文章编号:1671-4598(2017)07-0282-03 DOI:10.16526/j.cnki.11—4762/tp.2017.07.070 中图分类号:TP311 文献标识码:A

异构式分布下的 Internet 数据挖掘方法优化研究

林明方
(广东工程职业技术学院 信息工程学院,广州 510520)

摘要:为了提高异构式分布下的 internet 数据的利用率,增加 internet 的多样化使用功能和数据传输率,减少 internet 运行的时间,需要对异构式分布下的 internet 数据进行挖掘;当前的数据挖掘方法多是先采用 SOM 系统的可视化功能对异构式分布下的 internet 数据进行聚类,然后根据聚类结果的计算完成对异构式分布下的 internet 数据挖掘;但该方法存在操作过程复杂,internet 数据经常性丢失的问题;为此,提出了一种基于本体论的异构式分布下的 internet 数据挖掘优化方法;该方法首先对异构式分布下的 internet 数据进行预处理选取出数据特征,并利用特征选择决策系统对挖掘数据进行特征选择,在此基础上利用信息熵实现异构式分布下的 internet 数据的过滤,过滤过程中通过信息熵数据过滤的理论值减小的变动,得到最佳数据过滤值,最后以预处理中获得的各项数据信息为基础,采用决策树生成算法中的信息增益值的迭代计算结果对异构式分布下的 internet 数据进行高精度挖掘;仿真实验结果证明,所提方法提高了异构式分布下的 internet 数据操作的灵活度,增加了 internet 数据的可循环利用率,使异构式分布下的 internet 操作更加简洁化、高效率化,为该领域的研究发展提供了强有力的依据。

关键词:异构式分布;internet;数据挖掘方法;优化研究

Under the Heterogeneous Distribution of Internet Data Mining Method Optimization Research

Lin Mingfang
(School of Information Engineering, GuangDong Engineering Polytechnic, Guangzhou 510520, China)

Abstract: In order to improve the utilization rate of the Internet data under the heterogeneous type distribution, increase the diversification of the Internet use function and data transfer rate, reduce the operation of the Internet time, needs to be under the heterogeneous distribution of Internet data mining. The current data mining method is to adopt the visual function of SOM system under the heterogeneous distribution of Internet data clustering, then according to the clustering results of calculation with heterogeneous distribution of Internet data mining. But this method is a complex process of operation, regular Internet data missing problem. To this end, this paper proposes a heterogeneous distribution based on ontology of Internet data mining method. The method of first selection under the heterogeneous distribution of Internet data preprocessing the data characteristics, and use the feature selection decision-making system for mining data feature selection, on the basis of using information entropy under the heterogeneous distribution of Internet data filter, filter by information entropy in the process of data, the change of the theoretical value of reducing data filtering to get the best value, and finally on the basis of the pretreatment of the data obtained from the information, using the decision tree generation algorithm of iterative calculation results of information gain value to high precision under the heterogeneous distribution of Internet data mining. Simulation experimental results show that the proposed method improves the heterogeneous distribution of Internet data operation flexibility, increased the recycled utilization rate of Internet data, makes the heterogeneous distribution of the Internet more concise, efficient operation, the research in the field development provides a strong basis.

Keywords: heterogeneous distribution of; internet; data mining methods; optimization research

0 引言

随着计算机科学与互联网技术的不断发展以及普及,异构式分布下的 internet 数据分别在办公自动化、电子数据交换、远程交换、远程教育、电子公告板系统 BBS、电子银行、证券及期货交易、广播分组交换、信息高速公路、企业网络、智能大厦和结构化综合布线系统等社交性平台以及系统中都有着广泛的应用。因此异构式分布下的 internet 数据的发展受到了人们的广泛关注和高度重视[1-2]。异构式分布下的 internet 不仅可以支持不同协议的不同应用,将各具优势的产品或系统进行结合利用,而且还可以满足网络业务的多样化需求[3],提高 internet 中各项多功能平台和系统的利用率。由于异构式分布下的 internet 具有不确定性、多样性、灵活性等特点,所以需要对异构式分布下的 internet 数据进行挖掘。大多数的异构式分布下的 internet 数据挖掘方法在进行数据挖掘时无法对其进行迅速、有效、高精度的挖掘,导致异构式分布下的 internet 在运行或操作时经常出现丢包率过大,数据操作过程复杂,计算有误差等问题[4-5]。在这种情况下,如何减少异构式分布下的 internet 数据挖掘的丢包率,提高 internet 数据挖掘的精度成为了亟待解决的问题。而基于本体论的异构式分布下的 internet 数据挖掘优化方法可以对其进行灵活、方便、可靠、高精度的数据挖掘。是解决上述问题的可行途径[6],受到了该领域专家的广泛关注,并且已经成为了异构式分布下的 internet 数据挖掘领域研究学者的研究课题,同时也获得了很多优秀的

