文章编号:1671-4598(2017)12-0275-04 DOI:10.16526/j.cnki.11—4762/tp.2017.12.071 中图分类号:TP311 文献标识码:A

基于云平台的大数据资源挖掘技术研究

薛 蓓,周延怀,王晓兰
(南京师范大学 泰州学院,江苏 泰州 225300)

摘要:针对云平台下大数据资源挖掘过程准确率低、耗时长等问题,对大数据资源挖掘技术进行改进研究;利用MST聚类法对云平台数据集进行预处理,根据数据间的关联性来增强检测结果,并提高数据索引效率,将数据间的邻接矩阵作为边的权值,生成全图的MST,获取评价数据资源挖掘准确度的标准,并得到k个最小生成子树,其中的一个子树就是数据集最优聚类结果;实验结果表明,所提方法有效提高了大数据挖掘准确性,使得数据资源得到了更高效的利用。

关键词:云平台;数据资源;挖掘;技术改进

Research on Large Data Resource Mining Technology Based on Cloud Platform

Xue Bei, Zhou Yanhuai, Wang Xiaolan
(Taizhou College, Nanjing Normal University, Taizhou 225300, China)

Abstract: In order to solve the problem of low precision and long time consuming in mining large data resources under the cloud platform, the mining technology of large data resources is improved. Preprocessing of the cloud platform data sets using MST clustering method to enhance the detection results according to the relevance between data and data, improve the efficiency of the index, the adjacency matrix data as edge weights, generating graph MST, obtain evaluation data mining accuracy standard, and get k a minimum spanning tree. The results of the optimal clustering a sub tree, which is the data set. Experimental results show that the proposed method effectively improves the accuracy of large data mining, and makes data resources more efficient.

Keywords: cloud platform; data resources; excavate; technical improvement

0 引言

当今世界的科学技术发展迅速,已然成为了各国发展的经济支撑,科技创新的地位也越来越重要[1]。科技创新服务平台是经济社会中新的形式,可以根据资源整合来提升科技资源利用率,并加强“产学研”联合和发挥科技中介作用,促进科技成果转换,是国家科技创新结构中的重要构成部分[2]。依据当前的形式而言,对科技创新服务平台建设的力度加大,是适应科技迅速发展的必然趋势,同时也是推动科技社会迅猛前进的主要动力[3]。

目前人类正处于瞬息万变的环境中,经济发展与科技创新均发生了重大且深远的变革[4]。科学技术作为第一生产力,每个国家或者地区科技综合竞争力强弱,主要表现于科技资源整合水平、科技利用率和科技创新等方面[5]。每个国家或者地区总体的科技创新服务平台是国家重要的组成部分,科技创新服务平台的构建作为国家科技比较基础的条件平台,依据本地区或本国的实际情况构建科技创新服务平台,该问题是实现国家科技进步的关键,也是落实中央制度的具体行动[6]。

综上所述,对科技创新服务平台中的数据利用云平台进行存储,并实现数据资源的高效利用,需要对云平台数据的大数据资源进行挖掘[7]。

收稿日期:2017-10-14; **修回日期**:2017-10-24。
基金项目:2015年泰州市软科学研究计划项目(RKX201529)。
作者简介:薛 蓓(1985-),女,江苏泰兴人,硕士,助理研究员,主要从事计算机技术,教育管理方向的研究。
周延怀(1954-),男,江苏镇江人,大学,教授,主要从事物理学方向的研究。

1 大数据资源挖掘技术原理

在对云平台的大数据资源挖掘技术进行研究之前,首先对大数据资源挖掘技术原理进行分析。大数据资源挖掘技术原理主要包括新科技创新服务平台结构体系和大数据资源挖掘依据两部分。新科技创新服务平台结构体系是对数据资源进行挖掘的平台环境,对平台环境加以介绍,在充分掌握平台结构体系之后,能够更加准确地制定出数据资源挖掘技术的改进方案。大数据资源挖掘依据为数据资源挖掘技术的改进方案提供依据,并给出了数据挖掘技术实现的流程。

