基于加权动态网络的频繁模式挖掘研究*

肖港松,陈晓云
(福州大学 数学与计算机科学学院,福建 福州 350108)

摘 要:不同时刻的动态网络往往具有不同权重,针对加权动态网络的频繁模式挖掘,提出一种挖掘算法 WGDM,它适用于加权动态社会网络、生物网络等方面的频繁模式挖掘。WGDM 算法利用支持度的反单调性裁剪搜索空间,从而减少冗余候选子图,提高算法效率。通过实验测试了 WGDM 算法的性能,并根据中国实际股票市场网络,利用 WGDM 算法挖掘股票市场网络中有趣的频繁模式。

关键词:加权动态网络;加权图集;频繁子图;图挖掘
中图分类号:TP391 文献标识码:A 文章编号:1674-7720(2011)19-0007-04

Frequent pattern mining research based on weighted dynamic network

Xiao Gangsong, Chen Xiaoyun
(College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China)

Abstract: The dynamic network often has different weights at different times. In order to mining the frequent pattern from the weighted dynamic network, this paper presents a mining frequent subgraph algorithm WGDM bases on weighted graph datasets. It can be applied to mining weighted dynamic network, biology network and so on. Algorithm WGDM utilizes the decrease monotony of support to reduce the search space and the number of candidate subgraph. And make a test for the algorithm WGDM. Finally use the algorithm WGDM for mining the interesting frequent pattern from the Chinese stock market network.

Key words: weighted dynamic network; weighted graph datasets; frequent subgraph; graph mining

近年来,针对社会网络、生物网络等的挖掘研究越来越多(如社区识别、社区关系发现等)[1],尤其是针对犯罪团伙和恐怖分子活动网络的研究,引起了世界各国的重视[2]。实际中,网络往往随时间而变动,即网络是动态网络[3]。挖掘动态网络中的频繁模式,即可以发现变化网络中具有相对"稳定性"的频繁模式,这些模式在动态网络中往往也是比较有趣和重要的,这对研究动态网络很有意义。由于图具有结构关系,可用来表示事物之间复杂的相互作用关系,是基本的数据结构,因此网络可用图来表示,即一个网络可抽象成一个图,对网络的挖掘研究也就转化为对图的挖掘研究。

在实际中,一个动态网络在某个时刻表现出来的整体重要性可能并不一样,这就需要考虑各个时刻网络的不同权重,即考虑加权的动态网络。而挖掘加权动态网络的频繁模式,即是挖掘加权图集的频繁子图。

对图加权主要包括顶点、边和整个图的加权。当前,已经提出一些关于加权图集的频繁子图挖掘算法[4-7],如参考文献[4]、[6]提出的是基于顶点加权的频繁子图挖掘,而参考文献[5]、[7]则是基于边加权的频繁子图挖掘。

网络在某个时刻的重要性可以对整个图赋予不同权重来表示,无需考虑网络内部顶点和边的权重,有时也很难知道顶点和边的权重,针对这种整个图加权的挖掘,关于顶点或边加权的挖掘算法均不适用于这种挖掘。为此本文提出一种适用于整个图加权的频繁模式挖掘算法(简称 WGDM)。

1 相关概念和定义

一些图挖掘和动态网络的基本概念和定义[3-5]:

定义 1(标记图) 一个标记图可表示为一个四元组 $G=(V,E,S,L)$,其中,V 是顶点集合,$E\subseteq V\times V$ 是边集合,S 则是标记集合,$L:V\cup E\rightarrow S$ 是一个函数,用来分配顶点和边的标记。

定义 2(子图同构) 给定两个图 $G=(V,E,S,L)$和图

*基金项目:国家自然科学基金(No.61070020);福建省新世纪优秀人才项目(XSJRC2007-11)

