第 18 卷　第 1 期　2018 年 1 月
1671—1815(2018)001-0298-06

科　学　技　术　与　工　程
Science Technology and Engineering

Vol. 18　No. 1　Jan. 2018
ⓒ　2018　Sci. Tech. Engrg.

引用格式:杨　媛, 马　旭, 陈　琛. 一种多层次分布式网络数据挖掘方法的改进[J]. 科学技术与工程, 2018, 18(1): 298—303
Yang Yuan, Ma Xu, Chen Chen. An improved method of multi level distributed network data mining[J]. Science Technology and Engineering, 2018, 18(1): 298—303

一种多层次分布式网络数据挖掘方法的改进

杨　媛　马　旭　陈　琛
(宁夏师范学院数学与计算机科学学院,固原 756000)

摘　要　针对传统数据挖掘方法时间开销大的问题,提出一种新的多层次分布式网络数据挖掘改进方法,给出多层次分布式网络结构。介绍了常用的随机扰动数据挖掘方法,通过概率歪曲技术完成对多层次分布式网络中原始数据集的扰乱处理,对项集的实际支持度进行重构,经概率转换获取数据挖掘结果。随机扰动方法具有时间效率低的弊端,在时间复杂度要求较低的情况下,通过 XMASK 方法对随机扰动方法进行改进;在时间复杂度要求较高的情况下,给出相应的改进过程。对提出的多层次分布式网络数据挖掘方法进行实验测试,结果表明,该方法准确性高、挖掘时间短、效果优。

关键词　多层次　　分布式网络　　数据挖掘　　改进
中图法分类号　TP393;　　**文献标志码**　A

当前,计算机技术发展迅猛,大规模网络服务被广泛应用[1]。最初通常利用集中式结构为用户提供网络服务,然而服务器性能无法达到日益增加的大规模用户的需求,多层次分布式网络应运而生,其拓展性高,能够为大规模用户提供服务,被广泛应用[2,3]。在大规模数据中挖掘出有效的、具有潜在价值的、用户需要的数据无疑非常困难且有意义,所以文章针对多层次分布式网络数据的挖掘进行深入研究。

为了挖掘出有效信息,提出了一种多层次分布式网络数据挖掘改进方法,经实验验证,所提方法有很高的性能。

1　多层次分布式网络结构

多层次分布式网络结构将服务器群划分为几个组,将各组划分成三层,依次是中心服务器层,数据服务器层和用户层[4],详细结构如图 1 所示。

在多层次分布式网络中,中心服务器层仅含有一个中心服务器,而数据服务器层主要由策略服务器与数据服务器构成[5,6]。上述多层次分布式网络结构利用数据服务器的协同运行,共同为用户提供相应服务,有效避免了服务器负载过多、负载不均的问题。

2017 年 5 月 29 日收到　2016 年度宁夏自然科学基金(NZ16258)、2015 年度宁夏师范学院科研项目(NXSFZD1604)、2016 年度宁夏师范学院科研项目(NXSFYB1779)、宁夏高等学校科研项目(NGY2015124)、宁夏自然科学基金(NZ16260)和宁夏师范学院重点科研项目(NXSFZDT170)资助
第一作者简介:杨　媛(1981—),女,汉族,宁夏固原人,硕士,讲师。研究方向:领域为数据库、网络安全。E-mail:yangyuan2562@126.com。

