计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(3):186−190 [doi: 10.15888/j.cnki.csa.006171]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

异构大数据网络的多维关联细粒度数据挖掘算法①

陆江东, 郑　奋, 戴卓臣

(第二军医大学 基础医学部, 上海 200433)

摘　要: 为了解决具有密度高、关联复杂的异构大数据网络的数据挖掘效率低下问题, 基于多维关联架构, 结合细粒度控制提出了数据挖掘算法. 首先, 在异构大数据网络存储与转发、处理的数据个性特征和差异化的基础上, 给出了异构大数据网络数据定义和多维关联模型. 接着, 基于大数据网络多维关联初始化进程, 通过大数据单位重构、维度置换、细粒度化和粒度均衡等, 提出了多维关联细粒度数据挖掘算法. 最后, 通过与粗粒度算法、线性化结构数据挖掘算法对比了在不同网络规模和数据规模下的执行效率. 实验结果表明, 所提算法具有更佳的执行效率.

关键词: 异构大数据; 大规模网络; 数据挖掘; 细粒度; 多维关联

引用格式: 陆江东,郑奋,戴卓臣.异构大数据网络的多维关联细粒度数据挖掘算法.计算机系统应用,2018,27(3):186–190. http://www.c-s-a.org.cn/1003-3254/6171.html

Multidimensional Association Fine-Grained Data Mining Algorithm for Heterogeneous Big Data Networks

LU Jiang-Dong, ZHENG Fen, DAI Zhuo-Chen

(College of Basic Medical Sciences, The Second Military Medical University, Shanghai 200433, China)

Abstract: In order to improve the low efficiency of data mining with high density and complex heterogeneous data network, a data mining algorithm based on multi dimension association structure is proposed. Firstly, on the basis of the data personality characteristics and the differences of the data storage, forwarding and processing in the heterogeneous large data network, the data definition and multi-dimensional correlation model of the heterogeneous data network are given. Then, based on the large data network, the paper proposes a multidimensional association fine-grained data mining algorithm based on the reconstruction of the large data units, the dimension replacement, the granularity and the granularity. Finally, the efficiency of the algorithm is compared with the coarse grained algorithm and the linear structured data mining algorithm. The experimental results show that the proposed algorithm has better performance.

Key words: Heterogeneous big data; large scale network; data mining; fine-grained; Multidimensional association

引言

随着大数据的快速发展和延伸, 已经应用于工业[1]、农业[2]、物流[3]和财务管理[4]等各个领域, 成为高速可靠应用的关键技术之一. 因此, 实现大数据的网络化[5]和高速共享成为重要研究方向之一. 然而, 在大数据网络中, 如何实现异构性大数据的高效识别和精确挖掘[6], 保障大数据处理的有效性和可信性[7]成为当前亟需解决的关键问题.

在分析最佳特征子集及其派生增长速度的基础上, 文献[8]基于高维和流格式的数据馈送的大数据, 提出了一种新的轻量级特征选择. 该特征选择可用于挖掘飞流数据, 从而加速粒子群优化类型的群搜索效率, 提高

① 收稿时间: 2017-04-25; 修改时间: 2017-05-09; 采用时间: 2017-05-16; csa 在线出版时间: 2018-02-09

© 中国科学院软件研究所 http://www.c-s-a.org.cn

分析精度和缩短处理时间. 为了保护开采数据中的敏感信息, 文献[9]提出了 Rampart 框架分类的保护方法. 文献[10]从类的相关性和上下文线索出发, 提出了一种新的多媒体大数据挖掘系统的 MapReduce 框架基础. 从深度学习出发, 以个性化特征的分布式数据为处理对象, 文献[11]设计了一种适用于广域网络的粗粒度分布式深度学习方法, 在精度、通信和响应等方面性能优越. 文献[12]所设计的机会认知和类脑智能相结合的数据挖掘算法, 通过融合时间粒度和分割时间序列, 结合马氏距离, 预测和分析大数据的相关性. 基于关联映射和生物信息网络的多维, 文献[13]设计的数据挖掘算法不仅可以降低生物信息网络复杂数据挖掘的低效率和大数据挖掘速度慢等一系列问题. 文献[14]研究了云计算中细粒度数据访问控制问题, 并提出了一种新的访问控制策略实现细化和有效执行撤销用户的操作.

