

相似重复数据检测的数据清洗算法优化

蒋园, 韩旭, 马丹璇, 罗登昌

(长江水利委员会 长江勘测规划设计研究有限责任公司, 湖北 武汉 430010)

摘要:数据一直是各大企业竞争的对象,而企业在采集、处理以及最终录入数据库的数据中往往存在着相似重复的数据,这些数据也即“脏数据”。脏数据如果不进行处理,势必会影响后续数据的操作,最终影响到数据的质量。数据清洗是处理脏数据、提高数据质量的热门技术手段,而其中相似重复数据检测更是数据清洗中的重要方面,比如堤防工程的数据存在很多地名、经纬度、砖孔数据等等,录入到数据库时相似重复度很高。目前针对重复数据检测应用最多的是 SNM(基本邻近有序法)算法,主要是先将原有的数据集进行排序,再比较排序后相邻数据的相识度。但这种算法的时间复杂度很高。文中对 SNM 算法进行优化,首先将数据库记录的属性值进行分类,并结合三区间排序算法进行排序来减少比对范围,最后通过设定属性的权重并求和,根据记录相似度的结果来判断。实验结果证明了该算法的正确性。

关键词:脏数据;相似重复;数据清洗;SNM 算法

中图分类号:TP31

文献标识码:A

文章编号:1673-629X(2019)10-0079-04

doi:10.3969/j.issn.1673-629X.2019.10.017

Optimization of Data Cleaning Algorithm for Similar Duplicate Data Detection

JIANG Yuan, HAN Xu, MA Dan-xuan, LUO Deng-chang

(Changjiang Institute of Survey Planning Design and Research, Changjiang Water Resource Commission, Wuhan 430010, China)

Abstract:Data has always been the object of competition for large enterprises, and enterprises often have similar and repeated data in the data collected, processed and finally entered into the database, which is also known as “dirty data”. If these dirty data are not processed, they will affect the operation of subsequent data and ultimately affect the quality of data. Nowadays, data cleaning is a popular technical method to improve the data quality by processing dirty data, and similar duplicate data detection is more important in data cleaning. The data of many place names, latitude and longitude, brick hole data and so on are highly similar in Dike database. At present, the most widely used application for repeated data detection is the SNM (basic proximity ordered method), which mainly sorts the original data firstly, then compares the acquaintances of the adjacent data. However, the time complexity of this calculation is very high. In this paper, by optimizing the SNM algorithm, the database records are first classified to reduce the comparison range, and then the weight of the attributes is set to detect and judge the similarity of the records. Finally, an example is given to prove the correctness of the algorithm.

Key words:dirty data; similar repetition; data cleaning; SNM

0 引言

数据清洗专家研究学者对于数据清洗并未给出标准化的定义^[1],对它的解释大多基于字面意思,理解为通过某种技术方法发现数据集中错误或者不一致的数据也即脏数据并将其进行改正或者剔除^[2]。数据清洗主要有数据的标准化、解析、增强以及重复数据的归并^[3]。标准化主要是将数据在格式上和表达方式上分别进行规范化和同一化;解析针对字段的拆分以及合

并;增强主要是将原有缺少的数据进行补充以达到完整性^[4];重复数据的归并主要是针对相似重复的数据或者进行合并或者进行剔除^[5]。

数据清洗技术现在不仅在大数据互联网企业的海量数据挖掘中得到了广泛应用,而且很多工程数据也用到了数据清洗技术^[6],比如堤防工程数据。堤防工程数据中很多地名、经纬度数据不仅数据位数多、复杂,而且很多数据命名极易令人混淆,录入系统后出现

收稿日期:2018-11-20

修回日期:2019-03-22

网络出版时间:2019-06-26

基金项目:国家重点研发计划课题(2017YFC1502601)