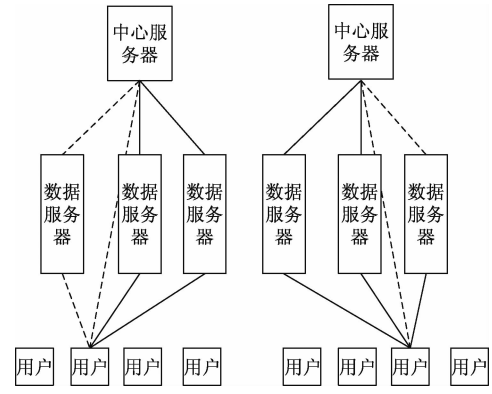

图 1　多层次分布式网络结构图

Fig. 1　Multi-level distributed network structure

将多层次分布式网络看作一个有向图 $G=(V,E)$,其中,V 为多层次分布式网络中所有文档节点集,有向边集 E 和节点间的超链接相应[7]。进一步对节点集进行分割,$V=(v_j,v)$ 中所有非叶子节点 v_j 代表网络文档,叶子节点 V 代表不同格式的文本文件。叶子节点 V 中所有节点均和一个可用的多层次分布式网站站点相应。

与传统集中式网络相比,多层次分布式网络中的信息存在易混淆、动态性的特点,因此,很难直接挖掘出其中的有效信息,需对其进行处理[8,9]。图 2 描述的是常用的数据挖掘过程。

2　网络数据挖掘方法及其改进

2.1　随机扰动挖掘方法

随机扰动方法是一种常用的数据挖掘方法,最早是由 Rizvi 提出的,其假设多层次分布式网络中的整个

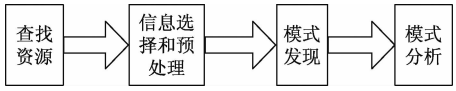

图 2　多层次分布式网络数据挖掘示意图

Fig. 2　Sketch map of multilevel distributed network data mining

数据集是超市购物篮[10],将待挖掘数据看作由 0 与 1 构成的二维稀疏布尔矩阵,其中 1 代表购买某商品,0 代表未购买。为了使挖掘出的数据集更加安全,随机扰动方法通过概率歪曲技术完成对多层次分布式网络中原始数据集的扰乱处理。将一个 0-1 数据库元组看作一个随机向量 $\boldsymbol{X} = \{X_i\}$, $X_i = 0$ 或 $X_i = 1$ 。

对 X_i 进行歪曲处理,则有:

$$Y_i = X_i X \text{ or } \overline{r_i} \tag{1}$$

式(1)中, $\overline{r_i}$ 为 r_i 的补; r_i 为符合贝努里分布的随机变量,分布律 $p(i_i = 1) = p$, $p(i_i = 0) = 1 - p$ 。通过异或计算的特性可看出,随机向量 $\boldsymbol{X}$ 经歪曲处理后,第 i 个分量 X_i 保持原值的概率是 p [11]。

通过随机扰动方法挖掘的数据集为多层次分布式网络数据集经概率转换产生的,因此需对项集的实际支持度进行重构[12]。假设与实际数据集相应的矩阵为 $\boldsymbol{T}$, $\boldsymbol{T}$ 通过歪曲转换后获取的矩阵为 $\boldsymbol{D}$,歪曲概率是 p 。$\boldsymbol{T}$ 中的第 i 列中 1 的数量为 c_1^T ,0 的数量为 c_0^T , D 中第 i 列中 1 的数量为 c_1^D ,0 的数量为 c_0^D 。通过前面分析的概率歪曲过程可知:

$$\boldsymbol{c}_1^T p + c_0^T (1 - p) = c_1^D \tag{2}$$

$$c_0^D p + c_1^D (1 - p) = c_0^D \tag{3}$$

因此有:

$$\boldsymbol{C}^T = \boldsymbol{M}^{-1} \boldsymbol{C}^D \tag{4}$$

式(4)中, $\boldsymbol{M} = \begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix}$, $\boldsymbol{C}^D = \begin{pmatrix} c_1^D \\ c_0^D \end{pmatrix}$, $\boldsymbol{C}^T = \begin{pmatrix} \boldsymbol{c}_1^T \\ \boldsymbol{c}_0^T \end{pmatrix}$ 。