收稿日期:2017-03-30; **修回日期**:2017-04-24。
基金项目:广东省科学技术厅项目(2014A010103008);广东省高等职业教育教学改革项目(20140116);广东省高等职业教育品牌专业建设项目—广东工程职业技术学院软件技术专业(2016gzpp031)。
作者简介:林明方(1981-),男,广州梅州人,硕士,高级工程师,主要从事嵌入式技术与应用、软件工程技术方向的研究。

方法[7]。

文献［8］提出了一种基于最近邻聚类的异构式分布下的 internet 数据挖掘方法。该方法采用改进的最近邻聚类算法对异构式分布下的 internet 数据进行训练，使异构式分布下的 internet 数据在满足挖掘精度要求前提下，减少隐藏的网络数据节点数，简化 internet 数据结构，加快异构式分布下的 internet 数据的学习速度，达到改善异构式分布下的 internet 数据学习效率的目的，在此基础上用最近邻聚类方法对异构式分布下的 internet 数据进行挖掘。该方法可以安全、稳定地对异构式分布下的 internet 进行数据挖掘，但是存在消耗时间过长的问题。文献［9］提出了一种基于垂直搜索的异构式分布下的 internet 数据挖掘方法。该方法首先利用垂直搜索的方式从异构式网络搜集数据，对得到的网络数据信息进行数据分类操作处理，将处理过后的结构化网络数据保存至异构式分布下的 internet 数据库中，然后对 internet 数据库中的网络数据进行分析，发现其中的规则和特征，最后根据异构式分布下的 internet 数据库中的网络数据的规则和特征对异构式分布下的 internet 数据进行挖掘。该方法可以快速的对异构式分布下的 internet 数据进行挖掘，但是存在数据挖掘精度较低的问题。文献［10］提出了一种基于粗集理论和神经网络结合的异构式分布下的 internet 数据挖掘方法。该方法首先利用粗集理论对原始网络数据进行属性约简化，然后使用神经网络对异构式分布下的网络数据实现学习与预测操作，完成网络数据属性的不一致约简化，最后利用粗集理论对神经网络知识的获取完成对异构式分布下的 internet 数据挖掘。该方法对异构式分布下的 internet 数据挖掘的容错能力较好，但挖掘完成速度较慢，而且过程繁琐。

针对上述产生的问题，提出一种基于本体论的异构式分布下的 internet 数据挖掘优化方法。该方法首先对欲挖掘的异构式分布下的 internet 数据进行预处理操作，使挖掘精度更高，挖掘速度更快，然后利用决策树生成的算法对异构式分布下的 internet 数据进行挖掘。仿真实验证明，所提方法可以高效精确地对异构式分布下的 internet 数据进行挖掘，且具有良好的可实现性。

1　异构式分布下的 internet 数据挖掘优化方法

1.1　异构式分布下的 internet 数据预处理

利用本体论对异构式分布下的 internet 数据进行挖掘，首先进行异构式分布下的 internet 数据预处理，在数据预处理中要确定异构式分布下的 internet 原始数据集的数据目标属性集合以及数据条件属性集合，其次将属性集合的取值范围区分为若干个小区间，一个 internet 数据离散符号对应一个数据属性集合小区间，由此得到一个异构式分布下的 internet 数据的特征选择决策系统，对特征选择决策系统中的相同数据记录进行合并，记作 (R,CRD)，建立异构式分布下的 internet 数据的特征选择决策系统，是对传统数据挖掘方法中没有进行数据特征选择而直接进行 internet 数据挖掘的方法进行了优化。