1.1 科技创新服务平台结构体系

科技创新服务平台结构如图1所示,其中主要分为:用户层,网络层,资源层,运输层和数据层[8]。用户层为平台的使用者,其中包含获取有关科技服务与资源的客户,还包含提供科技服务和资源,进而受益的供应商;网络层是平台的窗口,是其他层的线上媒介,展示供需信息的时,提供线上的交易平台;资源层将运营层当作中介,为用户层供给科技资源与服务,其中包括有形资产、无形资产持有者、专业技术服务执行者;运营层是平台核心,根据线下运营者,线下的服务执行者以及线上网站建设的维护者组成;数据层将云平台当作载体,实现后台数据资源的挖掘分析,跟踪并完善客户的需求,进而完成数据的推送和资源配置[9]。

1.2 大数据资源挖掘依据

根据科技创新服务平台结构体系中数据层中的大数据挖掘和分析需求,利用图2中的数据挖掘原理实现数据资源的挖掘。

由图2可知,首先对大数据资源进行获取,将获取的大数据资源分为两部分,一部分进行数据预处理备用,另一部分通过数据处理函数等的综合计算,对数据进行充分分析,再将分析好的数据进行分类,最终实现数据资源的挖掘。

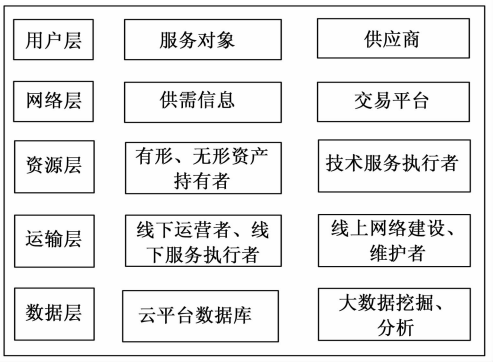

图 1 科技创新服务平台结构体系

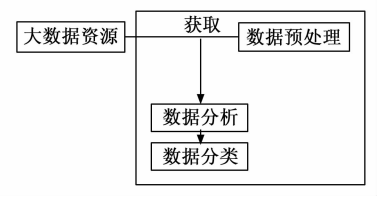

图 2 大数据资源挖掘原理

2 云平台下大数据资源挖掘技术改进研究

要使云平台中的数据利用率达到最大，通过图 2 的大数据资源萃取原理，利用 MST 聚类法实现云平台的大数据资源高效挖掘。通过 MST 聚类法的大数据聚类分析以高维大数据作为背景，于图论的基础上，对数据集进行预处理，根据量化各数据对象间的关联性组建邻接矩阵，将数据点当作顶点，各个数据间的邻接矩阵作为边的权值，组建一个全图，并生成此全图的 MST，依据实际的问题以及数据分布的状态，按照边权值由大到小分割 MST 的边，获得 k 个最小的生成树子树，其中的一个子树就是数据集中最优的聚类结果。详细过程如下：

2.1 数据邻接矩阵的建立

所谓的邻接矩阵就是根据数据组表示数据点间关联的数据矩阵[10]，假设图 G 代表赋权网络图，则能够将其定义为：

$$A_{ij}=\begin{cases} w_{ij} & \text{假设}(v_i,v_j)\text{或者}<v_i,v_j>\in E(G) \\ 0 & \text{假设}(v_i,v_j)\text{或者}<v_i,v_j>\notin E(G) \end{cases} \tag{1}$$

式中，w_{ij} 代表边的权值，图 3 和图 4 展现了从图生成为邻接矩阵全过程：

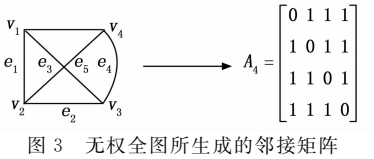

图 3 无权全图所生成的邻接矩阵

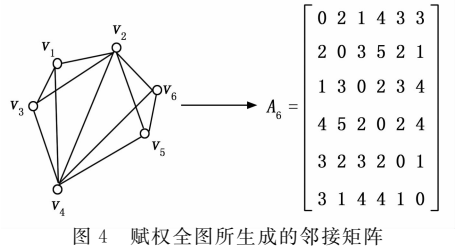

图 4 赋权全图所生成的邻接矩阵

根据图 3 和图 4 可得知，网络图邻接矩阵是对称矩阵，其中矩阵中的第 i 行第 j 列个元素就是赋权图内的顶点 v_i 和顶点 v_j 关联距离 w_{ij}。根据该矩阵可增强数据关联性检测效果，促使数据分类形状多样化。