求解式(2)~式(4)即可依据合成矩阵 $\boldsymbol{D}$ 获取实际矩阵 1-项集的支持度 c_1^T [13]。n-项集的真实度评价方式与单项集相似,则有:

$$\begin{cases} \boldsymbol{C}^T = \boldsymbol{M}^{-1} \boldsymbol{C}^D \\ \boldsymbol{C}^D = \begin{pmatrix} c_{2^n-1}^D \\ \vdots \\ c_1^D \\ c_0^D \end{pmatrix} \\ \boldsymbol{C}^D = \begin{pmatrix} c_{2^n-1}^T \\ \vdots \\ c_1^T \\ c_0^T \end{pmatrix} \end{cases} \tag{5}$$

式(5)中, c_k^D 为 n-项集 k 的数量;n-项集 k 代表 n 位二进制数的形式,与其相应的十进制值为 k 。c_k^T 的定义和 c_k^D 一致, $\boldsymbol{M}$ 为 2^n 阶矩阵[14];矩阵元素 $M_{i,j}$ 为把实际 n-项集 j 概率歪曲成 n-项集 i 的概率。对式(5)进行求解可获取 n-项集在实际矩阵中的支持度。

随机扰动方法时间效率复杂,需对其进行改进。

2.2　随机扰动挖掘方法的改进

常用的改进方法为 XMASK 方法。XMASK 方法通过低阶和高阶 $\boldsymbol{M}$ 间的递归关系对 $\boldsymbol{M}^{-1}$ 求解:

$$\boldsymbol{M}_k = \begin{bmatrix} p\boldsymbol{M}_{k/2} & (1-p)\boldsymbol{M}_{k/2} \\ (1-p)\boldsymbol{M}_{k/2} & p\boldsymbol{M}_{k/2} \end{bmatrix} \tag{6}$$

将求解 $\boldsymbol{M}^{-1}$ 的过程简化:

$$\boldsymbol{M}_k^{-1} = \frac{1}{2p-1} \begin{pmatrix} \boldsymbol{M}_{k/2}^{-1} \\ \boldsymbol{M}_{k/2}^{-1} \end{pmatrix} \begin{bmatrix} p\boldsymbol{E}_{k/2} & (p-1)\boldsymbol{M}_{k/2} \\ (p-1)\boldsymbol{E}_{k/2} & p\boldsymbol{E}_{k/2} \end{bmatrix} \tag{7}$$

XMASK 方法可将重构支持度的时间复杂度由 $O(8^n)$ 减少至 $O(2^n)$,从而使挖掘方法在时间性能上增强 2 个数量级,大大提高计算效率,其在求解概率矩阵方面的可靠性较高,然而在时间复杂度减少到一定程度时,XMASK 方法无法保证挖掘出的多层次分布式网络数据的可靠性[15,16]。

在时间复杂度要求较高的情况下,通过下述改进过程对随机扰动方法进行改进。

随机扰动挖掘方法是在经典 Apriori 方法的基础上实现的,首先生成频繁 1-项集,然后生成频繁 k-项集,最终获取强关联规则[17]。随机扰动方法和经典 Apriori 方法的不同之处就在于对项集的计数问题[18]。随机扰动算法需依据歪曲后数据集对实际多层次分布式网络中数据集项集的支持度进行估测,针对 k-项集,需研究 2^k 种变化情况,所以随机扰动挖掘方法的计算开销是非常大的。

通过随机扰动方法挖掘的多层次分布式网络数据集是二维稀疏布尔矩阵,然而在对歪曲项集进行计数时,因为布尔数据集的特性,所有数据集的计数均存在一定的联系[19]。举例说明,在对二次频繁集 $\{A,B\}$ 进行计算时,仅需获取“11”的数量即可,其他取值组合为

$$\begin{cases} 10: |A \cap \overline{B}| = |A| - |A \cap B| \\ 01: |\overline{A} \cap B| = |B| - |A \cap B| \\ 00: |\overline{A} \cap \overline{B}| = U - |B| - |A| + |A \cap B| \end{cases} \tag{8}$$

