P2P 网络中最大频繁项集挖掘算法研究 *

邓忠军[1], 宋 威[2], 郑雪峰[1], 王少杰[3]
(1. 北京科技大学 信息工程学院, 北京 100083; 2. 北方工业大学 信息工程学院, 北京 100144; 3. 国家信息技术安全研究中心, 北京 100094)

摘 要: 为解决 P2P 网络频繁项集挖掘中存在的全体频繁项集数量过多和网络通信开销较大这两个问题,提出了一种在 P2P 网络中挖掘最大频繁项集的算法 P2PMaxSet。首先,该算法只挖掘最大频繁项集,减少了结果的数量;其次,每个节点只需与邻居节点进行结果交互,节省了大量的通信开销;最后,讨论了网络动态变化时算法的调整策略。实验结果表明,算法 P2PMaxSet 具有较高的准确率和较少的通信开销。

关键词: 数据挖掘; P2P 网络; 最大频繁项集; 关联规则
中图分类号: TP311 **文献标志码**: A **文章编号**: 1001-3695(2010)09-3490-03
doi:10.3969/j.issn.1001-3695.2010.09.078

Research on maximal frequent itemset mining algorithm over P2P network

DENG Zhong-jun[1], SONG Wei[2], ZHENG Xue-feng[1], WANG Shao-jie[3]
(1. *School of Information Engineering, University of Science & Technology Beijing, Beijing* 100083, *China*; 2. *College of Information Engineering, North China University of Technology, Beijing* 100144, *China*; 3. *National Research Center for Information Technology Security, Beijing* 100094, *China*)

Abstract: The obstacles mainly lie in numerous frequent itemsets and huge communication cost. To solve the two problems, this paper proposed a maximal itemset mining algorithm P2PMaxSet. Firstly, only considered maximal itemset, which reduced the number of itemsets greatly. Secondly, only interchanged mining results between neighbor nodes, which saved communication cost. Finally, discussed adjust strategies for dynamic environment. Experimental results show P2PMaxSet is not only accurate but also with lower communication cost.

Key words: data mining; P2P network; maximal frequent itemset; association rule

频繁项集(模式)挖掘是数据挖掘研究中的一个重要内容,在关联规则、序列模式等方面有着广泛的应用[1]。随着网络技术的发展,数据趋向于以分布式的方式进行存储。特别是大规模 P2P 网络的兴起[2],为传统的频繁项集挖掘提出了新的挑战。

目前 P2P 网络中数据挖掘的主要工作集中于聚类算法[3,4],而 P2P 网络中频繁项集的挖掘则鲜有研究。Wolff 等人[5]最先提出了 P2P 网络中的关联规则挖掘算法,然而,他们的方法基于多数投票策略直接挖掘关联规则,省去了频繁项集挖掘的过程。与分布式频繁项集挖掘[6,7]不同,P2P 网络下的挖掘往往需要考虑成百上千个分布于不同节点的数据库。因此,在 P2P 网络中挖掘频繁项集就难免要考虑如下两个关键因素:a)挖掘什么样的项集,众所周知,传统的频繁项集挖掘最主要的问题之一就是结果过多,在 P2P 网络中更是如此;b)通信问题,节点间的消息传递会造成大量的通信开销。

为解决这两个问题,本文提出了一种 P2P 网络中最大频繁项集的挖掘算法。首先,只挖掘远少于全体频繁项集的最大频繁项集[8]来解决结果过多的问题。其次,网络节点只需与其直接相邻的邻居节点进行数据交换,从而节省了大量网络通信开销,解决了第二个问题。为适应 P2P 网络的动态性,还讨论了算法的调整策略。实验结果表明,本文所提出的算法是快速和有效的。

1 问题描述

1.1 最大频繁项集

设 $IS=\{i_1, i_2, \cdots, i_m\}$ 为一组由 m 个不同的项(item)组成的集合。集合 $X \subseteq IS$ 称做项集(itemset),将含有 k 个项的项集称为 k-项集。记 TDB 为事务(transaction)T 的集合,这里事务 T 是项集,且 $T \subseteq IS$。

定义 1 若非空事务数据库 TDB 的总事务数为 N,TDB 中包含项集 X 的事务数为 S,则 X 的支持度为 S/N,记为 sup(X)。如果 sup(X) ≥ min_sup,其中 min_sup 为给定的最小支持度阈值,则 X 是频繁项集。

定义 2 对项集 M,若不存在项集 X 使得 $M \subset X$,且 sup(X) ≥ min_sup,则频繁项集 M 是最大频繁项集。

性质 1 Apriori 性质[1]。频繁项集的所有非空子集也是频繁的;非频繁项集的所有超集也是非频繁的。

收稿日期: 2010-02-04; **修回日期**: 2010-03-29 **基金项目**: 国家"863"计划资助项目(2007AA012474);北京市优秀人才培养资助项目(2009D005002000009)

作者简介: 邓忠军(1963-),男,内蒙古赤峰人,高级工程师,博士,主要研究方向为网络安全、数据挖掘(deng.zj@163.com);宋威(1980-),男,讲师,博士,主要研究方向为数据挖掘;郑雪峰(1951-),男,教授,博导,主要研究方向为计算机网络、信息安全;王少杰(1976-),男,工程师,博士,主要研究方向为计算机网络.

