 200 patients. The reasons of most of wrong addresses were that patients wrote "a street" while there was no name of this street in standard data set. Through filtered key word could further enhance repaired level of address information. Through the comparison of thermodynamic diagram, the repaired data could provide more clear and centralized location information. **Conclusion:**The data governance method that adopts character-based cosine similarity can develop an effective method for repairing basicdata and implementing data mapping, and can provideaccurate data about basic information of patients for relevant departments of hospital.

[Key words] Address;Data governance;Cosine similarity;Elastic distance;Information gain

[First-author's address] Guangzhou Women and Children's Medical Center, Guangzhou 5100623, China.

地址信息的准确性对于病案上传、传染病上报、地理信息系统(geographic information system，GIS)分析、医院内部科研及发病地区统计至关重要。目前，多数医院对于患者地址信息的获得主要依赖于患者主动填写的地址，信息从挂号入口进入医院内部系统后，各使用部门所对照的规范数据集不相同。通常，病案上传要求5级地址，而传染病报病只需4级地址，两者包含的行政街道并不完全一致，每年更新的时间也不一致。为了实现数据治理的目标，本研究提出基于字符余弦相似度的地址数据治理方法，将患者填写的原始地址在不同的数据规范集进行数据映射[1-6]。

1 资料与方法

1.1 一般资料

待清洗的数据来自2018年1月至2019年4月广州市妇女儿童医疗中心的挂号患者，按挂号时间先后选用了其中100万例患者的地址信息作为分析集。

1.2 相关技术

将患者提供的地址信息中残缺地址与标准地址进行字符编码，通过余弦相似度以及弹性距离找到合适的映射地址。从热力图上看，数据治理后的地址比治理前提供了更多的信息增益，并且达到了90%的修复率。

1.2.1 余弦相似度算法

在两份文本相似度上，“余弦相似度”算法[7]被广泛使用。通过将两个文本出现的单词，建立A、B两个向量，并且计算这两个向量的余弦值。其数学表达为公式1：

$$Cos\_Similarity(A,B)=\cos(\theta)=\frac{A\cdot B}{\|A\|\|B\|} \quad (1)$$

式中A、B为文本编码后的向量；cos_Similarity(A,B)∈(0,1)代表了为两个文本的相似度，当两份文本越相

①广州市妇女儿童医疗中心数据中心 广东 广州 510623
②广州知汇云科技有限公司 广东 广州 511458
*通信作者：lianghuiying@hotmail.com
作者简介：张帆，男，(1985-)，本科学历，研究员，研究方向：数据分析。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

似，越接近1。

1.2.2 弹性距离算法

由于余弦相似度算法本身是在向量空间里面计算两份文本的相似度，并不考虑字符的前后次序关系。为了表示这一特征，并且最大程度从患者填写地址中提取连贯的片段信息，定义“弹性距离”这一数学量，其概念表达见图1。

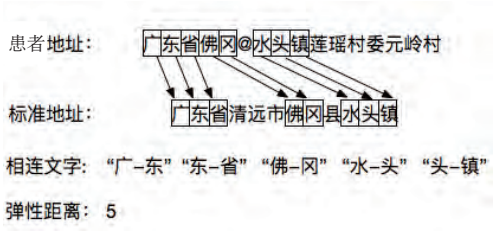


图1 弹性距离计算方法

将患者地址映射到标准地址相同的文字上，再检查标准地址上有多少个文字是相连的，最后数出相连的文字对数，标记为弹性距离。若有多种映射方式，则选取最大的作为映射距离，其具有同时表示文字的次序和文字的紧密程度的优点。即便患者写出“北京市朝阳区广州市”的地址信息，也能正确地根据“北京市朝阳区”这段紧密相连的文字得到更大的权重，排除“广州市”相对于标准地址无关的信息干扰。

弹性距离数学表达设患者地址文字序列为$A=a_1a_2\cdots a_i$，标准地址序列为$B=b_1b_2\cdots b_j$，患者地址文字第l块片段$C_l=c_{l1}c_{l2}\cdots c_{lk}$，其中$C_l\in A$且$C_l\in B$，则弹性距离为公式2：

$$D=\Sigma(length(C_l)-1) \quad (2)$$

因为文字片断匹配的过程满足动态规划的后无效性要求，所以可以使用动态规划方法求解。

1.2.3 地址综合得分

计算患者地址对于每个标准地址的余弦相似性，将结果按高分排在前面，取前10个相似的地址，再对10个地址分别计算弹性距离，并将弹性距离的得分在10个地址的组内归一化，再与余弦相似性相加作为10个地址的最后得分，取最高分作为映射地址：