式(8)中, A 、B 的取值可通过之前的 1-项集挖掘获取。在对 n-项集进行计算的过程中,可将上述规则改为

$$N(\overline{A_1}, \overline{A_2}, \cdots, \overline{A_m}, B_1, B_2, \cdots, B_n) = N(B_1, B_2, \cdots, B_n) + \sum_{k=1(x_1, x_2, \cdots, x_k) \subset \{1,2,\cdots,m\}}^{n} (-1)^k N(A_{x1}, A_{xm}, \cdots, A_{xk}, B_1, B_2, \cdots, B_n) \tag{9}$$

通过式(9),针对某 k 次候选频繁项集,仅需在

歪曲数据集合中查找取值都是 1 的项集数量,剩余组合数量可依据之前得到的频繁项集取值全是 1 的计数获取[20]。在对多层次分布式网络数据进行挖掘时,利用哈西链对所有取值均为 1 的项集数量进行保存,以便于后续挖掘。采用上述改进过程后,计算开销将大大降低。

3 实验结果测试

3.1 挖掘精度测试

为了验证本文提出挖掘方法的有效性,将采集的 3 个标准的多层次分布式网络文本数据集作为测试对象,主要包括 20Nc、IndustlySec 和文本数据集,其中 20Nc 为常见的文本数据集,其中含有 30 个新闻组的近 23 500 篇新闻,共 19 939 个文本;IndustlySec 为网页数据集,其中含有 96 763 个网页文本,共 116 个类别;Web 数据集中含有 46 个类,共 5 147 个网页。

实验通过典型的测试方式对本文方法的有效性进行测试,分别是训练测试和 k 重交叉校验方式,训练测试方式为常用的测试方式,其将数据样本划分为训练集与测试集两部分,利用测试集对数据挖掘方法进行测试。而 k 重交叉校验把数据样本分割为 k 份,每次实验均取其中的 $k-1$ 份当成训练集,将其余部分作为测试集进行测试。

为了不失一般性,针对各个实验均重复进行 20 次,10 次训练测试,10 次 k 重交叉校验测试,将 20 次实验结果均值作为测试结果。

下面将神经网络方法和贝叶斯方法作为对比,通过三种挖掘方法对前述 3 个文本数据集进行测试,将挖掘精度、查全度和重合度作为测试指标,三个指标的计算公式如下,三种方法下各指标比较结果见表 1。

挖掘精度为准确挖掘数据量占总挖掘数据量之比,计算公式如下:

$$p(\eta) = \frac{|\eta \cap R|}{|\eta|} \tag{10}$$

式(10)中,η 为总挖掘数据量;R 为真实待挖掘数据量。

查全度计算公式如下:

$$r(\eta) = \frac{|\eta \cap R|}{|R|} \tag{11}$$

重合度是采用挖掘方法挖掘出的数据和真实需挖掘数据间相似度的体现,重合度越高,挖掘效果越好,其计算公式如下:

$$\deg(r,c) = \frac{|r \cap c|}{|r|} \tag{12}$$

式(12)中,r 为真实的待挖掘数据;c 为采用挖掘方法挖掘的数据。

表 1 三种方法挖掘准确性比较结果

Table 1 Comparison results of mining accuracy of the three methods

方法	项目	数据集		
		20Nc	IndustlySec	Web
本文方法	精度/%	97.39	91.72	87.66
	查全度/%	96.75	83.89	86.21
	重合度/%	94.36	85.92	82.16
神经网络方法	精度/%	92.35	83.69	75.22
	查全度/%	90.16	79.33	72.51
	重合度/%	87.95	76.22	79.69
贝叶斯方法	精度/%	86.37	81.66	72.51
	查全度/%	81.59	72.62	68.76
	重合度/%	83.22	75.17	62.16