1.2　P2P 网络

令 $N_i(1\leqslant i\leqslant n)$ 为 P2P 网络中的节点,N_i 上的数据 $X_i\subseteq X$,称做 N_i 上的局部数据;其中,X 为整个 P2P 网络中所有数据的集合,称做全局数据。局部数据 $X_i(1\leqslant i\leqslant n)$ 与整个 P2P 网络上的全局数据 X 满足如下两个条件:a) $X_1\cup X_2\cup\cdots\cup X_n=X$;b)对 $\forall i\neq j, X_i\cap X_j=\varnothing$。

每个与节点 N_i 直接相连的节点称做 N_i 的邻居节点,记做 $\delta(N_i)$。这样整个 P2P 网络可以看做是一个具有 n 个节点的无向连接图,每个节点都有一个 ID,通过一条边与它的邻居节点相连。

为方便讨论,作如下假定:

a)在任意时刻,每个节点 N_i 的邻居节点的集合 $\delta(N_i)$ 是已知的。

b)网络中的消息传递是可靠的。N_i 向 $N_j(N_j\in\delta(N_i))$ 所传递的消息均能确保到达,除非节点 N_j 已被删除,或者不再是 N_i 的邻居。

本文所提出的 P2PMaxSet 算法旨在高效地从分布于不同节点的局部数据中发现最大频繁项集,最大程度地达到与在单一计算机上对全局数据挖掘最大频繁项集相同的效果。

2　P2P 网络中最大频繁项集挖掘算法

2.1　最大频繁项集挖掘算法

为方便说明,本节给出每个节点内部挖掘最大频繁项集的算法。

算法 1

```
maxSet(C,MFI)
if ((sup(C∪tail(C))≥min_sup) and MFI 中不存在 C∪tail(C)的超集) then
        C∪tail(C)→MFI;
return
for tail(C)中的每个 1-频繁项集 i do
    C_n = C∪i;
    if (sup(C_n)≥min_sup) then
    maxSet(C_n,MFI)
    if (tail(C) == Ø) then
      C→MFI;
```

说明:算法 1 中 tail(C)表示按照某种顺序排在项集 C 后面的 1-频繁项集的集合;MFI 为最大频繁项集集合;C 和 MFI 的初始值均为空集。

2.2　静态网络环境下的最大频繁项集挖掘算法

静态 P2P 网络环境下,最大频繁项集的挖掘如算法 2 所示。因为 P2P 网络中所有节点的地位相同,故只给出某个节点 N_i 的运行情况,其他节点类似。

算法 2　静态网络环境中的 P2PMaxSet 算法

```
(1)用算法 1 得到节点 N_i 内的最大频繁项集 MFI_i;
(2)for N_i 邻居节点集合 δ(N_i)中的每个节点 N_j do
(3)   发送 MFI_i;
(4)for δ(N_i)中每个节点 N_j 发送来的最大频繁项集 MFI_j do
(5)   for MFI_i 中每个项集 i do
(6)     for MFI_j 中每个项集 j do
(7)       if i⊂j then
(8)         MFI_i = MFI_i \i;
(9)       if j⊂i then
(10)        MFI_j = MFI_j \j;
(11)  MFI_i = MFI_i ∪ MFI_j;
(12)  if MFI_i 发生了变化 then
(13)  goto (2);
(14)else 节点 N_i 进入终止状态;
```

2.3　动态网络环境下算法的调整

由于 P2P 网络是动态变化的,本节分如下三种情况对算法 2 进行调整。

1)节点失效

若一个节点 N_j 离开网络,其邻居节点 $\delta(N_j)$ 将会发现这一变化;同样,若某一条边出现故障,则与该边相连的两个节点将检测到这一变化。具体处理步骤如下:a)若 N_i 与某节点 N_j 之间相连的链路需要拆除,则这两者之间的邻居关系就不存在了,N_i 需要把 N_j 从 $\delta(N_i)$ 中删除,N_j 也把 N_i 从 $\delta(N_j)$ 中删除;b)若某节点 N_j 离开网络时,其直接邻居节点 N_i 通过查看 $\delta(N_i)$ 发现;c)若节点 $N_j\in\delta(N_i)$ 离开网络,则 N_j 的邻居节点 $\delta(N_j)$ 成为 N_i 新的邻居节点,并把各自的局部最大频繁项集发给节点 N_i。