在数据特征选择之前要将数据的特征提取出来，在本文中利用最大间隔算法对异构式分布下的 internet 数据进行特征提取。假设 internet 数据服从某一特征分布 P，那么根据最大间隔算法计算异构式分布下的 internet 数据特征提取相似值 w 为：

$$w = \arg_w w_q^o \tag{1}$$

其中，O 代表异构式分布下的 internet 数据特征提取的预定义阈值，q 代表 internet 数据特征提取维数。

依据异构式分布下的 internet 数据特征提取相似值 w，得知 internet 数据特征的提取过程为：

输入：F，G，h

输出：$w_1,w_2,\cdots w_\delta$

其中，F 代表异构式分布下的 internet 数据特征提取数据集，G 代表异构式分布下的 internet 数据特征提取中一特征参数，h 代表异构式分布下的 internet 数据待提取特征维数，δ 代表 internet 数据特征提取中特征属性映射值。由此完成了对异构式分布下的 internet 数据特征提取。

将本体论应用于异构式分布下的 internet 数据的特征选择，就是要从异构式分布下的 internet 原始数据中提取出最能反映出异构式分布下的 internet 数据挖掘本质的特征，以下是异构式分布下的 internet 数据特征选择方法具体过程：

假设，输入：异构式分布下的 internet 数据条件属性集 A、数据决策属性集 B、数据决策系统 (R,CRD)。

输出：异构式分布下的 internet 数据生成分辨矩阵 H、数据约简集 $X(A,B)$。

1）若 n 代表 internet 数据决策系统中属性个数，则：

$$X(A,B) = \varphi \tag{2}$$

其中，X 代表异构式分布下的 internet 数据约简值，φ 代表异构式分布下的 internet 数据约简集。

2）假设一个 $n \times n$ 的 internet 数据属性集合矩阵 N；

3）根据本体论中分辨矩阵生成数据分辨矩阵，记录在（2）中的 internet 数据属性集合矩阵 N，则有：

$$for(j = i+2; i < n; j++) \tag{3}$$

$$if B(x_i) = B(x_i), N_{ij} \leftarrow \varphi \tag{4}$$

其中，i 代表数据条件属性个数，j 代表数据决策属性个数，N 代表 internet 数据属性集合矩阵。

4）将 2）中的每个 internet 数据属性子式集合添加到 $X_{LOP}(A,B)$ 中；

5）输出 internet 数据属性集合矩阵 N，约简集 $X_{LOP}(A,B)$。

综上所述，利用上述建立的异构式分布下的 internet 数据的特征选择决策系统，完成对异构式分布下的 internet 数据的特征提取和数据特征选择过程。

在异构式分布下的 internet 数据挖掘中为了提高数据挖掘质量，必须对 internet 数据进行过滤，基于本体论的异构式分布下的 internet 数据挖掘优化方法利用的是信息熵对 internet 数据进行过滤，利用信息熵的过滤理论值对 internet 数据过滤条件值 IT 进行输入：

$$IT = (U,A_t,V_x,I_x) \tag{5}$$

则输出为：

$$IT = (U,A_t,V'_x,I'_x) \tag{6}$$

其中，U 代表 internet 每一属性对应的数据属性值集，V 代表异构式分布下的 internet 数据挖掘样本的期望信息值，V' 代表 internet 过滤数据挖掘样本的期望信息值，I 代表 internet 数据过滤过程中的信息函数值，I' 代表 internet 数据过滤过程中的信息熵理论值，t 代表 internet 数据属性均值，x 代表 internet 数据属性值。

