数据挖掘中的数据品质问题及其挖掘

Problem and Mining about Data Quality in the Data Mining

蒋 渝 王蔚韬 张建高 何光辉
(重庆大学计算机学院 重庆400044)

Abstract This paper introduces the conception of data quality and the issues that the attention has not enough been paid to the data quality in data mining (DM). Then, it analyze and emphasize that the data quality is crucial for many applications in DM with real examples. Finally an example of the iatric diagnoses application is given to show how to improve the data quality.

Keywords KDD,DM,Data quality,DQM

一、引言

数据库中的知识发现(Knowledge Discovery in Databases,KDD)有时又叫数据挖掘(Data Mining,DM),它的各项技术在各个领域得到了应用,并得到广泛的重视。

建立数据仓库是数据挖掘工作的第一步。数据仓库被定义为面向主题、集成的、随时间变化的、数据稳定的,被用来组织决策的数据集合。数据仓库作为一个很重要的策略来为一个组织的从各种异构的信息来源进行结合,并进行在线分析(OLAP)以及数据挖掘。

不幸,数据挖掘中的数据品质未得到人们的足够重视。实际应用就有资料表明,因为在每一个代理处数据品质的缺乏,超过二亿美元的美国联邦贷款被损失了。在制造业中浪费的花费占总销售额的25%。在服务行业达到了40%。数据品质问题就体现出其在信息技术中的重要性。

数据仓库的基础数据来源复杂、多样。对面向相同的主题,不同基础数据的语义可能不同,格式不同,层次不同。并且,当前的数据收集和储存过程还没有完全地可靠和可信。数据仓库的数据品质问题必将影响下一步的数据挖掘。由于基础数据的品质问题和其它一些因素,数据挖掘的结果数据十分庞大,而实际可使用的信息的量不大,并且其使用效率也不高。所以,在数据挖掘的整个过程都发展相关的数据品质技术将可以有效地解决这些相关的问题,同时减小这些相关的现象。

二、数据品质的概念

数据品质对在很多应用的 KDD 工具将是一个决定性的因素。提高数据品质是 KDD 应用中一个急需解决的问题。

根据品质管理的原理,数据品质定义为是与用户期望的一致程度,也有定义为对满足使用要求的适应程度,英文是"fitness for use"。它包括很多方面,例如:数据合法性、数据一致性、数据准确度和数据完整性等等。

数据仓库必须有高水平的数据品质,才能提供高水平的服务。数据预处理技术可以改进数据的质量,从而有助于提高其后的挖掘过程的精度和性能。由于高品质的决策必然依赖于高品质的数据,因此数据预处理是知识发现过程的重要步骤。检测数据异常、尽早地调整数据,并归约分析的数据,将在决策过程得到高回报。实际上,数据挖掘的整个过程对数据品质都有很多要求,而所有这些要求有时是不同的,甚至是相反的。这样问题就产生了。所以,数据品质的这些问题将带来数据品质技术在数据挖掘中的新的应用。

三、数据品质的挖掘

这里主要提出一个提高数据品质的途径。数据品质的挖掘(data quality mining),缩写为 DQM。DQM 的目的是应用数据挖掘的方法在一个海量数据库中来发现、量化、解释和校正数据品质的不足。DQM 从教学和商业角度来看打开了数据挖掘方法的全新的、有前途的应用领域,而不是单纯的数据分析。当前 KDD 处理的是大量各个方面的基础数据,因此处理数据品质是一个决定性的问题,在实践中成为了实践挖掘的一部分。它有几个方面可以发展:

1. 用 DM 方法来测量和解释数据品质的不足。

2. 用 DM 方法来校正数据品质的不足。

对校正而言,重新收集是经常使用的方法。不幸的是重新收集常是不可能的或代价太高。所以,用 DM 方法来对收集进行发现、排除和猜测等以提高数据的品质。

3. 扩展 KDD 的过程模型来体现 DQM 潜在的能力。

虽然当前的过程模型也意识到数据品质的问题,但是 KDD 的过程模型应该有一个明确的数据品质阶段。同时,数据品质不仅是 KDD 的过程模型初始的一个阶段,还应是在实际的数据挖掘和应用中数据品质也是一个贯穿其中的重视因素。

4. 为纯的 DQM 发展专用过程的模型。

数据品质方面的问题并不只是 KDD 才有的,在实际工程中一样存在应用领域十分广阔。这种数据品质专用过程的模型需要从纯数据分析的基础上对数据品质的量化和提高有一个全新的发展。

下面举一个例子:

本例是用数据挖掘方法中的关联规则来进行数据品质挖掘。

关联规则挖掘发现大量数据中项集之间有趣的关联或相关联系。而我们现在是用关联规则挖掘发现关联或相关联系

蒋 渝 硕士生,研究方向为数据库技术,数据挖掘技术。王蔚韬 副教授,主要研究方向为数据库技术,系统工程,专家系统。何光辉 硕士生,研究方向为数据库技术,数据挖掘技术。

对事务数据进行数据品质处理。

假设有一个事务数据库D由数据库进行数据挖掘我们可得以下关联规则见表1。

表1

关联规则	置信度
肝胆湿热⇒脉弦数	80%
肝胆湿热⇒小便短赤	60%
肝胆湿热⇒胁肋胀痛	60%
肝胆湿热⇒大便溏结	50%
⋮	⋮

现在我们要对再进入系统的基础数据进行数据品质评价。

定义一个概念为得分用S表示$S\in R_t^+$，每个事务的得分S由前面数据挖掘得到的关联规则来计算。设R为关联规则的集合。对一个关联规则r=X⇒Y，其中X为前提记为body(r)=X，Y结论记为head(r)=Y。