2.2 大数据资源聚类

在这里把数据对象点间距离的关联性当作权值对两点间相关性赋权，组建出数据对象点间的邻接矩阵，获得邻接矩阵表示的全图，根据生成最小树法获得此权值全图中的一个 MST，并按 MST 边赋值大小分割最小树边，获得若干最小树子树，各子树就是一个最优 Cluster（簇），详细过程如下。

将数据集点进行初始化：

初始化就是将指标变量量纲相异或者数量级差别比较大的数据对象标准化，并统一数据对象类型以及单位，使该数据对象可以进行比较与计算。

针对待挖掘数值属性，这里根据各数据对象之间距离表示数据间的相关性。并选取欧式距离对数据点进行定义：$x_1,x_2,\cdots x_i,\cdots x_j,\cdots x_n;i,j=1,2,\cdots,n$，由此可得：$d(x_i,x_j)=\left[\sum_{i,j}(x_{ik}-x_{jk})^2\right]$，其中，$d(x_i,x_j)$ 代表欧式距离，x_{ik} 和 x_{jk} 代表两个数据点，将各数据对象点当作顶点，且将数据对象间距离当作权值，组建一个二维邻接矩阵，根据矩阵生成赋值全图 $G(V,E,W)$，$V=\{v_i,v_j,\cdots v_n\}$ 在图中表示着 n 个数据对象所建立的点集，$v_i,v_j,\cdots v_n$ 为全图节点，$E=\{e_{ij}\mid 1\leqslant i,j\leqslant n\}$ 为图中的 n^n 条边集合，式中的 e_{ij} 代表连接 v_i 和 v_j 间的边。

则所有获得的生成树权值与 W 最小的就是 MST，也就是满足：

$$W=\min\left\{\sum_{i,j}W(v_i,v_j)\right\}=\min\left\{\sum_{i,j}d(x_i,x_j)\right\} \tag{2}$$

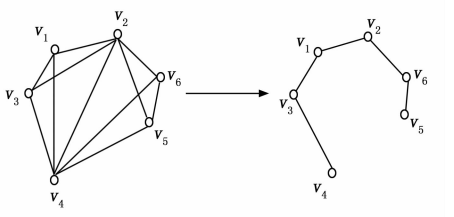

图 5 赋权全图 MST

由大至小切割 MST 赋值边，获得 MST 若干子树，也就是将 MST 中最大的赋权边 e_m 切割移除，e_m 满足下列条件：$e_m=\max\{W(v_p,v_q)\}=\max\{d(v_p,v_q)\}$。

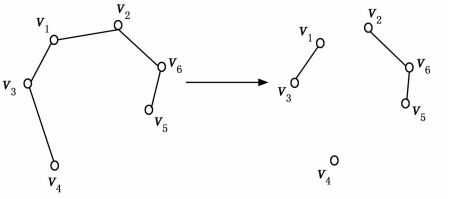

图 6 切割最小的生成树中最大两条边之后获得三个子树

经过切割，所获得的各子树边就是全局最优的一个类簇，假设在切割 k 条最大的边之后，会获得 $k+1$ 个类的最优类簇，照比传统的数据挖掘方法，得到的聚类结果更加准确。

3 实验结果分析

为了验证基于云平台的大数据资源挖掘技术的有效性和可行性，实验针对改进技术的数据关联性、聚类召回率、聚类时所出现的形状、数据索引效率及挖掘精度五项指标进行测试。首先给出实验数据的由来及实验平台环境，通过实验模拟制定

实验方案，执行实验操作，对实验结果进行分析。

并对分析所得结果进行总结，具体实验描述如下：

3.1　实验数据的由来

实验中，采用加利福尼亚的机器学习数据集，将两种不同的数据集划分为四组不同数量的数据集，并分别与本地的云平台进行连接，本地的实验环境是 Google App Engine SDK、AMD 双核 1.6、2 G 内存。将 Average－Linkage 聚类法、K－Means 聚类法以及 SOM 聚类法应用 MATLAB 软件完成实验模拟，将 MST 聚类法利用 LINGO 软件实现模拟分析。