$option_{address}=Head10\_address(Cos\_Similarity$

(患者地址，校准地址表)

$normallize_{score}=normallize(D(option_{address}))$

$best\_{address}=max(Cos\_Similarity+normallize_{score})$

1.3 研究方法

1.3.1 数据特点

患者手填地址有着和文本不同的特点。文本除虚词“的”、“了”外，多数词汇出现次数不止一次，因此可从词汇整体出现的概率去提取表示位置的实体，但地址信息除省、市、区前三级在标准地址表里出现频繁外，第四级的镇、乡、县、街道的命名则种类繁多，大量名词只在标准库中出现一次，而由于出现频率低，故无法有效提取与位置实体有关的词汇。对100万名患者填写的地址进行分析，发现标准地址具有统计特点：地址的长度一般为(21±6)个字符，当>2个标准差以后(9个字符、33个字符)，地址的质量堪忧，其统计特点无论是统计5000个地址，还是100万个地址，皆保持稳健。多种患者的不良地址类型见表1。

表1 不良地址类型举例

不良地址	问题类型
沿江路；河北；洋青	缩写
城西花园6栋505	以小区名字替代
44000000省44010000市44010500县440105013石溪南约183号	记录错误
*；—；0；同上	无意义文字
广东省广州市+	错误字符
新街1号；古市	对应多个地址
桥村16组；石庵小学	自编地址
gz	文字拼音缩写
南华街道	标准地址无对应

对于标准地址表《全国街道乡镇级以上行政区划》的分析则表现出以下规律：①规范地址使用3100个汉字，除建设兵团用编号标注外，极少使用阿拉伯数字；②使用“省”“市”“镇”“县”“乡”“州”“街道”“区”“农场”“建设兵团”“团”“单位”“公司”“场”“管理处”“园”“管委会”“自治州”等文字进行级别划分；③在个别地区并不使用省、市、区这种类型的行政区划，如“建设兵团直属单位兵团机关”。

1.3.2 数据处理流程

构建一份字向量字典，将规范地址表中的文字进行one-hot编码。通过字向量字典对患者填写地址与规范地址进行编码，去除数字和空格等非规范字符后，获得患者地址与规范地址向量；依次按算法处理患者地址，得到唯一地址，其流程见图2。

1.3.3 统计学方法

采用统计学软件IBM SPSS Statistics19.0对数据进行统计分析，数据不符合正态分布，采用t检验方式。

2 结果

通过地址质量较差的标准(字符数<9个，>33个)

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

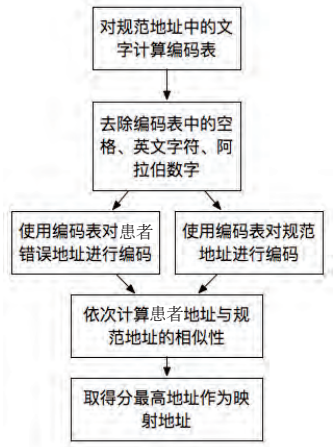


图2 数据处理流程

从100万名患者地址中筛选58 900个地址作为数据集。通过随机抽取样本，并且人工判断的方法检查算法的处理情况，通过*t*检验[7]进行估计，每200个地址以95%置信区间(confidence interval，CI)可以正确修复170～186个地址，见表2。

表2 不良地址修复统计

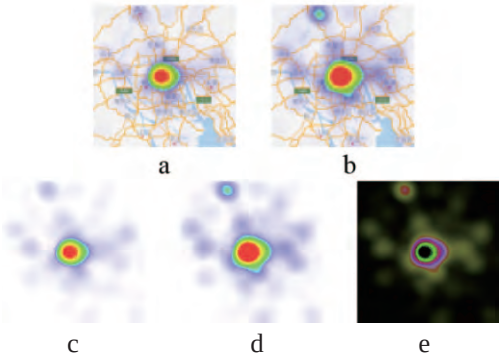
抽样批次	抽样数量(个)	修复后地址映射正确(个)	错误率(%)	修复率(%)
1	200	172	0.14	0.86
2	200	181	0.095	0.905
3	200	185	0.075	0.925
4	200	183	0.085	0.915
5	200	172	0.14	0.86