作者简介:蒋园(1990-),女,硕士,初级工程师,研究方向为通信与信息系统。

网络出版地址:<http://kns.cnki.net/kcms/detail/61.1450.TP.20190626.0829.028.html>

大量相似重复的数据,人工还不容易发现和剔除。这些存储在数据库中相似重复数据的存在不仅会浪费系统有限的存储空间,并且还会影响后续对一些地理检测数据情况的正确判断,因此相似重复数据的处理是一项重要的任务。

目前针对相似重复数据的算法有很多,比如 N-Grams 算法、聚类算法、排序和合并的算法等等^[7]。N-Grams 算法主要通过计算得到一个能代表数据记录的属性键值,并且键值生成一个映射数据库的哈希表,然后按照一定的算法,分析哈希表、计算哈希值、根据哈希值来判断记录之间的相似度,最后判定是否为相似的数据^[8]。聚类算法主要通过无监督的学习、迭代计算将相似或者相同的数据聚合到同一类中,同类数据相似度大,但不同类间的数据差异大^[9]。排序和合并算法基于聚类主要分为以个步骤:先依据某个特征将数据库中的数据表进行排序即分类,让相似数据记录聚集在小范围中,再在小范围内对记录进行相似度比较。排序和合并算法包括 SNM 算法(邻近排序算法)、MPN 算法(多趋近邻排序算法)以及优先权队列算法^[10]。多趋近邻排序算法可以将相似重复的数据聚集在较全的集合中,但该算法不仅要多次使用 SNM 算法还得多次创建不同的排序关键字^[11];优先权队列算法虽然可以将不同的相似重复记录放到不同队列中,从而减少比对的次数,但它使用的是单趟的优先权队列算法容易漏配,如果采用多趟优先权队列算法又造成计算时间过长。

1 改进算法 SVN

1.1 相关概念

在数据清洗中衡量一个算法对于相似重复数据的检测效率主要根据召回率以及误识别率^[12]。

准确率:对于一个数据集合,运用算法后实际检测出来的相似重复的记录数目占原有的相似重复的记录的比例^[13]。

全面率:一个数据集合运用算法后正确地检测出相似重复的记录占已经运用算法检测出相似重复的记录的比例,这个比例越低越好^[14]。

相似重复记录:如果在数据库系统中两条记录之间属性值高度相似或者相同,则认为是相似重复记录^[15]。

阈值:通常用 U 表示,判断一个效果临界值。文中当计算出来的相似度高于 U ,则认为是相似重复记录,小于则认为不是^[16]。

1.2 SNM 算法

SNM 算法是在数据清洗中针对相似重复数据问题应用比较成熟的算法,主要分为三个步骤:

(1)排序关键字的创建。首先在数据表中提取关键的属性或者属性的组合来区分记录,依据这些提取的属性或者属性的组合对记录进行划分时具有很强的区分度^[17]。

(2)按照创建的关键字也即关键的属性,将数据库中不同位置的相似重复的记录分配到相邻的位置也即顺序排序^[18]。

(3)针对数据集中的数据,设定一个具有一定宽度比如 X 大小的滑动窗口,窗口采用先进先出的队列方式来处理记录,第 X 进来的记录会与窗口中第一到 $X-1$ 的记录进行逐个比较,每比较完一条记录窗口就会向后移动,如果在比较的过程中存在两条相似重复的记录则会进行相应的合并处理。这种算法对检测速度有很大的改善,因为它的主要计算次数也只有 $N * (X - 1)$,实际中 X 不大,但是算法还是存在如下缺陷:过于依赖关键字的选取,关键字选取的好坏将直接影响后面的检测正确度; X 的大小影响计算的次数;滑动窗口中相似记录的判别多数采用笛卡尔积的计算方式,计算时间较长且最重要的是针对已经存入到数据库(如关系型数据库中的数据记录)直接运用 SNM 算法存在很大的弊端。因为数据记录中的数据随着时间的变化会不停地添加新的数据记录,如果此时还采用 SNM 排序算法进行比较,则会将旧数据再比较一次,这样会造成计算的浪费。因此,文中针对关系型数据存储的特点,将 SNM 算法进行优化来对相似重复数据进行清洗。