3.2　实验模拟

进行实验模拟时，将类簇划分为 3 个，可获得下列聚类效果如图 7 所示。

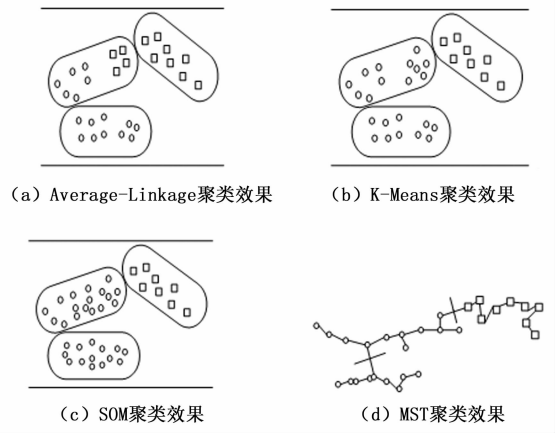

图 7　不同技术聚类效果对比

观察图 7 可知，图 7（a）～图 7（d）分别是采用 Average－Linkage 聚类法、K－Means 聚类法、SOM 聚类法及 MST 聚类法对数据进行聚类的效果图，前三种聚类法的聚类效果较为相似，最后一种聚类法的聚类效果比较特殊。Average－Linkage 聚类法的聚类精度较低，在 3 个数据集中，存在数据混杂的现象，且聚类数据量少。K－Means 聚类法的聚类精度相对较高一些，由图 7（b）可以看出，3 个数据集中没有混杂数据，但是聚类的数据量依然较少。观察图 7（c）可知，SOM 聚类法的聚类精度较前两种方法高，聚类的数据量也明显增多。采用 MST 聚类法进行数据的聚类，由图 7（d）可看出，它能够将 3 种数据集按照不同类别进行无缝聚类，不仅聚类精度高，聚类数据量大，且聚类密度高，有效节省了聚类空间。对比 4 种不同聚类方法的数据聚类效果，MST 聚类效果远远优于其他 3 种聚类方法，改进的云平台大数据资源挖掘技术正是应用这个方法对数据进行聚类，充分说明改进技术聚类效果更好，验证了改进技术的有效性。

将实验数据导至网格中，网格分为横纵坐标，纵坐标代表数据量，横坐标代表时间，观察改进技术数据关联性检测效果。

根据图 7 分析图 8，由图 7 已经得知不同聚类方法的数据聚类效果是不同的。据经验 Average－Linkage 聚类法并不需要先确定 k 值，不过数据挖掘程序一旦运行，就无法更正了，这也就影响了数据聚类的正确性；K－Means 聚类法的参数 k 值为随机给定的，由此致使聚类结果不一致，导致数据聚类的效果不理想；SOM 聚类法具有比较高的聚类准确度，不过查阅资料可知，该聚类法是基于欧式距离且处于反复的循环过程，这使得数据的维度越高，其数据聚类的收缩速度就越慢，严重耗时；改进技术的数据关联性检测效果与聚类效果直接相关，由图 8 可知，改进技术的数据关联性检测效果随着时间变化越来越显著，当实验时间为 40 s 时，数据库资源达到最高值为 78 万个，检测出改进技术的数据关联性较高。产生这种情况主要是因为改进技术通过量化各数据对象间的关联性组建邻接矩阵，以此增强了数据关联性检测效果，并依据实际的问题以及数据分布的状态，按照边权值由大到小分割 MST 的边，将其作为评价数据资源挖掘准确度的标准，提高了数据聚类正确性，进而提高改进技术数据关联性，实验结果表明，改进技术的数据关联性高。

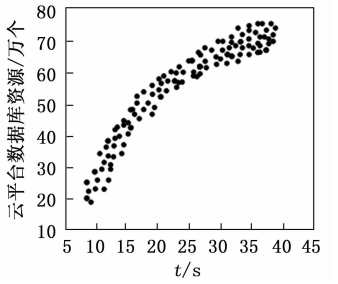

图 8　改进技术数据关联性检测效果

为了验证改进技术的数据聚类错误召回率，给出召回率计算公式。假设数据对象集能够分割成 k 个簇，其中 *Cluster* i 代表第 i 簇，n 代表 c_i 中的数据数量，a_i 代表准确聚类至 *Cluster* i，不过错误地聚类至其他数据类簇中的数据数量。则召回率表达式为：$REC = \frac{1}{k}\sum_{n=1}^{k} * \frac{a_i}{c_i}$。