使每个 internet 数据属性值 $x \in A_t$，对异构分布下的 in-

ternet 数据属性值进行排序，对于非 internet 数据属性值，假设其属性值为有序关系可以将其转化成数据数值序，此步骤在异构分布下的 internet 数据值排序过程进行了优化，使非 internet 数据属性也可以排序。完成排序后，要对每个 internet 数据属性值 $x \in A_t$ 依据信息熵的过滤信息函数执行以下步骤。

$$fori = 1toK - 1 \tag{7}$$

其中，K 代表利用信息熵对异构式分布下的 internet 数据过滤的最大指定过滤值。则利用信息熵对异构分布下的 internet 数据过滤可定义为：

$$H(C/X;V_1,V_2\cdots,V_t) = \sum_{j=1}^{i+1} p(U_j) \sum_{d\in V_D} p(d/U_j)\log(\mathbb{I}d/U_j) \tag{8}$$

其中，H 代表信息熵对异构分布下的 internet 数据过滤的定义值，p 代表 internet 数据属性概率分布值。

当异构分布下的 internet 数据不断增加时，信息熵数据过滤的理论值变动很小，则输出最佳过滤值，完成对异构分布下的 internet 数据的过滤。

综上所述，异构式分布下的 internet 数据预处理主要由数据特征选择和数据过滤组成，数据的预处理提高了数据挖掘的质量，减少了数据挖掘的时间。

1.2 异构式分布下的 internet 数据挖掘

在完成基于成本体论的异构式分布下的 internet 数据挖掘的数据预处理后，采用决策树算法对异构式分布下的 internet 数据进行挖掘。具体方法如下：

决策树算法依据自上而下的方式构造异构式分布下的 internet 数据挖掘决策树，分为 internet 数据决策树生成与 internet 数据决策树剪枝，本文对 internet 数据决策树剪枝不做研究。internet 数据决策树生成算法利用信息增益来选择异构式分布下的 internet 数据中最好的挖掘属性，信息增益具体计算方式如下：

假设有 m 个信息，挖掘的数据属性概率分布为：

$$p = (p_1, p_2, \cdots p_m) \tag{9}$$

则该异构式分布下的 internet 数据挖掘样本 S 的期望信息值：

$$V(S) = V(p) = \sum_{1}^{m} p_i \log_2 p_i \tag{10}$$

其中，S 代表异构式分布下的 internet 数据挖掘中的样本总数，m 代表信息增益中的信息个数。

给定的异构式分布下的 internet 数据挖掘样本 $s_i \in S$，internet 数据挖掘样本总数为 S_i，根据异构式分布下的 internet 数据挖掘类别属性值将 s_i 划分为 z 个数据类别属性子集，每个数据挖掘类别子集中包含的异构式分布下的 internet 数据挖掘样本数为 s_{ij}，则 internet 数据挖掘属性概率分布如式（10）所示：

$$p = (S_{i1}/S_i, S_{i2}/S_i, \cdots, S_{iz}/S_i) \tag{11}$$

根据公式（10）得知 internet 数据挖掘样本 s_i 的期望信息值为 $I(s_i) = I(p)$。异构式分布下的 internet 数据挖掘样本集 S 的熵为：

$$E(S) = \sum_{1}^{z} \frac{(S_{i1} + S_{i2} + \cdots + S_{im}) I(s_i)}{S} \tag{12}$$

异构式分布下的 internet 数据挖掘样本 S 的信息增益值为：

$$Y(S) = I(S) - E(S) \tag{13}$$

其中，Y 代表异构式分布下的 internet 数据挖掘的信息增益值，E 代表异构式分布下的 internet 数据挖掘样本集的熵。

对上述过程进行迭代计算，直到满足下列条件之一结束迭代计算：1）给定的 internet 数据结点的所有样本属于同一分类；2）没有多余的 internet 数据属性可以进一步划分数据属性样本；3）异构式分布下的 internet 数据分支属性样本为空。至此完成对异构式分布下的 internet 数据挖掘。

2 仿真实验结果与分析

为了证明基于本体论的异构式分布下的 internet 数据挖掘优化方法的有效性，需要进行一次仿真实验。在 Visual C 的环境下搭建异构式分布下的 internet 数据挖掘实验仿真平台。实验数据取自于 SPSS Clementine11.1 数据挖掘系统，在该实验中，利用本体论对 SPSS Clementine11.1 数据系统中的异构式分布下的 internet 数据进行高质量挖掘。表 1 是对基于文本论的异构式分布下的 internet 数据挖掘优化方法中特征选择数据量（万个）与其选择效率（%）之间关系的描述。