上述研究对于高密度大数据的冗余性和多样化关联关系未作深入研究, 这些因素将对大数据网络的数据识别和挖掘精度产生重要影响. 本文在上述分析的基础上, 提出了适用于大数据网络, 基于多维关联架构的细粒度数据挖掘算法.

1 异构大数据网络的多维关联模型

异构大数据网络存储与转发、处理的数据具有明显的个性特征, 同类数据一致性较好, 异构类数据差异化较强. 异构大规模网络, 由于多样化用户需求、网络存储设备差异化、大数据服务多样化和非线性数据关联等特性, 终端发送的大数据结构特征弱化, 关联度及其维数成为关键因素. 此处, 用户需求的差异化和网络存储设备的特性是弱化数据结构特征和导致高关联大数据的关键因素. 于是, 从用户需求出发结合多样性建立异构大数据网络, 同时参考被弱化的数据结构特征. 对于大数据网络存储设备, 特别是大数据服务器文件系统结合异构类数据差异化特性建立关联度模型, 目的是为了解决多维度的异构大数据映射关系和关联问题.

异构大数据网络的存储设备和大数据服务器对异构大数据的管理和处理基于网络文件管理架构. 在架构中, 对数据进行了详细定义 (详见表 1), 用于全面描述异构大数据. 在大数据服务器中任意抽样大数据, 表 1 所述定义占空比如表 2 所示, 在异构大数据网络中的占空比详见图 1.

分析图 1 发现, 异构大数据网络的占空比中第 4 属性即关联性占空比极高, 但是第 6 属性即关联性极低, 这样会严重制约异构大数据网络的数据识别与挖掘效率, 而且对于字符占比高的大数据的识别率极低. 其中, 各数据属性定义的占空比分配不均, 将会严重降低异构大数据网络的系统效率和执行力.

表 1 异构大数据定义

定义	描述	定义	描述
大小	占内存空间	属地	所属服务器编号
新建时间	初始化日期	结构	是否具有结构特征
线性	是否具有线性特征	字符数	数据中属于字符类型的符号个数
关联性	是否与其它数据关联	维度	关联的维度

表 2 大数据服务器占空比情况

编号	数据属性	占空比(%)
1	大小	6.7
2	新建时间	8.9
3	线性	21.45
4	关联性	56.7
5	线性	1.23
6	关联性	78.98
7	字符数	23.1
8	维度	78.9

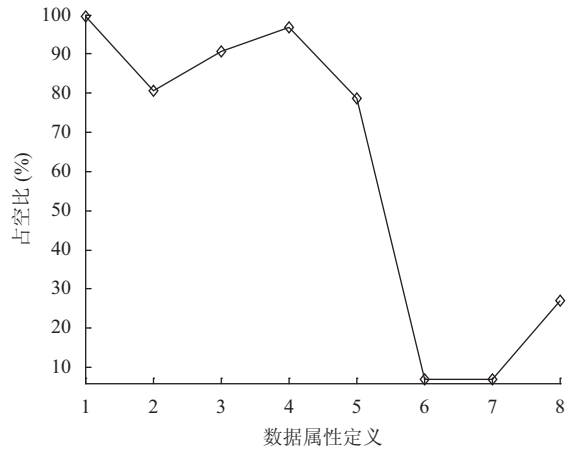

图 1 异构大数据网络占空比

因此, 在异构大数据网络中必须提高大数据维度和关联度属性的占空比, 并将这 2 个属性融合为一体, 有助于均衡异构大数据在大数据服务器和不同网络存储设备上的结构特性和非线性特征. 针对表 1 所述的数据大小、数据创建时间、数据所属设备、数据的结构和线性特征、字符数关联性和维度等定义, 进一步弱化大数据结构信息, 加强多维关联对大数据的描述

© 中国科学院软件研究所 http://www.c-s-a.org.cn

定义, 详见表 3.