根据上述召回率公式，利用软件完成聚类程序之前，在程序的最前面和程序的最后面，分别添加一时间函数，获得的时间差就是该聚类法执行聚类时的时间效率，聚类期间所出现的形状表达的就是数据挖掘方法所展现出的聚类样式是否多变。表 1、表 2 和表 3 分别代表数据聚类错误的召回率、聚类时所出现的形状以及经过挖掘之后的数据索引效率。并以此为依据检验改进技术的整体效果。

表 1　不同技术召回率对比

聚类方法	召回率/%
Average－Linkage	0.759
K－Means	0.845
SOM	0.912
MST	0.998

分析表 1 可知，应用 Average－Linkage 聚类法的数据挖掘技术，其数据聚类错误召回率为 75.9%；应用 K－Means 聚类法的数据挖掘技术，其数据聚类错误召回率为 84.5%；应用 SOM 聚类法的数据挖掘技术，其数据聚类错误召回率为 91.2%；改进技术采用 MST 聚类法，其数据聚类错误召回率为 99.8%。对比 4 种引入不同聚类方法的数据挖掘技术的实验结果，明显看出改进技术的数据聚类错误召回率最高，近乎接近了 100%，改进技术建立了数据对象间的关联性量化后产生的邻接矩阵，并对数据的均值、标准差、极差等项进行了计算，由此提高了数据聚类的召回率，实验结果验证了改进技术的有效性。

表 2 不同技术聚类时所出现的形状

聚类方法	聚类时出现的形状
Average－Linkage	球形
K－Means	凸形
SOM	球形或者凸形
MST	任何形状

分析表 2 可知，应用 Average－Linkage 聚类法的数据挖掘技术，其聚类时所出现的形状为球形；应用 K－Means 聚类法的数据挖掘技术，其聚类时所出现的形状为凸形；应用 SOM 聚类法的数据挖掘技术，其聚类时所出现的形状为球形或凸形；改进技术采用 MST 聚类法，其聚类时所出现的形状为任何形状。对比 4 种引入不同聚类方法的数据挖掘技术，其聚类时所出现的形状，明显看出改进技术聚类时出现的形状没有局限性，可对任意形状进行聚类，聚类范围广，提高了改进技术的数据聚类精度，改进技术对所生成的邻接矩阵进行了赋权，在增强数据关联性检测效果的同时，也使改进技术聚类时所出现的形状变得多样化，充分说明改进技术更优良。

表 3 不同技术挖掘后的数据索引效率

聚类方法	索引字数量/个	索引时间/s
Average－Linkage	2	0.5
K－Means	2	0.4
SOM	2	0.6
MST	2	0.1

在索引字数量为 2 个的情况下，对 4 种引入不同聚类方法的数据挖掘技术的索引效率进行测试，分析表 3 可知，应用 Average－Linkage 聚类法的数据挖掘技术，其索引时间为 0.5 s；应用 K－Means 聚类法的数据挖掘技术，其索引时间为 0.4 s；应用 SOM 聚类法的数据挖掘技术，其索引时间为 0.6 s；改进技术采用 MST 聚类法，其索引时间为 0.1 s。对比 4 种引入不同聚类方法的数据挖掘技术的实验结果，明显看出改进技术的数据索引效率最高，近乎是其他三种数据挖掘技术索引效率的五分之一，索引效率大幅度提升，这是因为改进技术依据实际问题及数据分布状态，按照边权值由大到小分割 MST 的边，从而实现挖掘后的数据索引效率的提升，实验结果验证了改进技术的实用性。

为了验证改进技术能够高精度地对大数据资源进行挖掘，以传统技术作为对照组，实验共进行 6 次，记录每次试验的不同技术数据资源挖掘情况，并计算其精度。挖掘精度对比实验，实验结果如下：