由表 1 中数据可以看出,针对 20Nc 数据集,本文方法的挖掘精度为 97.39%,高于神经网络方法的 92.35% 和贝叶斯方法的 86.37%,且查全度与重合度也明显高于其他两种方法,说明针对 20Nc 数据集,本文方法的挖掘准确性很高。虽然本文方法针对 IndustlySec 数据集和 Web 数据集的挖掘精度、查全度与重合度低于 20Nc 数据集,但与神经网络方法、贝叶斯方法相比,本文方法的挖掘准确性最高,验证了本文方法在挖掘精度方面的性能。

3.2 本文方法挖掘结果测试

将挖掘难度较高的 Web 数据集作为研究的对象,对改进前和改进后数据挖掘结果进行比较,结果分别见图 3 和图 4。

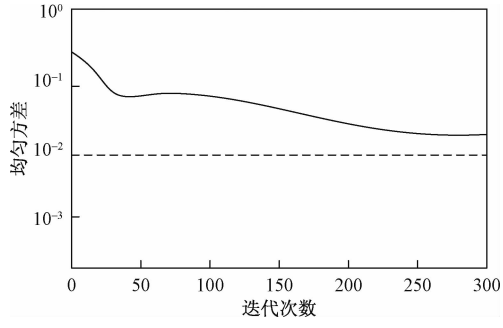

图 3 改进前数据挖掘结果

Fig. 3 Data mining results before improvement

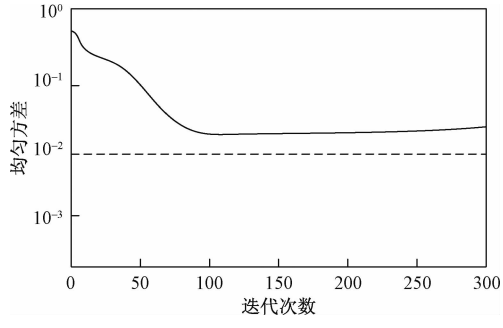

图 4 改进后数据挖掘结果

Fig. 4 Results of improved data mining

由图 3 和图 4 可以发现,改进后挖掘效果明显优于改进前,详细情况见表 2。

表 2　改进前后挖掘效果比较

Table 2　Comparison of mining results before and after improvement

指标	改进前	改进后
收敛速度	慢	快
是否有极小值	是	否
迭代次数	260	80
挖掘时间	长	短

经比较可以看出,改进后挖掘方法的收敛速度明显快于改进前,改进前需迭代 240 次才可收敛,而改进后只需迭代 80 次即可收敛,收敛速度明显提高,挖掘时间明显降低。且改进后算法不会出现极小值,这主要是因为改进后的方法挖掘准确性更好。

3.3　数据分布稀疏时挖掘结果

多层次分布式网络中的数据可能会呈稀疏分布,在数据分布稀疏时,主要研究本文方法、神经网络方法和贝叶斯方法的挖掘结果。

图 5 是一种数据分布较为稀疏的情况,图中的点呈两条斜线,两条斜线中各存在 6 个点,其中,斜线中相邻两点间的距离为 $\sqrt{2}$,而其中的点代表数据。

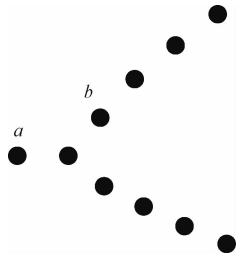

图 5　数据分布较稀疏时数据分布

Fig. 5　Data distribution when the data distribution is sparse

图 5 中 a、b 两个点依次运行本文方法、神经网络方法和贝叶斯方法,得到的结果分别如图 6 ~ 图 8 所示。

分析图 6、图 7 和图 8 发现,本文方法针对 a 点的离群因子值明显低于 b 点,而在神经网络方法和贝叶斯方法下,b 点离群因子大部分都高于 a 点,说明本文方法挖掘结果更加有效。

3.4　不同密度区域相互靠近时挖掘结果

在图 9 二维平面点图中,含有两个密度存在差异的区域,在该数据集中运行改进前和改进后的挖掘方法,对离群数据进行挖掘。为了便于分析,取挖掘方法挖掘出的离群因子值排前三位的点,用“×”标记,挖掘结果如图 10 和图 11 所示。

