doi: 10.3969/j.issn.1674-8425(z).2016.04.016

基于 SNM 改进算法的相似重复记录消除

余肖生, 胡孙枝

(三峡大学 计算机与信息学院,湖北 宜昌 443002)

摘 要:高质量的数据是构建数据仓库的最重要因素,低质量的数据可能对决策产生不利影响。来自不同数据源的相似重复记录是数据仓库构建中影响数据质量的主要问题之一,在源数据进入数据仓库之前尽可能地消除相似重复记录能很大程度地提高数据质量。为此,比较了现有的相似重复记录消除算法,改进了 SNM 算法,并通过实验比较了传统 SNM 方法与改进 SNM 算法。实验结果显示:在相似重复记录消除方面,SNM 改进算法具有明显的优势。

关 键 词:SNM 算法;SNM 改进算法;相似重复记录消除

中图分类号:TP311 **文献标识码**:A **文章编号**:1674 - 8425(2016)04 - 0091 - 06

Research on Eliminating Duplicate Records Based on SNM Improved Algorithm

YU Xiao-sheng, HU Sun-zhi

(College of Computer and Information Technology,
China Three Gorges University, Yichang 443002, China)

Abstract: High quality data is the most important factor to build the data warehouse. The low quality data may be bad for decision making. An approximately duplicate record from different data sources is one of the main data quality issues to build data warehouse. To eliminate approximately duplicate data as far as possible before the source data enters into a data warehouse can greatly improve the quality of data. Firstly, the existing approximately duplicate records elimination algorithms were compared, and then SNM algorithm was improved. The authors compared traditional SNM method and SNM improved algorithm by the experiment, and the results show: SNM improved algorithm has obvious advantages in eliminating duplicate records.

Key words: SNM algorithm; SNM improved algorithm; approximately duplicate records elimination

收稿日期:2016 - 01 - 18
基金项目:国家自然科学基金资助项目(71473185)
作者简介:余肖生(1973—),男,湖北监利人,博士后,副教授,主要从事信息管理与电子商务研究。
引用格式: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),2016(4):91 - 96.

Citation format: YU Xiao-sheng, HU Sun-zhi. Research on Eliminating Duplicate Records Based on SNM Improved Algorithm [J]. Journal of Chongqing University of Technology(Natural Science),2016(4):91 - 96.

在企业中,各级管理人员需要面对不同层次的大量信息,并需要分析这些信息,以便及时了解市场变化,做出正确有效的判断和决策。为了保证信息的正确性和有效性,企业通常利用长期积累的分散数据构建自己的数据仓库,然后利用数据挖掘工具从企业数据仓库中获得用于支持管理决策的战略信息[1]。由于长期积累的数据往往是海量的和分散的,存在数据错误、数据丢失、格式不统一、规则不一致等多种问题,因此导致从数据仓库挖掘出的信息不能有效地支持管理决策。高质量的数据可能是数据仓库成功的最重要因素[2],而低质量的数据可能对决策产生不利影响[3-4]。在数据仓库构建的众多数据质量问题中,来自不同数据源的相似重复记录占有相对较大的比例。数据仓库中的相似重复记录直接影响着信息的有效性,因此在源数据进入数据仓库之前尽可能地消除相似重复记录能很大程度提高数据质量,对成功构建数据仓库具有深远的意义。

本文比较了现有相似重复记录消除算法,并改进了 SNM 算法。通过实验比较传统 SNM 方法与改进 SNM 算法,结果显示:在相似重复记录消除方面,SNM 改进算法具有明显的优势。

1 现有相似重复记录消除算法的比较

相似重复记录是指对于现实世界中同一个实体,在各个数据源数据库或平面文件中存储时,由于可能出现格式错误、结构不一致、拼写差异等问题导致数据库管理系统没有正确识别而产生的两条或者多条不完全相同的记录[5]。相似重复记录是导致数据仓库构建中数据质量不符合标准的最常见的问题之一,是大部分低质量数据产生的源头。相似重复记录会损害数据的唯一性,产生数据冗余,导致资源浪费。因此,相似重复记录的消除成为数据仓库构建成功的关键因素之一。优先队列算法、Delphi 算法和 SNM 算法是目前常见的消除海量数据环境下数据库中相似重复记录的策略。