1.3 SVN 算法思想

文中提出的改进 SVN 的算法思想主要是避免重复比较过去的旧数据,从而减少计算量。处理过程如下:

(1)因为数据库的记录中并非所有的记录都有唯一的主键,可以对每个记录设定唯一的 id 编号,只不过这个标号值的权重设为 0 并不纳入到后面的算法计算。

(2)创建关系记录属性表。首先依据表的属性创建各种不同的属性记录库,再在属性记录库中针对每个属性创建属性值记录表,这些表中记录各个不同的属性值以及对应这个属性的记录。这里类似于倒排索引的思想,“单词”对应“文档”的倒排索引链表结构^[19],其中的记录可以用 id 编号表示

(3)设定数据记录为 $R = \{R_1, R_2, \dots, R_l\}$,其中 l 表示数据记录的个数,相应的某个记录为 $R_i (1 \leq i \leq l)$,并且两个记录在某个属性(比如 p 属性)之间的关系用式 1 表示:

$$M = A(R_{ip}, R_{jp}) = \{0, 1\} (1 \leq i \leq l, 1 \leq j \leq l, 1 \leq p \leq n) \quad (1)$$

其中, $R_{ip} = (ID_1, ID_2, \dots, ID_m)$, ID_m 表示针对某个属性值相似重复的相似记录; n 表示某个记录具有 n 个属性值的长度; p 为记录的第 p 个属性值。

当 $M=1$, 则判断两个记录在 p 属性具有很近的关系是重复关系, 当为 0 则判断在当前的属性下没有相似关系。依据此种关系就可以计算两个记录在某个属性值记录表下的相似度, 以及记录每个记录在此属性值记录表中的值。设定相似度如下:

$$Q_{ij,p} = M * W_p \quad (2)$$

其中, W_p 为记录中 p 属性的权值。 $\sum_{i=1}^n W_p = 1$ 计算完成后, 再到其他属性值记录表中将每个记录的属性值求相似度。最后将每个记录的属性相似度进行累加求和即为两条记录的相似度:

$$Z_{i,j} = \sum_{p=1}^n Q_{i,j,p} \quad (3)$$

其中, $Z_{i,j}$ 表示为 i, j 两条记录的相似度值。

(4) 设定相似阈值 U , 如果式 4 成立, 就可以判断两者之间的相似度。

$$Z_{i,j} \gg U \quad (4)$$

(5) 属性值记录表的数据在进行比较的过程中, 并不是对属性值记录表进行顺序比较, 如果从头到尾比对会增加时间长度。文中引用快速排序算法先进行排序再进行比较, 经过从大到小排序或者从小到大排序后, 一旦比较发现了有相等的属性值后出现不相等的值, 后面的数据就不用再进行比较, 从而缩短了计算时间。传统的快速排序算法主要将从数据组中随机选取一个数作为比对数, 再将其他数据与比对数进行比较, 也即分为两段比较方法, 比它小的数放在左边, 比它大的数放到后边, 最后多次递归比较来达到最后的排序。但是因为文中针对的是重复相似的数据, 所以很显然属性值记录表中存在很多的重复数据, 对这些重复数据再一次进行排序就会浪费计算时间。文中采取三区间比较方法让重复的数据不需要再参与排序, 只让大于或者小于选定的数参与排序。比如选定的数为 x , 只对大于 x 或者小于 x 的序列数进行排序, 如图 1 所示, 从左到右扫描使得指针 lt 维护 $[lo \dots lt]$ 中的数字比 x 小, 另外的指针 gt 维护 $[gt+1 \dots hi]$ 的数字大于 x , 以及指针 i , 使得所有 $[lt \dots (i-1)]$ 的数字与 x 大小相等。 $[i \dots gt]$ 之间的数字是还没有进行处理的数字, i 从 lo 开始进行扫描:

如果 $a[i] < v$, 那么交换 $a[lt]$ 和 $a[i]$, 相应的 lt 指针和 i 指针自增;

如果 $a[i] > v$, 那么交换 $a[i]$ 和 $a[gt]$, gt 指针自减;