于是, 异构大数据网络的多维关联具有图 2 所示的共享、存储与管理模型. 其中, 共享通过以维度为核心, 解决了大数据服务器占空比分配不合理的问题. 大数据存储以表 1 所定义的属性进行管理和查询.

表 3 多维关联定义

定义	描述
关联	异构大数据的属性定义的抽象属性与大数据空间之间的交互进程.
维度	根据异构大数据各定义属性建立的关联类.
多维关联	根据异构大数据定义及其占空比情况, 不同属性之间建立的多角度关联类型.

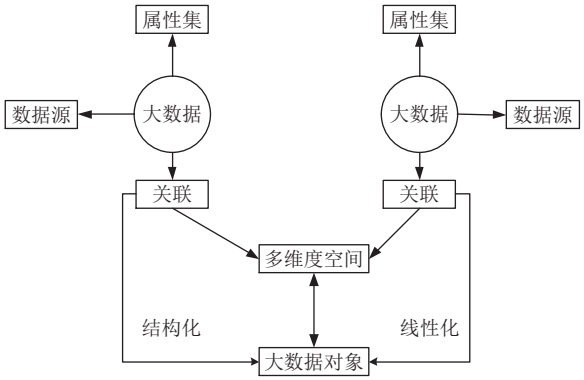

图 2 多维关联的异构大数据网络

综上, 异构大数据多维度空间的维度定义如公式 (1):

$$\begin{cases} MA(BD)=\dfrac{1}{len(BD)}\sum\limits_{i=1}^{len(BD)} f(i)|BD|^2 \\ BD=[bd_1,bd_2,\cdots,bd_{len(BD)}] \\ f(x)=x\cdot\sin\alpha \end{cases} \tag{1}$$

其中, 向量 BD 表示大数据源集合. 函数 $len(BD)$ 表示向量的长度. 函数 $f(x)$ 用于求解大数据关联度. $MA(BD)$ 表示多维空间的关联维度. α 表示维度之间的夹角.

多维关联系数 C_{BD} 可由公式 (2) 求得:

$$C_{BD}=\frac{\sin\alpha}{\ln(len(BD))} \tag{2}$$

2 多维关联细粒度数据挖掘算法

针对大数据网络不同特征的大数据, 线性化和结构化的弱化本质是多维关联. 在大数据网络服务器上, 大数据管理与处理的主要是大数据的内容与用户需求之间的关系, 特别是大数据一致性强度与数据健壮性对数据挖掘的影响, 详见图 3.

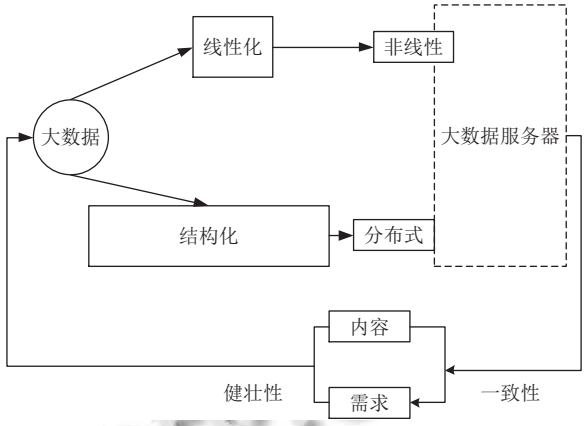

图 3 多维关联关系图

图 3 所述的大数据网络多维关联形式有助于数据挖掘. 对于大数据网络的各类用户或参与数据转发的协作终端的融合约束必须保持高度一致性. 因此, 对于多维关联下面给出形式语言描述定义. 这些定义适用于多个大数据源的交叉传输与识别. 为了更好地描述混合数据挖掘, 提高挖掘精度和识别效率, 对于形式描述语言进行多维定义. 而且, 混合异构大数据形式定义具有自主判断和决策能力, 通过异构形式结合分支进程实现. 为了保持一致性和健壮性, 在下述形式语言描述中, 逻辑上以交叉分支为主, 描述上以多个循环表达线性执行为主.