1.1 优先队列算法

假设 S 是一个数据集,S 中的记录都有键值,优先队列就是一种关于 S 的数据结构。优先队列包括最大优先队列、最小优先队列,支持 INSERT 等多种操作。优先队列算法中使用优先队列中的元素作为一组记录,每一个元素包含的这一组记录都是属于最新探测到的记录簇中的一部分。算法按照顺序匹配数据库中的记录,判定记录是否为优先队列中相关记录簇中的成员。若是,则扫描下一条;否则,这条记录将和优先队列中的记录进行比较,如果存在重复记录,那么就将该记录合并到匹配记录所在簇。如果不存在重复数据,则将该条记录加入一个新的簇,并进入优先队列,且具有最高优先级[6-7]。

1.2 Delphi 算法

Delphi 算法可用来判定两条或者多条记录是否相似,主要是利用文本相似度函数和共同出现相似度函数来进行相似重复记录的探测,并利用聚合策略减少记录比较次数[8]。对于“winxp pro”和“windows XP Professional”这样的等价错误,其识别效率较高。

1.3 传统 SNM 算法

SNM 算法[9-10]即邻近排序算法。SNM 算法的基本思想是:将数据集 R 中的所有记录按照相应指定的关键词(key)进行排序。绝大部分情况下,经过排序后的数据集中,如果存在相似重复记录,则认为它们是相邻的,且聚集在一定范围内,可在很大程度上提高匹配效率。另外,采用滑动窗口极大地减少了记录比较的次数,提高了比较速度,缩短了匹配时间。

1.4 现有相似重复记录消除算法的比较

综合上述几种常见的消除相似重复记录算法,可知它们各自都有自己的适用范围和应用环境。其优势和不足如表 1 所示。

数据仓库构建过程中,相似重复记录的消除首先要考虑针对海量数据的执行效率,在此基础上对算法进行改进以提高相似重复数据的探测率,得到更好的消除效果,进而提高数据仓库中的

数据质量。通过对几种常见的消除相似重复记录算法的比较,全面分析各自的优势与不足,对 SNM 算法进行讨论和改进。

表 1 几种常见消除相似重复记录算法的比较

消除相似重复记录算法	优势	不足
优先队列算法	利用两次排序增加相似重复记录聚合机会。	海量数据环境下效率较低。
Delphi 算法	利用聚合策略来减少记录比较次数。	对属性值等价的相似重复记录的消除效果不明显;海量数据环境下效率较低。
传统 SNM 算法	匹配时间短,海量数据环境下效率较高。	对长属性值或属性值中子串顺序不一致的情况,聚合效果不明显。

2 基于 SNM 算法的改进与实现

传统的 SNM 算法识别相似重复记录的做法是:对数据预处理后,选定关键属性,然后将记录生成记录字符串,并对其进行排序;排序后按照设定的窗口大小对窗口内记录进行记录匹配;最后根据设定的文本相似度判定是否为相似重复记录。SNM 算法的思想是尽量只对排序后邻近的记录进行匹配,从而大大减少比较次数和缩短比较时间,因此 SNM 算法对相似重复数据的匹配效果的好坏取决于排序后相似重复记录被排在相邻位置的邻近程度,相似重复记录越邻近,匹配效果就越好。然而,在对数据源的数据进行排序时,选择的排序字段不同对排序结果有很大影响。在实际数据中,往往有很大一部分记录的数据值不是单个的单词或词语,而是一个句子,如地址字段。对于属性值为句子之类的数据,如果直接排序,则相似重复记录很可能并非邻近,相反会分离得较远。有时候由于属性值的顺序规则不同,甚至较短的句子也有可能出现类似的问题。例如:有两条主要属性是(Name,Sex,Birthday,Phone,Address)的记录:(Wang Mei,F,1989-10-10,18671745011,Hubei Yichang Xiling University Road),(Mei Wang,W,1989-10-10,18671745011,University Road,Xiling,Yichang,Hubei)。无论按照 Name 属性排序,还是 Address 属性排序,其排序后的结果都会将这两条记录分离得很远,而事实上这两条记录属于重复数据。

