文章编号:1674-2869(2014)09-0066-04

大数据环境下的相似重复记录检测方法

殷秀叶
周口师范学院计算机科学与技术学院,河南 周口 466001

摘 要:大数据环境下的相似重复记录影响数据统计分析结果的准确性,需要过滤相似重复记录.对相似重复记录检测的研究现状做了介绍,在此基础上提出了属性加权的思想,对属性进行加权,并根据属性权值进行排序分组;在对属性加权时,考虑到一些字段的取值是一一对应的关系,权值相同,提出了同义属性的概念,在原数据集的基础上排除部分同义属性来缩减数据集,提高重复数据检测的效率,最后给出了相似重复记录判定的方法.考虑到大数据集给重复记录检测带来的挑战,将大数据集拆分成若干小数据集,充分利用MapReduce机制进行处理,将大数据集按照权重较大的属性取值进行分组,分割成若干个map任务,分别进行处理.实验结果表明,该方法能够有效地提高相似重复记录检测的效率.

关键词:相似重复记录;大数据;同义属性
中图分类号:TP393 **文献标识码**:A **doi**:10.3969/j.issn.1674-2869.2014.09.013

0 引 言

在信息化的时代,数据是企业成功的关键,而高质量数据是保证企业健康发展的前提,企业在对数据进行统计分析的过程中,难免会受到“脏、乱”数据的影响,在大数据环境下,原始数据中可能存在着大量的相似重复记录,这些记录将会影响数据统计分析的效率与准确性,如何有效的处理这些相似重复记录显得尤为重要[1].

相似重复记录常用的检测方法主要是排序—合并算法和二次聚类算法,这些算法在处理小数据集时有较高的效率,但随着数据量的不断增大,在大数据环境下,这些方法不能有效的提升检测的效率,如在数据量较大时,排序—合并算法有大量的I/O开销.另外,基于字符位置的敏感性问题,对排序的记录不能保证相似记录一定排在临近的位置,不能有效的提升重复记录检测的记忆率(recall,识别出的相似重复记录占整个数据集中所有重复记录的百分比)和准确率(precision,识别出的相似重复记录中正确的识别占识别出的相似重复记录的百分比).

1 相关工作

在相似重复记录检测上,已出现了一些成果,如早期的“排序—合并”算法,首先对记录进行排序,然后比较邻近记录的字段是否相等,在解决重复记录的检测上,有着较高的效率.

文献[2]通过等级法计算每个字段的权值,然后按照分组的思想[2],选择关键字段或字段某些位将大数据集分割成许多不相交的小数据集来解决数据量较大的问题,提高相似重复记录检测效率.

文献[3]提出了一种大数据量的重复记录检测方法,对于检测出的重复记录,保留一条记录作为主记录,其他重复记录中的信息合并到主记录[3].由于比较重复记录需比较记录间所有的属性字段,因此比较次数影响比较算法的效率,作者以减少记录间比较次数的思想,提高算法的效率.但由于有些记录中的字段存在互斥的值(如性别的取值),因此作者考虑了带限制规则的重复记录检测方法,引入了限制规则.

文献[4]通过计算属性的权重,确定每一属性对于记录相似性检测的重要性,然后多线程并发检测记录集,每个线程针对一个属性对记录集进行排序;最后在每个线程中检测相似重复记录并且合并所有的检测结果[4-5].

文献[6]提出了一种q-gram层次空间的聚类检测方法,将数据映射成q-gram空间中的点,并根据空间中的相似性通过层次聚类的方法来检测相似重复记录[6-7],有效的解决了传统的“排序-合

收稿日期:2014-06-12
基金项目:国家自然科学基金青年项目(61103143);周口师范学院青年科研基金项目(zknuc0215)
作者简介:殷秀叶(1984-),女,河南信阳人,助教,硕士.研究方向:大数据的检测效率.

并”算法中部分相似记录由于字符位置敏感而不能排在邻近位置的问题,提高了检测的精度,在大数据环境下有较好的伸缩性.

文献[8]通过对 CURE 算法进行改进,利用预抽样的概念,提高了随机抽样的准确性,并改进代表点的选择方法[8],通过基于距离影响因子的代表点选取策略,提高了代表点选取的合理性,提升了相似重复记录检测的准确性.

以上方法在不同的方面取得了不同的效果,但均未考虑在大数据环境下,充分利用 MapReduce 的思想,提升相似重复记录检测的效率.

2　相似重复记录检测方法

2.1　思想描述

由于在数据集中会存在很多意思相同但表示方式不同的字段,如区域和区号 (北京,010),都表示区域信息,只是一个是用编码的方式,一个是用汉字的方式,这部分字段的取值是一一对应的,在进行相似重复数据检测计算时,权值是相同的,把这种属性称为同义属性.

此种相似重复记录检测的总体思想是:考虑到大数据环境下数据量巨大的问题,在相似重复记录检测中,首先通过加权法计算属性的权重,为了提高属性加权计算的效率,先利用过滤法过滤掉同义属性,然后在计算属性权重及排序时,只按照同义属性中的一个进行权值计算,然后将计算结果直接作为两者共同的权值.