未能修复的地址在于患者使用了“街道”一词进行地址描述，并且该街道未出现在规范地址集里面。尽管余弦相似度与弹性距离得分均较低，但按算法流程依然筛选了一个映射地址，因此，对关键词进行过滤可以增强地址修复率。地址经过治理后呈现出良好的有序对照关系，方便病案室、科研人员进行下一步处理，并且在不同的数据规范中相互映射也有着良好的效果。对于列为“不详”的地址，医院在患者下次来院时给予提示，请患者填入现住址进行更正，见表3。

表3 地址治理后修复结果

患者地址	修复为
13580462495	不详
中国外籍null	外籍
44000000省44010000市44010600县440106001801	不详
洋青	广东省湛江市遂溪县洋青镇
广东省东莞市东莞本地东城街道办事处@广东省	广东省东莞市莞城街道办事处
*	不详
清远市清城区小市	广东省清远市清城区

从58 900的低质量地址中随机选取4800个地址进行修复，并于修复前后通过“百度地图开放平台”进行热力图信息增益对比[8-10]。当地址未治理前，根据热力图绘制的原理，地址被分散到各个不同的经纬坐标当中，结果不明显。而地址数据治理以后，因为更多的地址被归总映射到同一个经纬坐标下，因此热力图明晰可辨，见图3。



注：①图中采用Pixelmator图层叠加Exclusion绘制，图e=图c+图d-(图c×图d)/128；②a为地址修复前(含地图)，b为地址修复后(含地图)，c为地址修复前(不含地图)，d为地址修复后(不含地图)，e为信息增益

图3 治理后信息增益热力图

3 结论

本研究提出的简单且有用的地址映射算法对患者地址数据进行治理，由于“余弦相似性”及“弹性距离”均为一种统计上的先验规律，与其他先验规律得到的知识一样，其算法也有其限制因素：①当患者使用道路作为登记地址的时候，因为道路名称可能在全国其他省份使用，因此需要结合更多的信息进行校准；②当患者写错别字或字序颠倒时，需要先对原始数据进行清洗；③患者使用了楼盘的名称或未进入标准地址的道路名称时，算法无法找到正确的对应地址，而要结合GIS寻找该患者地址最大可能的位置，再匹配行政街道进行修正。

本研究拓展与改进患者地址数据治理规律的适普性，并从不同角度加以完善，可以使得患者地址数据治理工作更加完整。

参考文献

[1] 费晓璐，李嘉，黄跃，等.医疗大数据应用中的数据治理实践[J].中国卫生信息管理杂志，2018(5)：554-558.
[2] Holmes JH，Elliott TE，Brown JS，et al.Clinical research data warehouse governance for distributed research networks in the USA：A systematic review of the literature[J].J Am Med Inform Assoc，2014，21(4)：730-736.
[3] Hripcsak G，Bloomrosen M，Flatelybrennan P，et al.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

CT应用质量检测结果与分析*

庄晓璇① 刘鸿翔① 袁 田① 尚敬轩① 杨可邦① 胡红波①*

[文章编号] 1672-8270(2019)10-0097-03 [中图分类号] R812 [文献标识码] A

[摘要] **目的**：分析与探讨医疗机构大型医疗设备X射线计算机断层扫描(CT)应用质量管理现状，并提出针对性建议。**方法**：按照国家标准《X射线计算机断层摄影装置质量保证检测规范》，采用CT性能体模Catphon 500等检测工具，对CT设备的CT加权剂量指数、定位光精度、层厚偏差、CT值线性和对比度标度、空间分辨力、密度分辨力、场均匀性、噪声和水的CT值9项性能进行检测。选取2018年南部地区6个省份31家医疗机构在用的52台CT设备，应用质量检测中的检测数据和检测结果进行统计与分析，针对CT质量管理现状提出重视医疗设备日常维护和加强专业人才队伍建设的建议。**结果**：52台受检设备中，16台CT设备所有检测均合格(占30.80%)；结合性能检测结果和临床摄片评估结果，不合格CT设备有3台(占5.77%)；CT设备缺乏日常维护与调试和技术人员专业水平低，是造成CT质量检测不合格的主要原因。**结论**：医疗机构应重视医疗设备日常维护和调试，制定CT的保养和校准周期，建立规范化的操作流程，同时加强专业人才队伍建设，提高专业水平。