笔者将记录字符串单词化分割后再进行排序,较好地弥补了传统算法的缺陷。同样以上述两条记录为例,本文首先对不一致的属性进行预处理,示例中,对 Sex 属性,采用男性为“1”,女性为“0”,将记录中的 Sex 属性做归一化处理;其次选定关键属性(Name,Sex,Birthday,Address),并生成记录字符串分别为“Wang Mei 0 1989-10-10 Hubei Yichang Xiling University Road”,“Mei Wang 0 1989-10-10 University Road,Xiling,Yichang,Hubei”;然后针对记录字符串单词化处理并排序,得到结果字符串分别为“0 1989-10-10 Hubei Mei Road University Wang Xiling Yichang”,“0 1989-10-10 Hubei Mei Road University Wang Xiling Yichang”。经过该处理后的相似重复记录很大程度上增加了聚合的机会,再通过窗口内计算文本相似度就能很容易判定这两条记录是重复数据。因此,对记录字符串单词化处理后再排序能很大程度上将相似重复记录排到邻近位置,进而更好地消除相似重复记录。改进的 SNM 算法流程如图 1 所示,算法步骤及实现过程具体如下(以示例客户数据表为例):

1) 输入客户表记录,设定窗口大小 $S=3$,文本相似度阈值 $u=0.95$。客户数据表包括客户编号、姓名、性别、出生日期、手机号码、地址这 6 个属性。客户表记录中包含 4 条示例记录,如图 2 所示。

2) 数据预处理。客户表中的 Sex 和 Birthday 属性存在表示方式不一致的情况,对于这一类型的数据问题,通过数据预处理即可消除。

3) 选择关键属性。在判定两条或多条记录是否为相似重复记录时,并非所有属性都是关键属性。本文对客户表选择的关键属性是 Name,

Sex,Birthday,Address。

4）针对选择关键属性后的记录生成字符串记录,并存入字符串记录表中。

5）将字符串记录单词化处理,如图3所示。

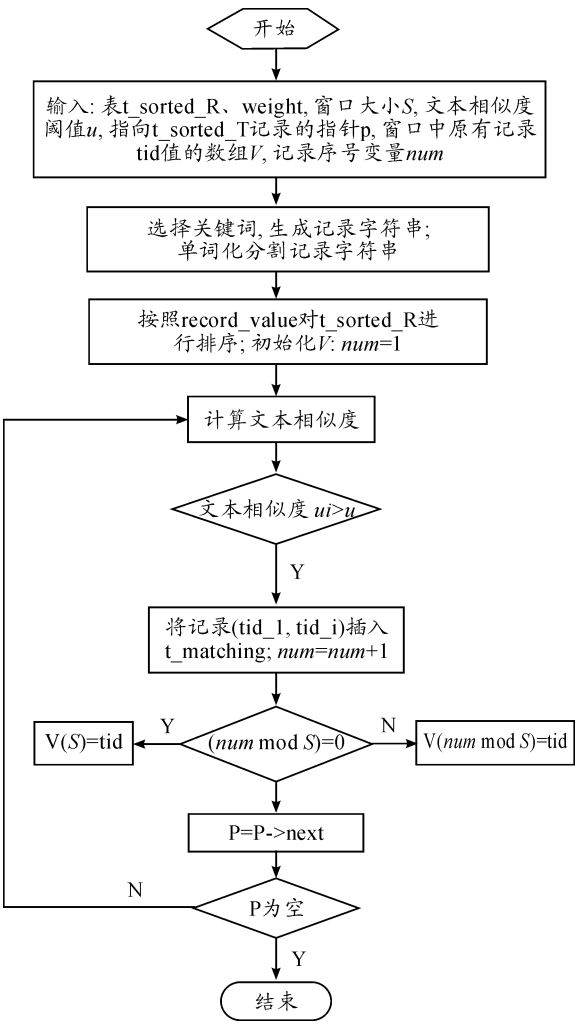

图1 改进的SNM算法流程

结果 消息

	ID	Name	Sex	Birthday	Phone	Address
1	1	Zhang San	1	1987-01-01	13800138000	Hubei Yichang Xiling University Road
2	2	Wang Mei	F	1989-10-10	18671745011	Hubei Yichang Xiling University Road
3	3	Mei Wang	W	1989-10-10	18671745011	University Road,Xiling,Yichang,Hubei
4	4	Zhang San	M	1987/01/01	13800138000	University Road,Xiling,Yichang,Hubei