表 1 internet 数据挖掘中数据特征选择与选择效率关系

特征选择数据量（万个）	数据选择效率（%）
1000	92.6
2000	93.3
3000	95.2
4000	92.4
5000	94.5

通过表 1 中的各项数据明显看出基于文本论的异构式分布下的 internet 数据挖掘优化方法是安全可靠的。在表中数据特征选择效率虽然随着特征选择数据量的增加而不断波动，但选择效率基本在 90%以上，更加说明了基于文本论的异构式分布下的 internet 数据挖掘优化方法的整体有效性。表 2 是对基于文本论的异构式分布下的 internet 数据挖掘优化方法中过滤数据量（万个）与过滤所用时间（s）的关系描述。

表 2 internet 数据挖掘中数据过滤与所用时间关系

过滤数据量（万个）	过滤所用时间（s）
100	3.11
200	3.63
300	4.10
400	4.56
500	4.89

表 2 中对基于文本论的异构式分布下的 internet 数据挖掘优化方法中过滤数据量与其所用时间的关系描述中过滤数据所用时间随着过滤数据量的增加波动相对较小，说明本文所提的数据挖掘优化方法时间消耗较少，进一步证明了基于文本论的异构式分布下的 internet 数据挖掘优化方法的可实现性。图 1 是对文献［9］所提挖掘方法与本文方法挖掘效率（%）的对比。

图 1 中文献［9］所提挖掘方法挖掘效率随着挖掘数据量的增加处于波动较大状态。本文所提数据挖掘优化方法挖掘效率处于平稳波动状态，且挖掘效率较高，明显优于文献［9］所提挖掘方法，这主要是因为利用本文所提方法进行数据挖掘

（下转第 289 页）

参考文献：

[1] 史红卫，史　慧，孙　洁，等．服务于智能制造的智能检测技术探索与应用［J］．计算机测量与控制，2017，(01)：1－48.

[2] 孙　娜，陶文华，李青苗，等．基于小波神经网络的船舶冷却水系统的传感器故障诊断［J］．测控技术，2007，3（2）：179－180．附.

[3] 马　超，张英堂，李志宁．基于 PSO－KELM 的发动机特征参数预测［J］．控制工程，2014，(S1)：28－32.

[4] 张　曦，陈世和，朱亚清，等．基于 KPCR 的发电机组参数预测与估计［J］．电力自动化设备，2010，(10)：54－57.

[5] 史书真．股价时间序列的分析与预测研究［D］．大连：大连理工大学，2013.

[6] Chalermarrewong T，See S，Achalakul T，editors．Parameter Prediction in Fault Management Framework［C］．Proceedings of The International Symposium on Grids and Clouds（ICGC 2012）－ 26 February － 2 March；2012；Taiwan.

[7] 韩路跃，杜行检．基于 MATLAB 的时间序列建模与预测［J］．计算机仿真，2005（4）：105－107.

[8] Nadja Saleck and Lueder von Bremen．Wind power forecast error smoothing within a wind farm［Z］．The Science of Making Torque from Wind，Conference Series 75（2007）012051.

[9] 张俊潇，邓长虹，陈允平．最小二乘估计法优化电力系统网络等值参数［J］．电力科学与工程，2004（2）：34－37.

[10] 许正福．轮机动力装置系统实验室冷却水系统设计研究［D］．大连：大连海事大学，2009.

[11] Tayman J，Swanson D A．On the validity of MAPE as a measure of population forecast accuracy［J］．PopulationResearch and Policy Review，1999，18（4）：299－322.

[12] 徐野男．基于时间序列的船舶冷却水系统状态参数预测分析［D］．大连：大连海事大学，2015.