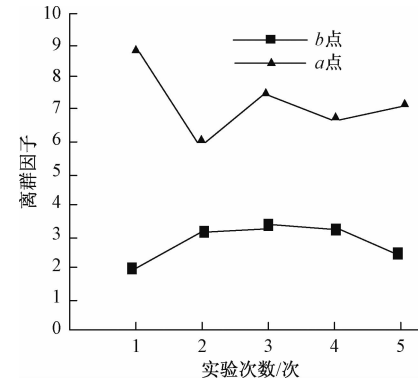

图 6　本文方法数据挖掘结果

Fig. 6　Results of data mining in this paper

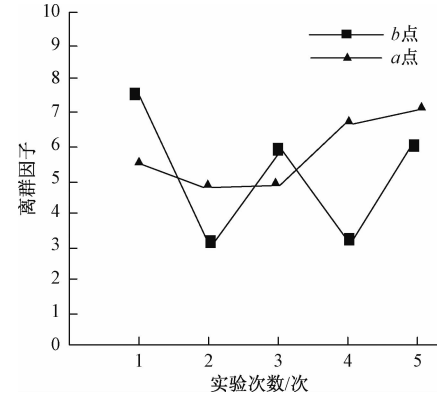

图 7　神经网络方法数据挖掘结果

Fig. 7　Data mining results of neural network method

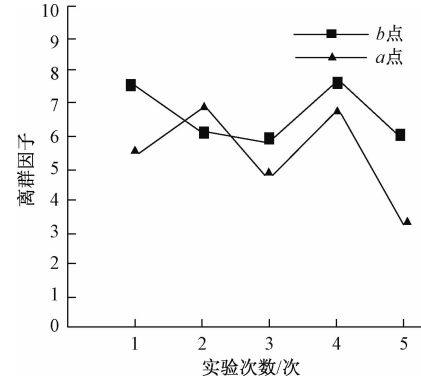

图 8　贝叶斯方法数据挖掘结果

Fig. 8　Data mining results of Bayesian method

由图 10 和图 11 可知,改进后挖掘方法挖掘的离群数据离群程度明显更高,进一步验证了本文方法的挖掘效果。

4　结论

提出一种新的多层次分布式网络数据挖掘改进方法,给出多层次分布式网络结构。介绍了常用的随机扰动数据挖掘方法,针对其弊端对其进行改进。实验结果表明,所提方法可靠有效。