图2 客户表记录

结果 消息

	ID	words
1	1	Zhang San 1 1987-01-01 Hubei Yichang Xiling University Road
2	2	Wang Mei 0 1989-10-10 Hubei Yichang Xiling University Road
3	3	Mei Wang 0 1989-10-10 University Road Xiling Yichang Hubei
4	4	Zhang San 1 1987-01-01 University Road Xiling Yichang Hubei

图3 单词化后的字符串记录

6）将单词化的子串进行排序。

7）为了最大限度地使相似重复记录处于邻近位置,将子串排序后的字符串记录表按照排序后的字符串进行排序。通过这一步的操作和处理,相似重复数据将处于邻近位置,即在算法的窗口之内。

8）根据设定的窗口大小以及文本相似度,对排序后的字符串记录计算文本相似度,消除相似重复记录。示例中消除相似重复记录后的结果见图5。

结果 消息

	ID	record_value
1	2	0 1989-10-10 Hubei Mei Road University Wang Xiling Yichang
2	3	0 1989-10-10 Hubei Mei Road University Wang Xiling Yichang
3	1	1 1987-01-01 Hubei Road San University Xiling Yichang Zhang
4	4	1 1987-01-01 Hubei Road San University Xiling Yichang Zhang

图4 排序后的字符串记录

结果 消息

	ID	Name	Sex	Birthday	Phone	Address
1	1	Zhang San	1	1987-01-01	13800138000	Hubei Yichang Xiling University Road
2	2	Wang Mei	0	1989-10-10	18671745011	Hubei Yichang Xiling University Road

图5 消除相似重复记录后的结果

3 实现方法与结果分析

3.1 实验环境和数据选择

考虑到真实数据涉及到商业机密,用来进行实验的数据获取比较困难,另外,实际数据中相似重复记录的总量不确定性也会对实验评价带来很大的困难,因此笔者利用来自Internet的测试数据生成器构造了用于本文测试的数据。构造的客户数据表主要包括ID,Name,Sex,Birthday,Phone,Address等6个属性。构造客户数据表之后,生成了10 000条客户记录,同时生成了8 000条相似重复记录,将其随机插入客户表中。

3.2 评价指标

笔者将算法消除相似重复记录的比例作为评价算法改进程度的指标。测试数据中相似重复记录的数量为已知量,因此通过算法消除的相似重复记录的比例很容易得到,且该百分比能在很大

程度上说明算法的性能和数据质量。

相似重复记录消除率表示算法可以消除的相似重复记录占数据表中所有相似重复记录的比例,定义为

$$\rho = \frac{N_V}{N} \times 100\% \quad (1)$$

其中:N_V 表示算法消除相似重复记录的数量;N 表示数据表中相似重复记录的总量。

3.3 结果分析

3.3.1 不同初始参数对消除结果的影响

根据算法流程可知,不同的初始参数对最终消除的相似重复记录的数量会产生影响。这里选择不同的窗口大小和文本相似度阈值进行实验和结果分析。

1)不同窗口大小 S 对消除结果的影响

为测试不同窗口大小对消除结果的影响,这里对文本相似度阈值取定值 $u=0.85$。测试结果如表 2 所示。

表 2 不同窗口大小消除结果

窗口大小 S	3	5	10	15	20	30	40
相似重复记录消除率/%	56	61	68	72	75	76	76.5

由图 6 的实验结果可知:在本文实验的数据中,相似重复记录消除率随窗口大小的增加而升高,当窗口增大到一定程度时,相似重复记录消除率上升缓慢并逐渐趋于平稳。可见,针对本文实验数据,最优窗口大小为 $S=20$。

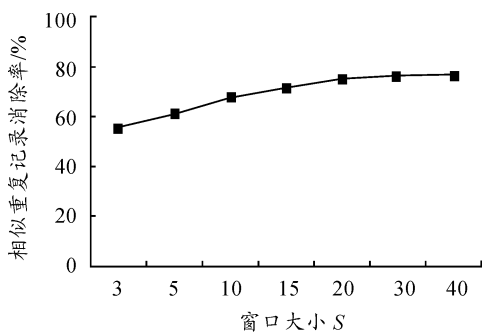

图 6 不同窗口大小消除结果折线

2)不同文本相似度阈值 u 对消除结果的影响

为了测试不同文本相似度阈值对消除结果的影响,这里对窗口大小取上述最优值 $S=20$,测试结果如表 3 所示。

表 3 不同文本相似度阈值消除结果

文本相似度阈值 u	0.75	0.8	0.85	0.9	0.95
相似重复记录消除率/%	86	81	75	70	58

由图 7 的实验结果可知:在本文实验的数据中,相似重复记录消除率随文本相似度阈值的增大而降低,当文本相似度阈值增大到一定程度时,相似重复记录消除率降低缓慢并逐渐趋于平稳,即文本相似度要求越严格,探测到的相似重复记录比例会越低。由上述实验结果可见,针对本文实验数据,可选择文本相似度阈值大小为 $u=0.85$。