观察图 9 可知，经过 6 次对比实验，采用文献［7］技术对数据进行挖掘，其数据挖掘精度随实验次数的增大逐渐减小，但减小的幅度并不大，曲线基本保持平稳状态，其平均数据挖掘精度为 35%，精度较低。采用文献［8］技术对数据进行挖掘，其数据挖掘精度随实验次数的增大基本保持不变，曲线十分平稳，平均数据挖掘精度为 18%。采用改进技术对数据进行挖掘，其数据挖掘精度初始值就已达到 80%，且曲线十分稳定，只有在第 4 次实验时，出现了挖掘精度最低值为 75%，在第 6 次实验时，出现了最大挖掘精度为 85%。对比文献［7］技术、文献［8］技术及改进技术可以明显看出，改进技术的数据资源挖掘精度远远高于文献［7］技术、文献［8］技术的数据资源挖掘精度，且通过每一次实验结果的对比，可以看出改进技术不仅挖掘精度较高，且均能稳定在 80%左右，充分说明改进技术的稳定性更好，实用性更强。

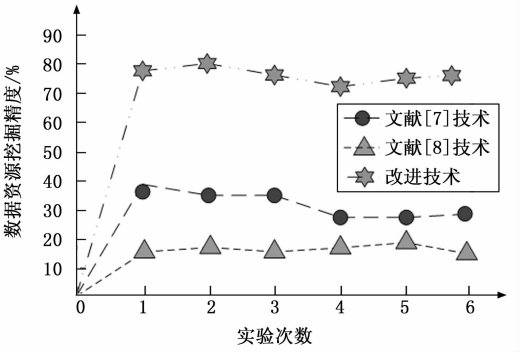

图 9 不同技术数据资源挖掘精度对比

综合以上实验结果可知，改进的云平台大数据资源挖掘技术通过引进 MST 聚类方法，其数据关联性好，数据聚类错误召回率高，聚类时出现形状多样化，数据索引效率高，且挖掘精度高，具有一定的有效性和实用性。

4 结论

根据互联网＋的科技创新服务平台，通过 O2O 模式把科技服务当作一种商品，并充当科研机构以及企业间的中介与桥梁，能够有效地满足企业创新服务需求，同时也激发了企业创新的活力，大大提升了平台的效能，不过在市场推动机制还未完全建立时，存在平台发展后劲亟待加强等问题，要保障平台稳定发展，就需要对其中的云平台中大数据资源进行挖掘。

提出一种 MST 数据聚类挖掘法，根据图论理论，利用数据间的关联性分析建立邻接矩阵，采用各个数据间邻接矩阵边的权值建立全图，并产生全图的 MST，按边权值大小对 MST 进行切割，直到获得最优簇。并通过实验证明，该方法具有可行性。

目前大众对大数据的连接以及运用只是停留在初期，云平台大数据越来越呈现出迅猛增长的趋势，由此该文未来会在更加高频以及高维复杂的数据挖掘上作进一步地研究和分析。

参考文献：

[1] 吴晓英，明均仁. 基于数据挖掘的大数据管理模型研究［J］. 情报科学，2015，32（11）：131－134.
[2] 欧阳秋梅，吴　超. 从大数据和小数据中挖掘安全规律的方法比较［J］. 中国安全科学学报，2016，26（7）：1－6.
[3] 部凯英，杨宜勇. 中国互联网＋社会保障信息系统构建——基于大数据挖掘视角［J］. 经济与管理研究，2016，37（5）：83－89.
[4] 马昱欣，曹震东，陈　为. 可视化驱动的交互式数据挖掘方法综述［J］. 计算机辅助设计与图形学学报，2016，28（1）：1－8.
[5] 申　琢，谭章禄. 基于数据挖掘的煤矿大数据可视化管理平台研究［J］. 中国煤炭，2016，42（12）：86－89.
[6] 张继荣，王向阳. 基于 XML 数据挖掘的 Apriori 算法的研究与改进［J］. 计算机测量与控制，2016，24（6）：178－180.
[7] 董本清，彭健钧. 复杂网络数据流中的异常数据挖掘算法仿真［J］. 计算机仿真，2016，33（1）：434－437.
[8] 王　琰. 一种多层安全相关属性标定偏好数据挖掘模型［J］. 科技通报，2015，31（12）：176－178.
[9] 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［J］. 电子设计工程，2016，24（11）：4－7.
[10] 梁风兰. 基于数据挖掘的农产品质量特性波动溯源方法［J］. 科学技术与工程，2017，17（3）：268－272.