$G'=(V',E',S',L')$，这两个图的子图同构即是一个单射函数 $f:V\to V'$，函数满足：(1) $\forall v\in V, L(v)=L'(f(v))$；(2) $\forall(u,v)\in E;(f(u),(f(v))\in E'$ 且 $L((u,v))=L'(f(u),(f(v))$，也称此单射函数 f 为 G 在 G' 中的一个嵌入。如果存在从 $G\sim G'$ 的子图同构，则称 G 为 G' 的子图，G' 为 G 的超图，记为 $G\subseteq G'$。

定义 3 (动态网络) 在用图 $G=(V,E)$ 表示的网络中，顶点集 V 和边集 E 随时间变化而变化的网络称为动态网络。

下面给出本文对加权图集、加权动态网络、加权图集中子图的支持度和频繁子图的定义。

定义 4 (加权图集) 给定一个图的集合 $D=\{G_1,G_2,\cdots\cdots,G_n\}$，对 D 中的图 $G_1,G_2,\cdots\cdots,G_n$ 分别赋予权重 $w_1,w_2,\cdots\cdots,w_n$(权重为非负实数)，则称 D 为加权图集。

定义 5(加权动态网络) 加权动态网络即是对不同时刻的网络赋予权重的动态网络，权重为一非负实数，由该时刻网络的重要性来决定权重大小。

定义 6 (支持度) 给定加权图集 $D=\{G_1,G_2,\cdots\cdots,G_n\}$ 和图模式 g，如果图集 D 中包含图 g 的图为 $G_{i1},G_{i2},\cdots\cdots,G_{in}$，各图对应的权重分别为 $w_{i1},w_{i2},\cdots\cdots,w_{ik}$，则图 g 的绝对支持度为：

$$\sup(g,D)=\sum_{i=1}^{\infty}w_{ij} \tag{1}$$

定义 7(频繁子图) 给定加权图集 $D=\{G_1,G_2,\cdots\cdots,G_n\}$ 和一个实数阈值 min_sup，如果子图 g 在加权图集 D 中的支持度 $\sup(g,D)\geqslant$min_sup，则称该子图 g 为频繁子图。

2 挖掘加权图集中的频繁子图

2.1 频繁子图挖掘

(1) 频繁子图挖掘的难点之一在于会产生数量庞大的候选子图，使得搜索空间巨大。本文提出的 WGDM 算法具有如下性质，从而可利用该性质来裁剪搜索空间。

性质：给定加权图集 $D=\{G_1,G_2,\cdots\cdots,G_n\}$，则一个图模式 g 的支持度是它所有超图支持度的上界。

证明：设图 g′是图 g 的任意超图、加权图集 D 中包含图 g′的图分别为 $G_1,G_2,\cdots\cdots,G_k$，对应的权重大小分别为 $w_1,w_2,\cdots\cdots,w_k$。由定义 7 可知，图 g′的绝对支持度为 $\sup(g',D)=\sum_{i=1}^{k}w_i$；又由图 g 是图 g′的子图可知，图集 D 中包含图 g 的图至少包括 $G_1,G_2,\cdots\cdots,G_k$ (这里可不妨设包含图 g 为 $G_1,G_2,\cdots\cdots,G_k$ 和 $G_{k+1},G_{k+2},\cdots\cdots,G_m$)，其中图 $G_{k+1},G_{k+2},\cdots\cdots,G_m$ 的权重分别为 $w_{k+1},w_{k+2},\cdots\cdots,w_m$，所以图 g 绝对支持度为：$\sup(g,D)=\sum_{i=1}^{k}w_i+\sum_{j=k+1}^{m}w_j\geqslant\sup(g',D)$。其中，当图集 D 中包含图 g 和 g′的图相同时，等号成立。同理可证明图 g 的相对支持度也是不小于图 g′的相对支持度，所以图 g 和图 g′在 D 中的支持度满足：$\sup(g')\leqslant\sup(g)$。

由 WGDM 的性质可得，如果图 g 是非频繁子图，则其所有的超图也不是频繁子图，即可裁减掉图 g 的所有超图，如图 1 所示。

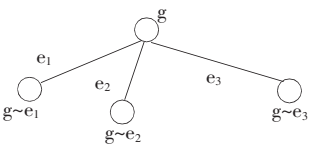


图 1 剪枝

子图 g 可扩展的超图包括 $g\sim e_1$、$g\sim e_2$、$g\sim e_3$。首先计算子图 g 的支持度 support(g)，若小于最小支持度，则剪掉 g 的所有超图。