大数据对象 BO 形式描述:

BO (name, time, process, uniformity, robustness){
  name: $BD(i)$;
  time: $len(BD)$;
  process: $f(x)$;
  uniformity: $|BD-MA(BD)|$;
  robustness: C_{BD};
  return BO($C_{BD}\cdot\sin\alpha$)
}

大数据网络多维关联初始化进程:

while ($len(BD)$-k>0){
  终端发送大数据序列;
  网络存储设备接收序列;
  大数据存储;
  计算得出维度;
  建立关联矩阵;
  构建多维关联空间;
}

其中, k 表示网络存储设备的缓存最大值.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

多维关联矩阵生成进程:

```
while(1){
  根据维度分析大数据线性趋势;
  根据大数据内容分析粒度情况;
  if (内容与需求保持一致)
    break;
  else
    根据需求重建粒度和大数据内容;
}
```

对于上述返回值, 通过式 (3)–(6) 所述的单位重构、维度置换、细粒度化和粒度均衡等 4 个步骤, 实现多维关联的细粒度重置. 其中, 矩阵 TF 表示大数据源.

$$TF(len(BD))=\begin{bmatrix} t_1 & \cdots & t_n \\ \vdots & \cdots & \vdots \\ t_m & \cdots & t_{mn} \end{bmatrix}\begin{bmatrix} \sin\alpha \\ \vdots \\ \sin\alpha \end{bmatrix} \tag{3}$$

$$TF(len(BD))=\begin{bmatrix} (MA_1,\sin\alpha) & \cdots & (MA_n,\sin\alpha) \\ \vdots & \cdots & \vdots \\ (MA_m,\sin\alpha) & \cdots & (MA_{mn},\sin\alpha) \end{bmatrix} \tag{4}$$

$$TF(len(BD))=\begin{bmatrix} (MA_1,\sin\alpha) \\ \vdots \\ (MA_m,\sin\alpha) \end{bmatrix}\cdot\begin{bmatrix} f(1) & \cdots & f(n) \\ \vdots & \cdots & \vdots \\ f(m) & \cdots & f(mn) \end{bmatrix} \tag{5}$$

$$TF(n,m)=\begin{cases} TF(len(BD)), len(BD)-mn>0 \\ \dfrac{mn}{\sin\alpha}TF(len(BD))^{\mathrm{T}}, len(BD)-mn\le 0 \end{cases} \tag{6}$$

多维关联细粒度数据挖掘算法描述如下.

输入: 大数据源 BD
输出: 挖掘反馈向量 MT

```
begin:
  for c_bd 属于 C_BD
    |BD|^2;
    f(i)
    Σ_{i=1}^{len(BD)} f(i)|BD|^2;
  end
  C_BD= sinα / ln(len(BD));
  if len(BD)−mn>0
    return TF(n, m)*MT;
  else
    return mn/sinα TF(len(BD))^T*MT;
end
```

3 算法性能分析

为了更好地验证和分析本文所提出的适用于异构大数据网络, 基于多维关联细粒度的数据挖掘算法记为 MAFG-H 的执行效率, 特别是针对高密度、关联复杂的大数据网络数据挖掘的效率低下问题. 大数据网络参数设置详见表 4. 所提出的 MAFG-H 算法分别与粗粒度挖掘算法记为 CG-DM 和线性化结构数据挖掘算法记为 LS-DM.