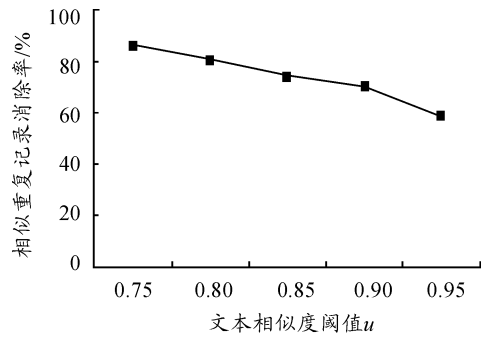

图 7 不同文本相似度阈值消除结果折线

3.3.2 改进 SNM 算法与传统 SNM 算法的消除效果比较

为了比较改进 SNM 算法和传统 SNM 算法的消除效果,采用本文中的测试数据,并设定文本相似度阈值 $u=0.85$ 进行不同窗口大小下的对比实验。消除效果对比见表 4。

表 4 改进 SNM 算法与传统 SNM 算法消除效果对比

窗口大小 S		3	5	10	15	20	30	40
相似重复记录消除率/%	改进的 SNM 算法	56	61	68	72	75	76	76.5
	传统的 SNM 算法	37	46	54	61	67	69	71.5

从图8显示的结果可知:相同窗口大小的情况下,改进SNM算法相比传统算法有较好的相似重复记录消除率,说明算法改进有一定的效果。

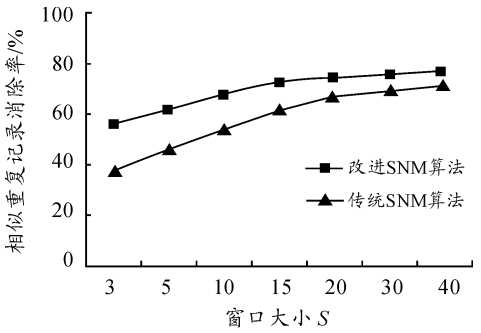

图8 改进SNM算法与传统算法对比消除结果折线

参考文献:

[1] KIMBALL R, REEVES L, ROSS M, et al. The Data Warehouse Lifecycle Toolkit:The Definitive Guide to Dimensional Modeling [M]. Indiana: Wiley Publishing Inc,2013.

[2] LOSHIN D. Data Quality ROI in the Absence of Profits [J]. Information & Management,2003(9):22.

[3] HUANG K,LEE T,Y W WANG,et al. Quality Information and Knowledge[M]. NJ:Prentice-Hall,1999.

[4] CLIKEMAN P M. Improving information quality [J]. Internal Auditor,1999(3):32-33.

[5] SINGH R, SINGH K. A descriptive classification of causes of data quality problems in data warehousing[J]. International Journal of Computer Science Issues,2010.

[6] 张建中,方正,熊拥军,等.对基于SNM数据清洗算法的优化[J].中南大学学报(自然科学版),2010(6):2240-2245.

[7] 陈爽,刁兴春,宋金玉,等.基于伸缩窗口和等级调整的SNM改进方法[J].计算机应用研究,2013(9):2736-2739.

[8] 叶焕倬,吴迪.相似重复记录清理方法研究综述[J].现代图书情报技术,2010(9):56-66.

[9] MAURICIO A HERNÁNDEZ,SALVATORE J S. Real-world Data is Dirty:Data Cleansing and The Merge/Purge Problem[J]. Data Mining and Knowledge Discovery,1998,2(1):9-37.

[10] HERNANDEZ M,STOLFO S. The Merge/Purge Problem for Large Databases[C]//Proceedings of the ACM SIGMOD International Conference on Management of Data. San Jose,California:[s. n.],1995:127-138.

(责任编辑 杨黎丽)