(2) 频繁子图挖掘的另外一个难点在于子图同构检测[8-9]。参考文献[9]提出的 GASTON 算法利用一种内嵌列表(Embedding List)记录了顶点和边在图集中的具体位置，在子图扩展时可以快速地从内嵌列表中找出可扩展的顶点和边以及进行同构检测，较好地解决了子图同构检测问题；而且该算法将一个复杂的图挖掘问题分割成三个比较简单的子问题，即先列举出路径(Path)、再列举由路径扩展出的树 (Non-cyclic Tree)、最后列举由路径或树扩展后的具有循环的图(Cyclic Graph)。

GASTON 算法虽然不能挖掘加权图集的频繁子图，不过其同构检测的方法与分解成三个子问题的策略很有意义。本文采用其策略方法来进行同构检测，并将加权图集挖掘也转为挖掘路径、树和循环图的三个步骤。

2.2 算法描述

首先计算 WGDM 算法加权图集中子图的支持度，其计算步骤如下：

算法 1 计算子图支持度 $\sup(g,D)$

输入：加权图集 D，子图 g，内嵌列表。

输出：子图 g 的支持度 $\sup(g,D)$。

(1)初始化 $\sup(g,D)=0$；

(2)利用内嵌列表(Embedding List)找出 D 中包含子图 g 的所有图 $G_{i1},G_{i2},\cdots\cdots,G_{ik}$。

(3) 找出 $G_{i1},G_{i2},\cdots\cdots,G_{ik}$ 各图对应的权重：$w_{i1},w_{i2},\cdots\cdots,w_{ik}$。

(4)For $j=1,2,\cdots,k$ do

$\sup(g,D)\leftarrow\sup(g,D)+w_{ij}$

(5)输出子图 g 支持度 $\sup(g,D)$。

计算加权子图支持度的实例如图 2 所示。图中，动态网络在 t_1、t_2、t_3 时刻形成的无向网络图(本文针对的是顶点和边均有标记的无向加权动态网络图)，对应的权重分别为 w_1、w_2、w_3。假设权重 $w_1=1$、$w_2=2$、$w_3=3$，从图 2 可看出，路径图 $P(v_1\sim v_2\sim v_3)$ 只出现在 t_1 和 t_3 时刻的网络图中，所以其绝对支持度为：

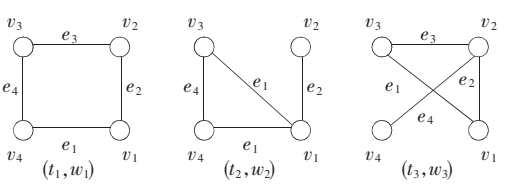


图 2 加权动态网络

$\sup(P)=w_1+w_3=1+3=4$

结合 GASTON 算法[9]的策略方法，下面给出挖掘加权图集中频繁子图的算法步骤：

算法 2 挖掘频繁路径(Path)

输入：加权图集 D，图编码，内嵌列表，最小支持度 min_sup，路径 P。

输出：频繁路径(Path)。

(1) 事先由算法 1 计算加权图集中所有顶点和边的支持度，删除小于 min_sup 的顶点和边。

(2)由算法 1 计算出路径 P 的支持度，如果其支持度 support(P)<min_sup，则停止扩展，剪掉其所有超图；否则从内嵌列表选取可扩展的边 l，构造新图 $g\leftarrow l+P$。

(3)如果新图 g 还是路径，则转至步骤(2)。

(4)如果新图 g 是树则转至算法 3。

(5)如果新图 g 是具有循环的图则转至算法 4。

算法 3 挖掘频繁树(Tree)