由于大数据环境中有很多这样的字段,多体现为维度字段,因此能够直接减少属性加权计算时的数据集合,提升计算的效率. 另外,由于同义属性比较直观,因此这部分内容可由数据库操作者直接指定,并不会给加权计算带来额外的任务和开销. 权重计算完成后,按照权值大小将数据集合进行排序,然后利用 MapReduce 的思想,将数据集合按照不同的字段拆分成若干个小的数据集合,且相同的字段也划分若干个集合,由 MapReduce 分别利用不同的 Map 任务处理各个小的数据集合中的相似重复字段,最终利用 Reduce 任务合并处理结果,提升重复记录检测的效率.

2.2　权重计算

设数据表 T 中有 m 个属性,用向量 $\boldsymbol{C}=\{C_1, C_2, \cdots, C_m\}$ 来表示,其中 C_m 表示数据表 T 的第 m 个属性,用向量 $\boldsymbol{W}=\{W_1, W_2, \cdots, W_m\}$ 来表示属性的权重,其中 W_m 表示属性向量 $\boldsymbol{C}$ 中第 m 个属性的权重. 考虑到属性的取值变化情况,取值越多,说明权重越大,因此,通过属性取值的变化情况来衡量属性的权重.

考虑到大数据环境下数据量过大的问题,采用随机抽样法,每次选取固定数量的数据来统计在固定数量内每个属性的取值变化个数,设 count_{ij} 为数据表 T 第 i 个属性 C_i 在第 j 次随机抽取的 m 条记录内取值的变化个数,k 为随机抽样的次数,则 C_i 的权重 W_i 可表示为:

$$W_i=\frac{1}{k}\sum_{1}^{k}\mathrm{count}_{ij} \tag{1}$$

假设数据表 T 中的 m 个属性有 $2n$ 个同义属性,则计算权值的数据表 T' 的属性数即为 $T-\{n\}$,即假设数据表中有 20 个属性,而同义属性有 8 个,则在计算权重时,可以去掉至少 4 个同义属性,原本 20 个属性的数据表,现在最多只需要对 16 个属性进行操作,效率能够提升 20%. 同义属性越多,效率提升的也越多.

2.3　相似重复数据处理

对于任意的属性 C_k,其在第 i 条记录的取值与第 j 条记录的取值的相似程度用 $S(C_{ik}, C_{jk})$ 表示,C_{ik} 和 C_{jk} 的取值越接近,则 $S(C_{ik}, C_{jk})$ 的取值越大,$S(C_{ik}, C_{jk})$ 的取值范围为 $[0,1]$,可以用式(2)来表示.

$$S(C_{ik}, C_{jk})=\frac{\sum_{x=1}^{t} se(p, C_{jk})}{|C_{jk}|} \tag{2}$$

其中 $|C_{jk}|$ 表示第 j 条记录中字段 C_k 的长度,$se(p, C_{jk})$ 表示第 i 条记录中字段取值的某一个子串 p 与第 j 条记录中字段 C_k 取值的每一个对应子串的匹配程度,t 代表子串的个数.

对于两条记录 C_i 和 C_j 的相似度,可用如式(3)表示.

$$\mathrm{Same}(C_i, C_j)=\frac{1}{m}\sum_{k=1}^{m} S(C_{ik}, C_{jk}) * W_K \tag{3}$$

其中 Same (C_i, C_j) 表示记录 C_i 和 C_j 的相似度,m 表示关系表 T 中属性的个数,W_k 表示关系表 T 中属性 C_k 的权值.

按照属性的权值对数据表进行排序,权值越大,排序越靠后,然后将大数据集按照属性拆分成若干个小的数据集,利用 MapReduce 的处理机制,将各个小的数据集交给 MapReduce 中的若干个小的 Map 任务去处理,为了提高相似性检测的合理性,数据集的拆分也多次进行,以保证相同的数据能够被分配到一组,提高相似重复字段检测的记忆率. Map 任务处理完成后,通过 Reduce 任务将处理结果合并,最终输出处理后的结果.

3 实 验

为了验证算法的效率,实验采用 Hadoop 平台,将算法与**文献[4]**中提出的算法进行了比较,实验的数据来自于高校半年内学生在餐厅的刷卡信息,共有数据500多万条.

实验采用的环境为64位 AMD 皓龙6274 CPU,196G 的内存,代码采用 JAVA 语言实现,在 Hadoop-1.0.4 上运行.

文献[4]提出的利用等级计算权重,然后利用优先队列算法进行相似重复记录检测的方法是效率较高的一种算法,其思想是:首先利用组合赋权的思想,为每个属性确定一个等级,再根据每个属性的等级计算权重,然后采用多线程并发执行的方式检测相似重复记录,并采用加速法减少多余的编辑距离计算,采用优先队列的重复聚类思想代替固定窗口大小的滑动窗口方法.该算法具有较高的检测精度和较小的时间复杂度.因此,选择该算法(以 IWM 简称该算法)与本文提出的算法(SPM)进行比较,数据量共510万条,共有同义属性3对,字段14个.比较结果如表1所示.