表 4 大数据网络参数

参数	值	参数	值
数据会话数	[1, 10]	数据传输速率(Mb/s)	[100, 1000]
终端数	[5, 10]	服务器存储空间(TB)	8
冗余数据比例(%)	[10, 50]	网络传输延迟(ms)	[1, 5]

数据挖掘算法采用 Visual C++ 6.0 实现. 执行该算法的服务器内存型号为 DDR4 2400, 容量是 8 GB*2; CPU 为 Intel Xeon E3 v3、主频位 3.4 GHz. 操作系统位在 Linux 环境下 Ubuntu server. 在表 4 所述的大数据网络中抽样采集三组数据, 然后组合为独立的实例数据. 三种算法的执行效率结果如图 4–7 所示. 图 4 分析了随着数据会话数的增加, 三种算法执行效率的表现; 图 5 对比了发送大数据的终端数对算法性能的影响; 图 6 给出了三种算法在不同网络延迟下的性能表现; 冗余数据的占比对三种算法性能的执行效率影响详见图 7.

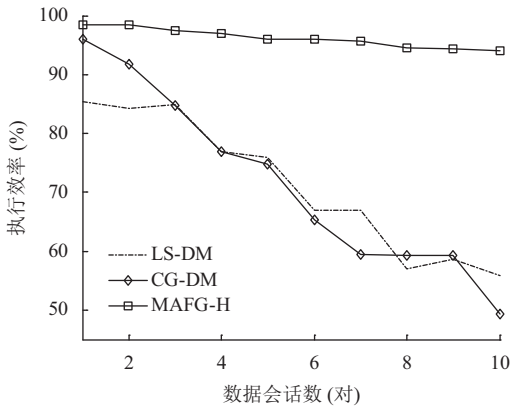

图 4 会话数对执行效率的影响

分析上述结果发现, LS-DM 算法难以对分布式、非线性大数据源的重构, 所以执行效率非常低. 高密度、关联复杂的大数据使得 CG-DM 算法的数据挖掘效率随着会话数和终端数的增加, 明显下降, 冗余数据比例对该算法的性能制约更为明显. MAFG-H 算法建立多维关联空间, 重构异构大数据网络的数据源, 解决非线性和分布式数据问题, 并且基于细粒度为大数据网络的各类用户或参与数据转发的协作终端的融合约束的高度一致性提供保障, 因此始终具有较高的执行效率.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