输入：加权图集 D，图编码，内嵌列表，最小支持度 min_sup，树 T。

输出：频繁树。

(1) 由算法 1 计算出树 T 的支持度，如果其支持度 support(G)<min_sup，则停止扩展，剪掉其所有超图；否则从内嵌列表选取可扩展的边 l，构造新图 $g\leftarrow l+T$。

(2)如果新图 g 还是树，则转至步骤(1)。

(3)如果新图 g 是具有循环的图则转至算法 4。

算法 4 挖掘频繁循环图(Cyclic Graph)

输入：加权图集 D，图编码，内嵌列表，最小支持度 min_sup，图 G。

输出：频繁图。

(1) 由算法 1 计算出图 G 的支持度，如果其支持度 support(G)<min_sup，则停止扩展，剪掉其所有超图。

(2) 否则从内嵌列表选取可扩展的边 l，构造新图 $g\leftarrow l+G$，转至步骤(1)。

(3)输出所有频繁图。

从算法 2~算法 4，先找出频繁路径，如果该路径扩展成树，则转至找频繁树；如果扩展成图，则转至寻找频繁循环图。在寻找频繁树时，如果树扩展成循环图则转至寻找频繁循环图；最后找出频繁循环图。其实，路径和树都是无循环的特殊的图，所以最后输出的加权频繁子图也包括路径和树。

3 实验

3.1 算法性能测试

本文测试使用的数据集是有关分子生物活性信息的真实数据集 NCI-H23，这个数据集可以从以下网址获得：http://www.cs.ucsb.edu/~xyan/dataset.htm。

NCI-H23 数据集包括具有活性和无活性两种类别的图集，其中顶点有 60 多种标记，边有 2 种标记。假设无活性的图权重为 1，而具有活性的图权重为 2。本文选取 200 个具有活性和 200 个无活性的图，然后组成了一个具有 400 个图的加权图集。

算法测试用的PC机使用Intel Pentium(R)2.6 GHz CPU和512 MB 的内存，操作系统为 Red Hat Linux，算法使用 C++语言实现，并用 g++编译。实验结果如图 3 所示。

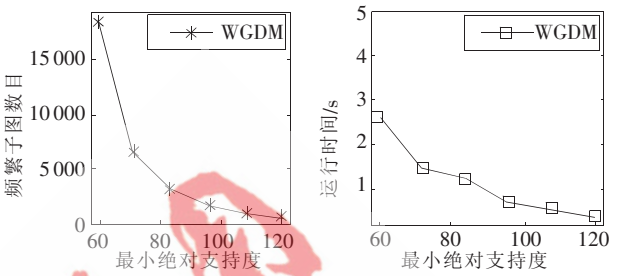


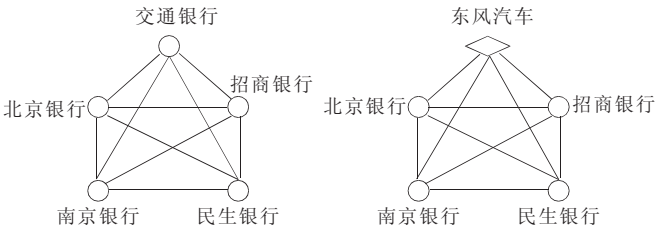
图 3 性能测试

从图 3 可以看出，当支持度比较小时，算法挖出到的频繁子图数目非常大，如在最小绝对支持度为 60 时，可挖掘到 18 673 个频繁子图，这比最小绝对支持度为 120 时挖掘到的 675 个频繁子图多了 27 倍；运行时间则是随着最小支持度的增加而减少，在最小绝对支持度为 96 时，运行时间只需 0.69 s，总体上算法具有良好的效率。