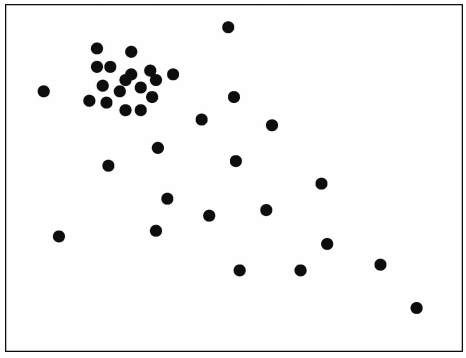

图 9 不同密度区域数据

Fig. 9 Different density region data

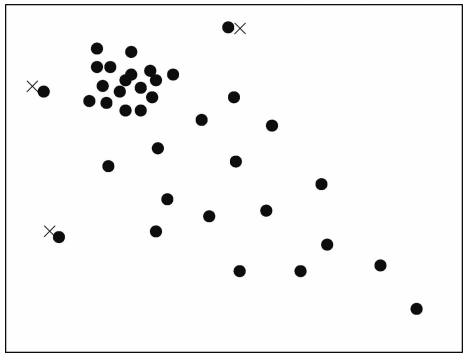

图 10 改进前挖掘结果

Fig. 10 Results of improved mining

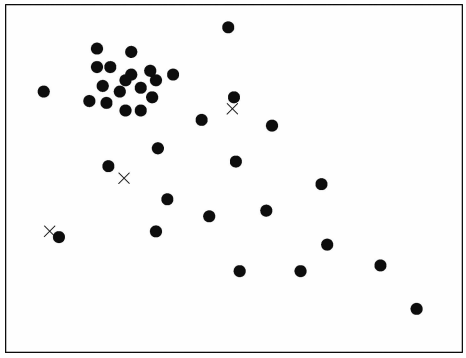

图 11 改进后挖掘结果

Fig. 11 Results of improved mining

参 考 文 献

1 卢琦蓓,郭飞鹏. 基于改进型 FP-Tree 的分布式关联分类算法. 山东大学学报(理学版),2014;49(1):71—75
Lu Qibei, Guo Feipeng. Distributed associative classification algorithm based on improved FP-Tree. Journal of Shandong University (Natural Science), 2014; 49(1):71—75

2 任高举,白亚男. 多媒体智能教学系统中特定数据挖掘方法研究. 电子设计工程, 2016; 24(11):4—7
Ren Gaoju, Bai Yanan. Research on specific data mining methods in multimedia intelligent tutoring system. Electronic Design Engineering, 2016; 24(11):4—7

3 王 昊,师 卫,李 欢. Hadoop 下基于贝叶斯网络的气象数据挖掘研究. 电子器件,2016,39(4):841—846
Wang Hao, Shi Wei, Li Huan. The research of meteorological data mining using bayesian network based on Hadoop. Chinese Journal of Electron Devices, 2016; 39(4):841—846

4 王 茜, 刘书志. 基于密度的局部离群数据挖掘方法的改进. 计算机应用研究, 2014; 31(6):1693—1696
Wang Qian, Liu Shuzhi. Improvement of local outliers mining based on density. Application Research of Computers, 2014; 31(6):1693—1696

5 Chen C, Li Y, Yan C, *et al*. An improved multi-resolution hierarchical classification method based on robust segmentation for filtering ALS point clouds. International Journal of Remote Sensing, 2016; 37(4):950—968

6 邓 欣,宁 芊. 基于开源的分布式山洪监测数据系统设计. 计算机测量与控制, 2016; 24(10):54—56
Deng Xin, Ning Qian. Distributed flash flood monitoring system design based on open source. Computer Measurement & Control, 2016; 24(10):54—56

7 刘丽娇,陶俊才,肖晓军,等. 电信大规模社交关系网络图数据挖掘研究. 电信科学,2015;31(1):23—31
Liu Lijiao, Tao Juncai, Xiao Xiaojun, *et al*. Research on large-scale social network graph data mining in telecommunication. Telecommunications Science,2015;31(1):23—31

8 巩树凤,张岩峰. EDDPC:一种高效的分布式密度中心聚类算法. 计算机研究与发展,2016;53(6):1400—1409
Gong Shufeng, Zhang Yanfeng. EDDPC:an efficient distributed density peaks clustering algorithm. Journal of Computer Research and Development,2016;53(6):1400—1409

9 高 静,董振华,郭 峰. 网络差异数据的优化挖掘模型仿真分析研究. 微电子学与计算机,2016;33(7):136—139
Gao Jing, Dong Zhenhua, Guo Feng. Simulation analysis of network difference data optimization mining. Microelectronics & Computer, 2016; 33(7):136—139

10 蔡斌雷,郭 芹. 面向多数据流的同现模式快速挖掘方法. 河北大学学报(自然科学版),2016;36(3):318—326
Cai Binlei, Guo Qin. Fast mining co-occurrence pattern over multiple data streams. Journal of Hebei University (Natural Science Edition), 2016; 36(3):318—326

11 Bachega L, Hariharan S, Bouman C, *et al*. Distributed signal decorrelation and detection in multi view camera networks using the vector sparse matrix transform. IEEE Transactions on Image Processing, 2015; 24(12):6011—6024