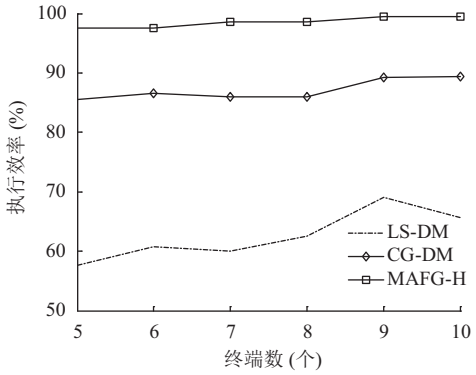

图 5 终端数对执行效率的影响

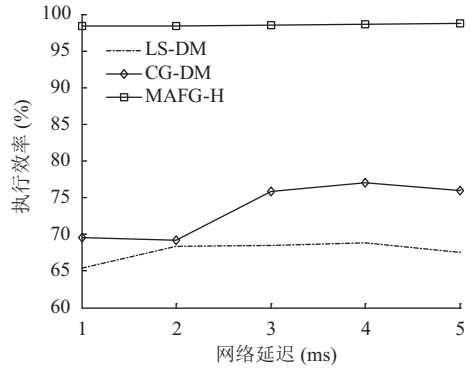

图 6 延迟对执行效率的影响

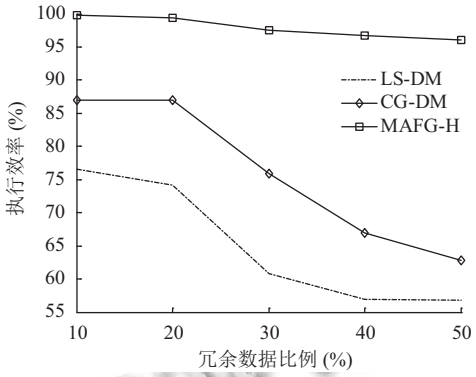

图 7 冗余数据对执行效率的影响

4 结束语

异构大数据网络面临着存储管理难、转发延迟高、处理精度低和执行效率低等问题, 为了保障大数据一致性和归一化异构类数据差异性, 提出了适用于异构大规模网络的多维管理细粒度数据挖掘算法及其机构. 一方面, 建立了异构大数据网络的存储设备和大数据服务器的大数据管理和处理架构. 另一方面, 设计了大数据网络多维关联形式. 最后, 将异构大数据形式定义进行融合, 实现自主判断和决策, 以保持大数据网络的一致性和健壮性为目的, 提出了多维关联细粒度数据挖掘算法. 基于大数据网络的仿真实验和数学分析, 从终端数、会话数、实时性和冗余数据等角度分析对比了执行效率, 证明了所提算法的有效性和可行性.

参考文献

1 彭宇, 庞景月, 刘大同, 等. 大数据: 内涵、技术体系与展望. 电子测量与仪器学报, 2015, 29(4): 469–482.
2 李秀峰, 陈守合, 郭雷风. 大数据时代农业信息服务的技术创新. 中国农业科技导报, 2014, 16(4): 10–15.
3 梁红波. 大数据技术引领物流业智慧营销. 中国流通经济, 2015, 29(2): 85–89.
4 张红英, 王翠森. 大数据时代财务分析领域机遇与挑战. 财会通讯, 2016, (5): 84–85.
5 Chen BY, Yuan H, Li QQ, *et al*. Spatiotemporal data model for network time geographic analysis in the era of big data. International Journal of Geographical Information Science, 2016, 30(6): 1041–1071. [doi: 10.1080/13658816. 2015.1104317]
6 Zhang YF, Chen SM, Wang Q, *et al*. i2 MapReduce: Incremental MapReduce for mining evolving big data. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(7): 1906–1919. [doi: 10.1109/TKDE.2015.2397438]
7 Durocher D. Big data and technical credibility [President's Message]. IEEE Industry Applications Magazine, 2015, 21(2): 4. [doi: 10.1109/MIAS.2014.2375011]
8 Fong S, Wong R, Vasilakos AV. Accelerated PSO swarm search feature selection for data stream mining big data. IEEE Transactions on Services Computing, 2016, 9(1): 33–45.
9 Xu L, Jiang CX, Chen Y, *et al*. A framework for categorizing and applying privacy-preservation techniques in big data mining. Computer, 2016, 49(2): 54 –62. [doi: 10.1109/MC.2016.43]
10 Yan YL, Shyu ML, Zhu QS. Supporting semantic concept retrieval with negative correlations in a multimedia big data mining system. International Journal of Semantic Computing, 2016, 10(2): 247–267. [doi: 10.1142/S1793351X16400092]
11 盛益强, 赵震宇, 廖怡. 用于个性化数据挖掘的粗粒度分布式深度学习. 网络新媒体技术, 2016, 5(6): 1–6.
12 宋小芹, 王莉丽, 张卫星. 基于机会认知的类脑智能数据挖掘机制. 计算机仿真, 2016, 33(11): 375–378. [doi: 10.3969/j.issn.1006-9348.2016.11.082]
13 唐晓东. 基于关联规则映射的生物信息网络多维数据挖掘算法. 计算机应用研究, 2015, 32(6): 1614–1616, 1620.
14 Yuan Q, Ma CG, Lin JY. Fine-grained access control for big data based on CP-ABE in cloud computing. In eds: Wang HZ, Qi HL, Che WX, *et al*. ICYCSEE 2015: Intelligent Computation in Big Data Era. Berlin Heidelberg. Springer. 2015. 344–352.

© 中国科学院软件研究所 http://www.c-s-a.org.cn