2)增加节点

若网络中增加了一个节点 N_j,其处理过程如下:a)N_j 内的数据执行算法 1 得到局部最大频繁项集 MFI_j;b)按照算法 2 执行以下的流程。

3)节点数据发生变化

当网络中某个节点 N_i 的数据发生变化时,其处理过程如下:a)若 N_i 处于非终止状态,则不需要改变方法;b)若 N_i 处于终止状态,则需要把 N_i 重新激活,并执行算法 1 重新计算 MFI_i;c)若 N_i 的邻居节点处于终止状态,则需要激活并且按照增加节点的过程进行相应的处理。

3　性能评测

3.1　模拟器

本文使用了 Internet 拓扑生成器 BRITE(www.cs.bu.edu/brite)来模拟产生 P2P 网络结构,BRITE 可生成图来代表网络拓扑结构,图中边的权重代表通信延时。本文使用 BRITE 中扁平层自治系统(autonomous system,AS)及 Waxman 模型来模拟 P2P 网络。在模拟网络中,两个节点 u 和 v 互连的概率由式(1)计算:

$$P(u,v)=\alpha e^{-d(u,v)/\beta L} \tag{1}$$

其中:$0<\alpha<1$,$0<\beta<1$;$d(u,v)$ 表示连接节点 u 和 v 的边的权重;L 表示网络内两个节点间的最大距离。

在网络拓扑结构的构造阶段使用了增量式的 Waxman 模型,每个步骤中所产生的新节点按照式(1)与两个已存在的节点相连,其中 $\alpha=0.15$,$\beta=0.2$,用该模型所产生的网络的半径与网络中的节点数量满足对数关系。BRITE 中其他参数的设置为 $HS=1\,000$,$LS=100$,这两个参数的含义是网络平面的大小;最大带宽和最小带宽分别为 maxBW = 1 024,minBW = 10。

本文所使用的数据是 T10I4D100K,该数据集可由 FIMI 资源库(http://fimi.cs.helsinki.fi)下载。数据集有 870 个项目,100 000 条事务,每条事务的平均长度为 11。为检验算法性能,把数据平均分布于 P2P 网络的节点中。

3.2　准确率验证

本文把 P2PMaxSet 挖掘得到的最大频繁项集与集中挖掘所得到的结果进行比较,从而来考察算法的准确率,其结果如图 1 所示。这一实验使用了两个参数 α 和 β。其中,α 表示在

集中挖掘时是最大频繁项集,而在 P2PMaxSet 的结果中也是最大频繁项集的百分比;β 表示在 P2PMaxSet 的结果中是最大频繁项集,而在集中挖掘时也是最大频繁项集的百分比。

由图 1 可以看出,参数 α 一直保持着较高的百分比。也就是说,在集中式挖掘下被认为是最大频繁项集,在 P2PMaxSet 挖掘时也往往被认为是最大频繁项集。而 β 参数随节点数量的增加有所下降。也就是说,随着网络中节点数量的增加,在 P2PMaxSet 的结果中是最大频繁项集,而在集中挖掘时也是最大频繁项集的百分比有所下降。

3.3　通信效率验证

本文分别在含有不同节点数量的情况下进行实验,比较了本文提出的算法 P2PMaxSet 与文献[5]提出的 Majority-Rule 的通信总量(图 2)和平均通信量(图 3)。

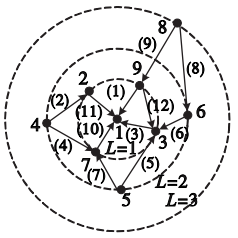

图1　多可用下一跳产生实例

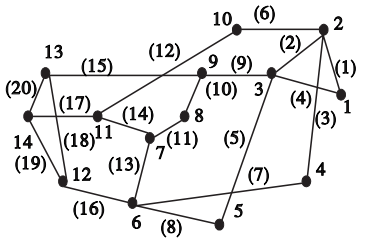

图2　NSFNET拓扑

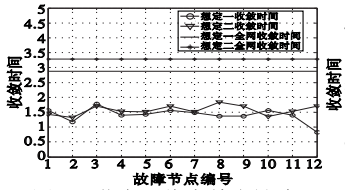

图3　9节点网络中单个链路故障收敛时间

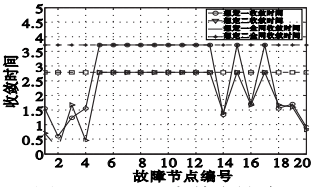

图4　NSFNET中单个链路故障收敛时间

通过图 2 和 3 可以看出,P2PMaxSet 算法的通信总量和单个节点平均通信量基本与节点数量的增长呈线性关系,且均比 Majority-Rule 少一个数量级以上。这主要是因为 P2PMaxSet 算法只挖掘最大频繁项集,而 Majority-Rule 则是要挖掘所有的关联规则。

4　结束语

本文提出了一种 P2P 网络中最大频繁项集的挖掘算法 P2PMaxSet。除了最大频繁项集可明显减少结果数量外,每个节点只与其邻居节点进行挖掘结果的传递和调整,也节省了通信开销。为更好地适应 P2P 网络的特性,还讨论了算法的动态调整策略。实验结果表明,P2PMaxSet 算法是快速和有效的。

参考文献:

[1] 程舒通,徐从富. 关联规则挖掘技术研究进展 [J]. 计算机应用研究, 2009, 26(9):3210-3213.