表1 IWM 算法与 SPM 算法的综合比较

Table 1 Comprehensive comparison of IWM algorithm and SPM algorithm

	查准率/%	查全率/%	消耗的时间/min
IWM	90.8	92.3	92.4
SPM	90.6	92.6	68.3

比较结果表明,在相同的机器上,由于本算法利用同义属性降低了数据集,并利用 Hadoop 平台加快了数据检测的效率,因此在时间复杂度上有较好的表现,在算法的查准率和查全率上,两种算法的表现趋向一致.但由于不同数据集合中同义属性存在的数量不同,因此本文提出的算法的时间开销与同义属性的多少有关,同义属性越多,则时间复杂度越低,反之越高,但只要同义属性存在,都能不同程度的提高相似重复数据检测的效率.

4 结 语

以上对相似重复记录检测进行了相关介绍,通过识别同义字段,直接减少了数据集的属性个数,降低了在权重计算阶段的工作量.另外,论文对相似重复数据的判定给出了公式,并将数据集按照属性的权重进行拆分,将拆分后的若干小数据集通过 MapReduce 进行处理,充分提高了相似性重复记录检测的速度.对于同义字段的公式判定,将在后续工作中进行研究.

致 谢

本研究在选题及论文写作的过程中,陈立勇老师提出了很多宝贵的意见,谨致谢意.感谢国家自然科学基金委员会对项目的开展提供的支持.

参考文献:

[1] 李建中,刘显敏.大数据的一个重要方面:数据可用性[J].计算机研究与发展,2013,50(6):1147-1162.
LI Jian-zhong, LIU Xian-min. An important aspect of big data: data usability[J]. Journal of Computer Research and Development, 2013, 50(6): 1147-1162. (in Chinese)

[2] 李星毅,包从剑,施化吉.数据仓库中的相似重复记录检测方法[J].电子科技大学学报,2007,36(6):1273-1277.
LI Xing-yi, BAO Cong-jian, SHI Hua-ji. A method for detecting approximately duplicate database records in data warehouse[J]. Journal of University of Electronic Science and Technology of China, 2007, 36(6):1273-1277. (in Chinese)

[3] 庞雄文,姚占林,李拥军.大数据量的高效重复记录检测方法[J].华中科技大学学报,2010,38(2):8-11.
PANG Xiong-wen, YAO Zhan-lin, LI Yong-jun. Efficient duplicate records detection method for massive data[J]. Journal of Huazhong University of Science and Technology, 2010, 38(2): 8-11. (in Chinese)

[4] 周典瑞,周莲英.海量数据的相似记录检测算法[J].计算机应用,2013,33(8):2208-2211.
ZHOU Dian-rui, ZHOU Lian-ying. Algorithm for detecting approximate duplicate records in massive data[J]. Journal of Computer Application, 2013, 33(8):2208-2211. (in Chinese)

[5] 敖莉,舒继武,李明强.重复数据删除技术[J].软件学报,2010, 21(5):916-929.
AO Li, SHU Ji-wu, LI Ming-qiang. Data deduplication techniques[J]. Journal of Software, 2010, 21(5):916-929. (in Chinese)

[6] 韩京宇,徐立臻,董逸生.一种大数据量的相似记录检测方法[J].计算机研究与发展, 2005,42(12):2206-2212.
HAN Jing-yu, XU Li-zhen, DONG Yi-sheng. An approach for detecting similar duplicate records of massive data[J]. Journal of Computer Research and Development, 2005, 42(12):2206-2212. (in Chinese)

[7] 邱越峰.一种高效的检测相似重复记录的方法[J].计算机学报,2001,24(1):69-77.

QIU Yue-fen. An efficient approach for detecting approximately duplicate database records[J]. CHINESE J. COMPUTERS, 2001, 24(1): 69-77. (in Chinese)

[8] DEAN J, GHEMAWAT S. MapReduce: simplified data processing on large clusters[C]// In Proceedings of the 6th Symposium on Operating Systems Design and Implementation, New York: NY, 2004.

Method for detecting approximately duplicate database records in big data environment

YIN Xiu-ye

School of Computer Science and Technology, Zhoukou Normal University, Zhoukou 466001, China

Abstract: The accuracy of the data statistical analysis is affected by approximately duplicated records in big data environments, so the approximately duplicated records need to be filtered. We introduced the current research of approximately duplicated records and proposed the weighted attribute idea, weighting the attributes and grouping them according to the weights. Considering that some field's relationship is one to one, we proposed synonymous property. We excluded some synonymous property on the basis of the original dataset to reduce the dataset and improve the efficiency of detection of approximately duplicated records . Finally synonymous property was proposed. Big datasets were split into a number of small datasets considering the challenge of approximately duplicated records in big dataset. Taking full advantage of MapReduce processing mechanism, big datasets were grouped according to the weight of the larger attribute values, and then divided into a number of map tasks to process. Experiment shows that this method can improve detection efficiency of approximately duplicated records effectively.

Key words: approximately duplicated records; big data; MapReduce; synonymous property

本文编辑:陈小平