如果 $a[i] = v$, i 指针自增。

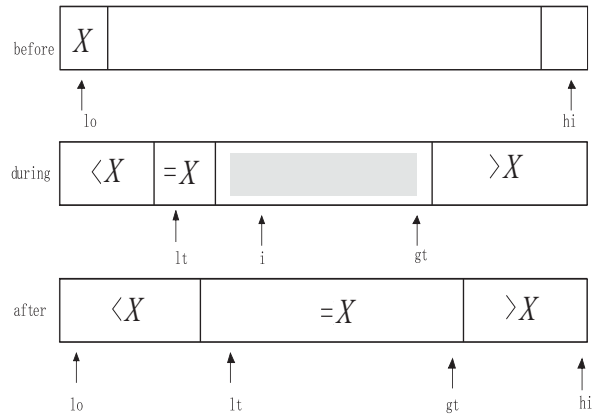


图 1 三区间快速排序示意

1.4 算法的实现步骤

(1) 将数据库中的数据记录标记相应的属性记录标号, 并且在数据库中增加属性记录库, 对应的库中针对每个属性增加属性值记录表。

(2) 将数据库中录入的所有的堤防工程数据记录对应属性找到属性值记录表, 将属性值和记录编号通过编码程序录入, 从而完成属性之间的关系表。

(3) 拿到当前记录, 依据当前记录的属性值从属性记录库中找到对应的属性记录表, 依据属性记录表的值读取经过三分区快速算法排序后的数据属性值, 通过对比来判断当前属性是否相似重复。当相似重复, 则用 1 乘以相应的权重, 当不重复, 即为 0, 不用乘以相应的权重, 并且将当前的记录分别和每一个属性值记录表进行比较计算, 最后将每个记录的属性相似度求和即为最终的记录相识值。

(4) 根据相似值和规定阈值之间的大小来判断两个记录之间的相似度大小。

(5) 当新增记录时, 将记录的值增加到属性库中的每个属性值记录表, 然后只需要重复步骤 3~5, 计算新增的记录和属性表中其他记录之间的相似度。

2 实验与结果分析

在相同的实验环境下, 应用传统的基于 SNM 算法的数据清洗方法和优化后的 SVN 算法, 分别对某存储有 7 万条数据记录的堤防工程的数据库进行相似重复数据的数据清洗处理, 记录表中包含 7 条属性值, 实验中随机测试 3 万, 7 万数据记录, 并且人为地在每组里面设置 245, 3 487 条重复记录数据, 统计分别在两种算法下的准确率和全面率以及添加新记录后的运行时间。

2.1 实验过程

分别对传统 SNM 算法和改进后的 SVN 算法在准确率以及全面率上进行对比, 并且对比添加记录后运行时间, 结果如表 1 和表 2 所示。

表 1 准确率和全面率的对比

度量标准	SNM		SVN	
	3 万	7 万	3 万	7 万
准确率/%	82.5	79	95	85
全面率/%	81	80	100	100

表 2 记录添加运行时间

度量标准	SNM		SVN	
	3 万	7 万	3 万	7 万
查询时间/s	220	310	90	280

2.2 结果讨论

通过表 1 可以发现,在针对准确率和全面率时,改进后的 SVN 算法相较传统的 SVN 算法有明显的改善,虽然随着数据量的增加 SVN 在准确率上是下降的,但是改善后的还是较改善前的高。

表 2 表明,当添加记录时改进后的 SVN 算法的运行时间还是比之前的要小。因为传统的 SVN 算法需要将所有的数据进行新旧比对,这会消耗大量的计算时间,虽然优化后的 SVN 算法需要将属性表中的数据进行排序,还要添加新的属性到属性表中增加了时间复杂度,但文中的三区间快速算法已经较之前的快速算法有了很大的改善。而且本来增加的记录的数据量相比已经存在于数据库中的记录已经很少了。因此,改进的 SVN 算法是有效的。