12 宋顶利,张 昕,于复兴. 分布式优化 Apriori 算法的交通运行状态数据分析模型. 科技通报,2016;32(10):202—206
Song Dingli, Zhang Xin, Yu Fuxing. Data analysis model of traffic running based on the distributed optimal apriori algorithm. Bulletin of Science and Technology, 2016; 32(10):202—206

13 陆莉莉, 张永潘, 谈海宇, 等. 大数据分类挖掘算法及其概念漂移应用研究. 计算机科学与探索, 2016; 10(12):1683—1692
Lu Lili, Zhang Yongpan, Tan Haiyu, *et al*. Research on classification algorithm and concept drift based on big data. Journal of Frontiers of Computer Science & Technology, 2016; 10(12):1683—1692

14 高梦超, 胡庆宝, 程耀东, 等. 基于众包的社交网络数据采集模型设计与实现. 计算机工程, 2015; 41(4):36—40

Gao Mengchao, Hu Qingbao, Cheng Yaodong, *et al*. Design and implementation of crowdsourcing-based social network data collection model. Computer Engineering, 2015; 41(4):36—40

15 李文昊, 李海芳. 确定性分布式数据库中长事务处理方法研究. 科学技术与工程, 2016; 16(13):92—95
Li Wenhao, Li Haifang. Research on the method of long transaction in deterministic distributed database. Science Technology and Engineering, 2016; 16(13):92—95

16 Corbellini A, Mateos C, Godoy D, *et al*. An architecture and platform for developing distributed recommendation algorithms on large-scale social networks. Journal of Information Science, 2015; 41(5): 686—704

17 于彦伟, 齐建鹏, 陆云辉, 等. 时空轨迹大数据分布式蜂群模式挖掘算法. 计算机工程与科学, 2016; 38(2):255—261
Yu Yanwei, Qi Jianpeng, Lu Yunhui, *et al*. Distributed swarm pattern mining algorithm in big spatio-temporal trajectory data. Computer Engineering and Science, 2016; 38(2):255—261

18 冯　勇,尹洁娜,徐红艳. 基于垂直频繁模式树带有负载均衡的分布关联规则挖掘算法. 计算机应用, 2014; 34(2):396—400
Feng Yong, Yin Jiena, Xu Hongyan. Distributed rules mining algorithm with load balance based on vertical FP-tree. Journal of Computer Applications, 2014; 34(2):396—400

19 李　宁,罗文娟,庄福振,等. 基于 MapReduce 的并行 PLSA 算法及在文本挖掘中的应用. 中文信息学报, 2015; 29(2):79—86
Li Ning, Luo Wenjuan, Zhuang Fuzhen, *et al*. MapReduce based parallel probabilistic latent semantic analysis for text mining. Journal of Chinese Information Processing, 2015; 29(2):79—86

20 钱晓军, 范冬萍, 吉根林. 物联网差异数据库中的故障数据快速挖掘仿真. 计算机仿真, 2016; 33(1):301—304
Qian Xiaojun, Fan Dongping, Ji Genlin. Differences between the internet of things in the database failure data rapid excavation simulation. Computer Simulation, 2016; 33(1):301—304

An Improved Method of Multi Level Distributed Network Data Mining

YANG Yuan, MA Xu, CHEN Chen
(Math and Computer Science College, Ningxia Normal University, Guyuan 756000, China)

[**Abstract**] Aiming at the problem of high cost of traditional data mining methods, a new method to improve the multi-level distributed network data mining was proposed. The random disturbance data mining method commonly used, through the probability distorted technology to disrupt the original data processing multi-level distributed network set, to reconstruct the actual item set support, the probability of conversion and acquisition of data mining results. Random perturbation method has the disadvantages of low time efficiency, the time complexity is low, the random perturbation method improved by XMASK method; the time complexity is high, the corresponding improvement process. The experimental results show that the proposed method has high accuracy, short mining time and good effect.

[**Key words**] multi level　distributed network　data mining　improvement