3.2 股票市场网络的挖掘应用

结合中国股票市场，利用本文提出的算法挖掘股票市场网络中的频繁模式。一般股票价格会随着时间变化，不同时段股票跌幅或涨幅不一样。本文抽取 20 支股票，这些股票来自电子行业、啤酒行业、金融银行等领域，然后以一个季度为一个时段，统计这些股票在 2010 年四个季度里的涨跌情况，其中在每个季度里，分四种情况划分成四种网络：涨幅超过 40%的股票网络、涨幅在 40%以内的股票网络、跌幅在 20%以内的股票网络以及跌幅超过 20%的股票网络。股票网络中，顶点表示股票，不同股票，标记也不同，而股票间的关联就是边，不同股票的边标记也不同，同一个网络中的任意两支股票均有一条具有标记的边相连。在实际中，对于涨幅比较高或者跌幅比较大的情况应给予额外关注，为此对涨幅超过 40%和跌幅超过 20%的网络加大权重，本文设定这两种网络权重为 2，而其他两种网络则给予 1 的权重。总共得到 9 个网络图组成的图集，其中有 3 个网络图属于涨幅超过 40%或者跌幅超过 20%，给予的权重为 2，其余 6 个网络图权重为 1。利用本文 WGDM 算法挖掘这个加权动态网络图集的频繁模式，而用 GASTON 算法挖掘无加权动态网络图集(即所有图权重都为 1)，其中设定绝对最小绝对支持度 min_sup 为 4 时，可以发现两种具有 5 个顶点的频繁模式如图 4 所示。

实际中，相同行业的公司、企业的发展趋势比较有相同之处，其股价也较有可能同涨同跌。如图 4 所示，本文挖掘出的频繁模式，都是由银行组成，而 GASTON 算



(a)本文算法(WGDM) (b)GASTON 算法

图 4 挖掘的频繁模式对比

法挖掘出的频繁模式由银行和汽车两个不同行业组成。所以本文算法的挖掘结果,与实际比较吻合,进一步验证了本文算法的有效性。

挖掘加权动态网络的频繁子图困难在于产生的候选子图数量过多,而且子图同构检测问题也会影响算法的效率。对此,本文算法利用支持度的反单调性对搜索空间进行裁剪,并采用参考文献[7]的策略将挖掘图划分成挖掘路径、树和循环图的三个子问题,减少了候选子图数量和子图同构检测次数,提高了算法效率。而且将算法应用于实际的股票市场网络,挖掘结果也验证了本文算法的有效性。本文算法还可进一步拓展应用到其他网络的频繁模式挖掘。

参考文献

[1] RADICCHI F, CASTELLANO C, CECCONI F, et al. Defining and identifying communities in networks[J]. PNAS, 2004, 101(9): 2658-2663.

[2] XU J J, CHEN H C. CrimeNet explorer: a framework for criminal network knowledge discovery [J]. ACM Transactions on Information Systems, 2005, 23(2).

[3] BERGER-W T Y, SAIA J. A frameworkfor analysis of dynamic social networks [C]. KDD'06. Philadelphia: [s.n.], 2006: 523-528.

[4] 耿汝年,董祥军,须文波.基于全局图遍历的加权频繁模式挖掘算法[J].计算机集成制造系统,2008,14(6):1220-1229.

[5] 王映龙,杨珺,周法国,等.加权最大频繁子图挖掘算法的研究[J].计算机工程与应用,2009,45(20):31-34.

[6] 封军,郑诚,郑晓波,等.基于加权有向图的权频繁模式挖掘算法[J].微型机与应用,2010,29(20):4-7.

[7] Jiang Chuntao, COENEN F, ZITO M. Frequent sub-graph minjing on edge weighted graphs[C]. DaWak'10 Proceedings of the 12th international conference on Data Warehousing and knowledge discovery, Spinger-Verlag, 2010:77-88.

[8] 高琳,覃桂敏,周晓峰.图数据库中频繁模式挖掘算法研究综述[J].电子学报,2008,36(8):1603-1609.

[9] NIJSSEN S, KOK J N. Aquick start in frequent structure mining can make a difference [C]. Proceeding of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004). Seattle, WA, USA: Springer-Verlag, 2004: 4571-4577.

(收稿日期:2011-06-08)

作者简介:

肖港松,男,1986 年生,硕士研究生,主要研究方向:数据挖掘,模式识别。

陈晓云,1970 年生,女,博士,副教授,主要研究方向:数据挖掘,机器学习,模式识别。