3 结束语

无论是人工录入还是程序读入数据到数据库中,总会存在很多相似重复的记录数据,如果单靠数据库管理员来检测,则效率非常低下,尤其是堤防工程数据大多是编码编号,数据位数也较多、复杂,因此好的检测算法非常必要。文中针对传统 SNM 算法在关系型数据库添加新记录后进行相似重复数据检测时需要重复比对旧的数据增加时间复杂度的缺点进行改进,提出一种 SVN 算法。该算法主要通过数据库中增加属性记录库和属性值记录表,然后将数据表中的记录值以及对应的权重添加到对应的属性值记录表中,在这中间还需要通过三区间排序算法将属性值进行从大到小的排序,最后新添加记录时可以将其属性添加到对应的属性记录表中和三区间排序算法排序后的属性值进行属性比对,并且按照相应的权重计算相似度,最后将相似度按照大小进行排序。实验结果表明,该算法提高了相似重复数据检测的准确率和全面率。同时该算法也存在不足之处,因为随着数据量的增加,准确

率跟传统算法相比也没多大改善,而且系统需要额外建立属性库和属性表,当数据量是海量数据时,这将会占用一部分存储空间,这方面尚需进一步完善。

参考文献:

[1] 时天元. 排序算法对比研究[J]. 通讯世界,2018(9):267-268.

[2] 王旭东,段敬,温志坚,等. 基于相似重复记录的 N-Gram 算法的改进与应用[J]. 现代计算机,2018(25):78-82.

[3] 张培根,黄树成. 一种用于中文数据清洗的近邻排序算法[J]. 计算机应用与软件,2018,35(8):286-288.

[4] 宋国兴,周喜,马博,等. 关键属性组的相似重复记录检测方法研究[J]. 科学技术与工程,2017,17(19):65-71.

[5] 董楠楠,于悦,喻兰,等. 一种基于相似性的分布式重复数据删除方法[J]. 中国新通信,2018,20(12):53-55.

[6] 赵月琴,范通让. 科技创新大数据清洗框架研究[J]. 河北省科学院学报,2018,35(2):35-42.

[7] WANG Xuyang,ZHANG Pengyuan,NA Xingyu,et al. Handling OOV words in mandarin spoken term detection with an hierarchical n-gram language model[J]. Chinese Journal of Electronics, 2017,26(6):1239-1244.

[8] 尹薇. 时间序列清洗关键技术的研究[D]. 哈尔滨:哈尔滨工业大学,2018.

[9] 刘齐锐. 基本排序算法研究[J]. 通讯世界,2018(5):295-296.

[10] 邱耀儒,沈明. 浅谈大数据的数据处理[J]. 电子世界,2018(10):100.

[11] 王芳. 中文重复记录清洗的相关算法的研究[D]. 青岛:青岛大学,2018.

[12] ZHANG Ningning. Research on data cleaning method based on SNM algorithm[C]//Proceedings of 2017 IEEE 2nd advanced information technology, electronic and automation control conference. Beijing:IEEE,2017:5.

[13] 孙海雪. 地质大数据发现与文本信息分析[D]. 北京:中国地质大学(北京),2018.

[14] 李超,刘辉. 一种基于关联分析与 N-Gram 的错误参数检测方法[J]. 软件学报,2018,29(8):2243-2257.

[15] 李军. 基本近邻排序算法的改进与应用[J]. 宁夏师范学院学报,2017,38(3):72-77.

[16] 李军. 一种相似重复记录检测算法的改进与应用[J]. 成都工业学院学报,2017,20(2):17-20.

[17] 王江. 数据清洗技术研究及清洗框架的设计与实现[D]. 呼和浩特:内蒙古大学,2016.

[18] 余肖生,胡孙枝. 基于 SNM 改进算法的相似重复记录消除[J]. 重庆理工大学学报:自然科学版,2016,30(4):91-96.

[19] ZHANG Yuejun,WANG Pengjun,LI Gang. An isolated SNM model for high-stability multi-port register file in 65 nm CMOS[J]. Journal of Semiconductors,2017,38(9):72-77.