定义事务对规则的违反度，对一个事务T有四种情况：

1)body(r)⊂T∧head(r)⊂T；

2)body(r)⊂T∧head(r)⊄T；

3)body(r)⊄T∧head(r)⊂T；

4)body(r)⊄T∧head(r)⊄T；

对第3和第4种情况前提都不满足，规则也就没有意义不存在违反或不违反的情况。第1种情况是满足规则的。第2种情况就只满足前提是违反规则的情况。

这样我们定义事务违反规则的公式为：

$$V(T,r)=\begin{cases}1 & \text{如 body}(r)\subset T\ \wedge \text{head}(r)\\ & \quad\subset T,\text{即第2种情况；}\\ 0 & \text{其它三种情况；}\end{cases}$$

事务的得分是本事务违反规则集合R的和，记为：

$$S(T,R)=\sum_{r\in R}C(r)^t\times V(T,r)$$

其中：C(r)是规则r的置信度，t是一个指数参数。

当t=7时，事务号1的得分：$S(1,R)=0.8^7+0.6^7+0.6^7+0.5^7$，其它同理如表2。

从表2可以看出t的不同值，得分的分布不同。t值越小，得分分布越疏。同时，t的不同值可以反映不同的情况。如事务3不违反的是置信度比较大为0.9的脉弦数的项，违反其它三项置信度比较小的分别为0.6、0.6和0.5。而事务2则反之。从表2我们可以看出在不同的t值下它们的得分是不同的，并且变化趋势也不同。这样在实际应用中进行调整就能从得分上体现使用者想要知道和关心的方面。再通过设置得分S的阈值就可以衡量事务数据的品质。

表2

事务号	事务内容	t=7时，得分	t=5时，得分
1	肝胆湿热⇒脉沉数　小便不利　腰痛　大便燥结	0.33	0.7
2	肝胆湿热⇒脉沉数　小便短赤　胁肋胀痛　大便溏结	0.21	0.33
3	肝胆湿热⇒脉弦数　小便不利　腰痛　大便燥结	0.13	0.37
4	肝胆湿热⇒脉弦数　小便短赤　胁肋胀痛　大便溏结	0	0
5	膀胱湿热⇒脉数　小便短赤　腰痛　舌红苔黄	0	0

我们举的例子只是清楚地表示应用数据挖掘技术进行数据品质的挖掘的过程。它的应用领域将十分广泛。

总结　现在，由于基础数据的品质问题和其它一些因素，数据挖掘的结果数据十分庞大，而实际可使用的信息的量不大，并且其使用效率也不高。在数据挖掘的整个过程都发展相关的数据品质技术将可以有效地解决这些相关的问题，数据挖掘在数据品质在数据挖掘的应用工具将是一个决定性因素。数据品质及其技术将使数据挖掘得到进一步的发展。

参考文献

1 Han Jiawei, Kambr M. 数据挖掘——概念与技术(影印版). 高等教育出版社, 2001

2 Srikant R, Agrawal R. Mining generalized association rules. In: Proc. of the 21st Conf. on Very Large Databases (VLDB'95), Switzerland, 1995

3 Jeusfeld M, Quix C, Jarke M. Design and Analysis of Quality Information for Data Warehouses. In: Proc. of Conf. on Information Quality Cambridge, Mass., 1997. 98～112

4 Theodoratos D, Bouzeghoub M. Data Currency Quality Factors in Data Warehouse Design(DMDW'99), 1999

5 张华珠主编. 中医学. 中国科技出版社, 1992

(上接第137页)

3 张学工译. 统计学习理论. 北京：清华大学出版社, 2000. 9

4 Kecman V. Learning and soft computing, The MIT Press, Cambridge, MA, 2001

5 de Freitas N, Milo M, Clarkson P. Sequential support vector machines, Neural Networks for Signal Processing IX. In: Proc. of the 1999 IEEE Signal Processing Society Workshop, 1999. 31～40

6 Barabino N, Pallavicini M. Support vector machines vs multi-layer perceptrons in particle identification. In: Proc. of the European Symposium on Artificial Neural Networks'99, 1999. 257～262

7 Osuna E, Freund R. Training Support Vector Machine : an Application to Face Dection. In: Proc. of CVPR'97, Puerto Rico, 1997

8 Drucker H, Wu D, Vapnik V. Support vector machine for spam categorization. IEEE Trans. on Neural Networks, 1999, 10: 1048～1054

9 Suykens J A K. Nonlinear modeling and support vector machines. IEEE instrumentation and measurement technology conference, Budapest, Hungary, 2001. 21～23

10 Mukherjee S, Osuna E, Girosi F. Nonlinear prediction of chaotic time seris using support vector machines. In: Proceedings of IEEE NNSP'97, 1997. 24～26

11 Mangasarian O L, Street W N. Breast cancer diagnosis and prognosis via linear programming. Operations Research, 1995, 43(4): 570～577

12 Burges J C. A Tutorial on Support Vector Machines for Pattern Recognition. Bell Laboratories, Lucent Technologies. 1997