（上接第 284 页）

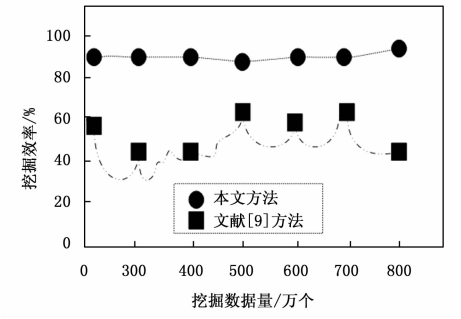

图 1　不同方法下挖掘效率对比

时对异构式分布下的 internet 数据挖掘时，利用最大间隔算法对 internet 数据进行特征提取、依据 internet 数据的特征选择决策系统对 internet 数据进行特征选择，以及采用信息熵对 internet 数据进行过滤的异构式分布下的 internet 数据挖掘预处理工作。为异构式分布下的 internet 数据挖掘打下了坚实基础，有利于对异构式分布下的 internet 数据进行高效率挖掘。图 2 是文献［8］所提挖掘方法与本文所提方法误差率（%）的对比。

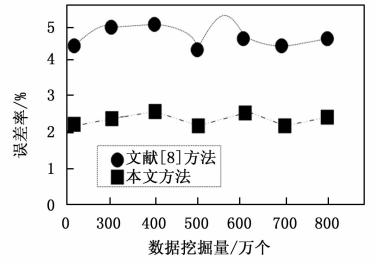

图 2　数据挖掘误差率

图 2 中本文所提基于文本论的异构式分布下的 internet 数据挖掘优化方法挖掘的误差率明显低于文献［8］所提挖掘方法，本文所提方法数据挖掘误差率在额定的数据挖掘数量持续增加中波动状态相对稳定，且一直在 5% 以下。主要是因为 internet 数据决策树的生成在数据挖掘过程中起着不可或缺的辅助作用，提高了异构式分布下的 internet 数据挖掘的准确率，有效增加了本文所提方法的可行性和优化性。

仿真实验证明，本文所提基于文本论的异构式分布下的 internet 数据挖掘优化方法可以精确地对异构式分布下的 internet 数据进行挖掘，保障了 internet 数据挖掘的整体有效性，提高了数据挖掘的速度，为该领域的研究发展提供了可靠依据。

3　结束语

采用当前方法对异构式分布下的 internet 进行数据挖掘时，无法高精度、高效率地实现异构式分布下的 internet 数据挖掘，存在挖掘误差率高、速度慢、不安全的问题。提出一种基于文本论的异构式分布下的 internet 数据挖掘优化方法。通过仿真实验证明，所提方法可以精准地对异构式分布下的 internet 数据进行挖掘，具有良好的应用价值，是切实可行的。

参考文献：

[1] 吕　佳，陈东生．基于聚类算法的服装感性数据挖掘方法［J］．纺织学报，2014，35（5）：108－112.

[2] 王　磊，张永坚，贾继鹏，等．基于 Hadoop 的公共建筑能耗数据挖掘方法［J］．计算机系统应用，2016，25（3）：34－42.

[3] 刘青风，李红兰．基于数据挖掘方法的风力涡轮机状态监测技术研究［J］．计算机测量与控制，2014，22（5）：1336－1339.

[4] 柳萌萌，赵书良，韩玉辉，等．多尺度数据挖掘方法［J］．软件学报，2016，27（12）：3030－3050.

[5] 丁骋骋，邱　瑾．性别与信用：非法集资主角的微观个体特征——基于网络数据挖掘的分析［J］．财贸经济，2016，37（3）：78－94.

[6] 杨丹丹．搜索引擎及网络数据挖掘相关技术研究［J］．数字化用户，2014，20（11）：126－126.

[7] 方永美，熊俊涛，杨振刚，等．基于贝叶斯网络数据挖掘的蔬菜质量安全分析［J］．湖北农业科学，2016，55（23）：6253－6257.

[8] 肖志军．一种面向社会网络的热点话题数据挖掘算法［J］．计算机应用与软件，2014，31（6）：24－28.

[9] 许学添，邹同浩．网络数据库中隐蔽数据快速挖掘方法研究［J］．电子设计工程，2016，24（24）：15－18.

[10] 余国清，周兰蓉．一种公共网络攻击数据挖掘智能算法研究［J］．计算机测量与控制，2016，24（10）：190－193.