[2] TAYLOR I J. From P2P to Web services and grids[M]. London: Springer-Verlag, 2005.

[3] KANTERE V, TSOUMAKOS D, SELLIS T K, *et al*. GrouPeer: dynamic clustering of P2P databases [J]. Information Systems, 2009, 34(1): 62-86.

[4] DATTA S, GIANNELLA C R, KARGUPT A, *et al*. Approximate distributed K-means clustering over a peer-to-peer network[J]. IEEE Trans on Knowledge and Data Engineering, 2009, 21(10): 1372-1388.

[5] WOLFF R, SCHUSTER A. Association rule mining in peer-to-peer systems[J]. IEEE Trans on Systems, Man, and Cybernetics, Part B, 2004, 34(6): 2426-2438.

[6] BOUTSINAS B, SIOTOS C, GEROLIMATOS A. distributed mining of association rules based on reducing the support threshold[J]. International Journal on Artificial Intelligence Tools, 2008, 17(6): 1109-1129.

[7] YI Xun, ZHANG Yan-chun. Privacy-preserving distributed association rule mining via semi-trusted mixer[J]. Data & Knowledge Engineering, 2007, 63(2):550-567.

[8] SONG Wei, YANG Bing-ru, XU Zhang-yan. Index-MaxMiner: a new maximal frequent itemset mining algorithm[J]. International Journal on Artificial Intelligence Tools, 2008, 17(2): 303-320.

(上接第 3489 页)较好的定位效果。系统只需利用蓝牙锚节点之间的信息来构造回归模型,不需要人工收集标定样本,同时能将实时收集的锚节点之间的信号样本更新模型参数,实现自适应免标定定位,便于推广和应用。下一步的工作将考虑利用更好的机器学习算法,如半监督学习方法等,将定位区域采集到大量未标定样本加入到训练样本中,进一步提高蓝牙定位精度。

参考文献:

[1] JONATHAN R, GEORG G, HASSAN K, *et al*. A critical evaluation of location based services and their potential[J]. Journal of Location Based Services, 2007, 1(1): 5-45.

[2] HOSSAIN A K M, WEE-SENG S. A comprehensive study of Bluetooth signal parameters for localization[C]//Proc of the 18th International Symposium on Personal, Indoor and Mobile Radio Communications. Washington DC: IEEE Computer Society, 2007: 1-5.

[3] KING T, LEMELSON H, FARBER A, *et al*. BluePos: positioning with Bluetooth[C]//Proc of IEEE International Symposium on Intelligent Signal Processing. Washington DC: IEEE Computer Society, 2009: 55-60.

[4] KOTANEN A, HANNIKAINEN M, LEPPAKOSKI H, *et al*. Experiments on local positioning with Bluetooth[C]//Proc of International Conference on Information Technology: Coding and Computing. Washington DC: IEEE Computer Society, 2003: 297-303.

[5] ZHOU Sheng, POLLARD J K. Position measurement using Bluetooth [J]. IEEE Trans on Consumer Electronics, 2006, 52(2): 555-558.

[6] LIM H, KUNG L C, HOU J C, *et al*. Zero-configuration, robust indoor localization: theory and experimentation[C]// Proc of the 25th IEEE International Conference on Computer Communications. Washington DC: IEEE Computer Society, 2006: 1-12.

[7] FERNANDEZ T M, RODAS J, ESCUDERO C J, *et al*. Bluetooth sensor network positioning system with dynamic calibration [C]// Proc of the 4th International Symposium on Wireless Communication Systems. Washington DC: IEEE Computer Society, 2007: 45-57.

[8] ROMAN R. Kernel-based regression and objective nonlinear measures to assess brain functioning [D]. Scotland: University of Paisley, 2001.

[9] 全勇, 杨杰. 一种基于核 ridge 回归的解耦控制系统[J]. 上海交通大学学报, 2003, 37(9): 1421-1425.

[10] BERNHARD S, TALF H, *et al*. A generalized presenter theorem [C]//Proc of the 14th Annual Conference on Computational Learning Theory and 5th European Conference on Computational Learning Theory. Berlin: Springer-Verlag, 2001.