[关键词] 大型医疗设备；X射线计算机断层扫描；应用质量检测；质量保证

DOI: 10.3969/J.ISSN.1672-8270.2019.10.027

Testing results and analysis of application quality of CT/ZHUANG Xiao-xuan, LIU Hong-xiang, YUAN Tian, et al//China Medical Equipment,2019,16(10):97-99.

[Abstract] Objective: To analyze and discuss the current situation of application quality managementof X-ray computed tomography (CT) of large medical equipment in medical institutions and put forward some corresponding suggestions for them. **Methods:** According to the national standard < The quality assurance test specification of X-ray computed tomography device>, adopting the CT performance phantom (Catphon 500) and other detection tools to implement detection for 9 performances of CT value of CT equipment included CT weighted dose index, precision of location light, deviation of thickness, linear of CT value, contrast scale, resolving ability of space, resolving ability of density, field homogeneity, noise and water. 52usedCTdevicesof 31 medical institution from 6 southern provinces in 2018 were selected, and detection data and detection results of them in quality detection were applied to implement statistics and analysis. Andaimed at the current situation of CT quality management to put forward some suggestions that paid attention to the routine maintenance for medical equipment and strengthened the construction of professional team.**Results:** In 52 tested CT devices, all of detection items of16 CT equipment were qualified (accounting for 30.80%). Combined with the results of performance testing and the results of clinical photographassessment, there were 3 CT devices were unqualified (accounting for 5.77%).The lack of routine maintenance and debugging of CT devicesand the low professional level of technical personnel were the main reasonsthat leaded to the disqualification of quality testing of CTdevices.**Conclusion:** The medical institution should pay more attention to the routine maintenance and debugging of medical equipment, and formulate the cycle of maintenance and calibration of CT equipment, and establish the standardized operation processof CT equipment. At the same time, the construction of professional talent team should be strengthened so asto improve their professional level.

[Key words] Large medical equipment; X-ray computed tomography (CT); Detectionof application quality; Quality assurance

[First-author's address] Station for Supervision and Inspection of Drug and Instrument, Guilin Joint Logistic Support Center, Guangzhou 510080, China.

*基金项目：国家重点基础研究发展计划(973计划)(2016YFC0103400)“医学成像与放射治疗中的质量控制体模研发”

①桂林联勤保障中心药品仪器监督检验站 广东 广州 510081

*通信作者：huluobu04@gmail.com

作者简介：庄晓璇，女，(1992-)，本科学历，助理工程师，研究方向：生物医学工程。

Health data use,stewardship,and governance:ongoing gaps and challenges:a report from AMIA\"s 2012 Health Policy Meeting[J].J Am Med Inform Assoc,2014,21(2):204-211.

[4] Mcglynn EA,Lieu TA,Durham ML,et al. Developing a data infrastructure for a learning health system:the PORTAL network[J].J Am Med Inform Assoc,2014,21(4):596-601.

[5] 徐衡.流动人口肺结核病例在国家网络报告中的地址讨论[J].江苏卫生事业管理,2017,28(2):153-155.

[6] 王玉贵,朱雪.谈病案首页中病人地址填写存在的问题及对策[J].中国病案,2013,14(9):27-28.

[7] Singhal A.Modern information retrieval:A brief overview[J].Bulletin of the IEEE Computer Society Technical Committee on Data Engineering,2001,24(24):35-43.

[8] 方积乾.现代医学统计学[M].北京:人民卫生出版社,2002.

[9] 吴志强,叶锺楠.基于百度地图热力图的城市空间结构研究—以上海中心城区为例[J].城市规划,2016(4):33-40.

[10] 熊雪晨,金超,白鸽,等.住院病人空间分布可视化表达与实证研究[J].中国卫生资源,2016,19(4):284-288.

收稿日期：2019-01-24

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net