

# 基于社交网络的图数据挖掘应用研究

李桃陶,周 斌,王忠振

(国防科学技术大学 计算机学院,湖南 长沙 410073)

**摘要:** 社交网络数据的高度复杂性给数据挖掘研究带来了巨大的挑战,而社交网络数据挖掘更注重实体之间相互关联的特点,使得图数据挖掘技术的研究与应用逐渐成为该领域的热点。传统数据挖掘,如聚类、分类、频繁模式挖掘等技术逐渐拓展到图数据挖掘领域。文中首先介绍了现阶段图数据挖掘算法(其中包括图查询、图聚类、图分类和图的频繁子图挖掘)的研究内容和存在的问题;其次介绍了图形数据库研究现状,以及对比了主流图形数据库管理系统的优劣;最后介绍了图挖掘技术在社交网络中的应用。

**关键词:** 图挖掘;图查询;图分类;图聚类;图形数据库;社交网络

中图分类号: TP39

文献标识码: A

文章编号: 1673-629X(2014)10-0006-06

doi: 10.3969/j.issn.1673-629X.2014.10.002

## Research on Graph Data Mining Application Based on Social Network

LI Tao-tao, ZHOU Bin, WANG Zhong-zhen

(College of Computer, National University of Defense Technology, Changsha 410073, China)

**Abstract:** The high complexity of the social network data brings a huge challenge for data mining research. The social network data mining pays more attention to the relationship between entities, and the research and application of graph mining technology is gradually becoming a hotspot in the field. Traditional data mining, such as clustering, classification, frequent pattern mining technology has gradually extended to graph mining field. In this paper, introduce the present the research content of graph data mining algorithm (including graph query, graph clustering, graph classification and frequent sub-graph mining) and the existing problems first. Second introduce the research status of graph database, and compare the advantages and disadvantages of the mainstream graphics database management system. Finally introduce the application of graph mining technology in social network.

**Key words:** graph mining; graph query; graph classification; graph clustering; graph database; social network

## 0 引言

社交网络的发展和普遍应用,使得大量研究人员转向对网络关联关系的研究,如在社交网络用户的关注、转发、评论、提及等关联关系中隐藏着许多有价值的知识,以图数据结构来建立社交网络中不同类型的复杂关系,挖掘这些数据中隐藏的知识在近几年已经成为研究的热点,在社交网络中用户影响力分析、企业广告营销、社区发现等方面都有应用的前景。文中通过图数据理论研究、图数据库研究和在社交网络的应用三个方面对图数据挖掘技术进行探讨。

## 1 图数据理论研究

本节给出图数据相关定义,并对目前国内外基于图挖掘算法所展开的工作进行总结,介绍了图数据挖

掘的研究现状与主要研究问题。

### 1.1 图数据定义

图由一系列的点以及连接点的边构成。一个图通常表示为  $G(V, E)$ , 其中  $V$  表示顶点的集合,称为图  $G$  的顶点集;  $E$  是集合  $V \times V$  的一个子集,即边的集合,称为图  $G$  的边集。在社交网络中可用图的顶点集来代表用户节点,而图的边集,多用于表示用户之间的关系或相互关联。

无向图: 一个图  $G(V, E)$ ,  $E$  中的边通常由两个顶点表示,若两个顶点是无序的,则该图是无向图。若顶点有先后顺序,则该图是有向图,边通常表示为带箭头的连线。

确定(标号)图: 可以被表示为  $G = (V, E, \Sigma V, \Sigma E, I)$ , 其中  $V$  是顶点集,  $E$  是边集,  $E \subseteq V \times V$ ,  $\Sigma V$  和  $\Sigma E$  分

收稿日期: 2013-12-06

修回日期: 2014-03-13

网络出版时间: 2014-07-28

基金项目: 国家“973”重点基础研究发展计划项目(2013CB329602)

作者简介: 李桃陶(1983-),男,广东广宁人,硕士研究生,研究方向为数据挖掘;周 斌,硕士研究生导师,研究方向为数据挖掘。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20140728.1228.041.html>

别是顶点和边的标号集,  $I$  是标号的映射函数。若节点上的标号互不相同, 则称为唯一标号图。

不确定图: 可以被表示为  $G=(V,E,\Sigma V,\Sigma E,I,P)$ , 其中  $V$  是顶点集,  $E$  是边集,  $E\subseteq V\times V$ ,  $\Sigma V$  和  $\Sigma E$  分别是顶点和边的标号集,  $I$  是标号的映射函数,  $P$  是边的存在可能性函数, 取值范围为  $(0,1]$ 。确定图可看成边存在可能性为 1 的不确定图。

子图: 对于两个图  $G=(V,E,\Sigma V,\Sigma E,I,P)$ 、 $G'=(V',E',\Sigma V',\Sigma E',I',P')$ , 如果  $V'\subseteq V, E'\subseteq E$ , 就称  $G'$  是  $G$  的子图。若  $E'\subseteq E, V'=V$ , 则称  $G'$  是  $G$  的生成子图。

图同构: 一个标号图与另一个图同构, 当且仅当满足以下条件的映射函数成立  $f:V(G)\rightarrow V(G')$ :

- (1)  $\forall u\in V, (I(u)=I'(f(u)))$ ;
- (2)  $\forall u\in V, v\in V, ((u,v)\in E\Rightarrow(f(u), (v))\in E')$ ;
- (3)  $\forall (u,v)\in E, (I(u,v)=I'(f(u), f(v)))$ 。

由上述公式可以得出: 如果两个图在拓扑结构上是相同的, 则称两个图是同构的。

1.2 图数据挖掘算法

图数据挖掘技术问题得到了广泛的研究<sup>[1]</sup>, 文中根据研究的进展主要将图数据挖掘算法归纳为图查询、图聚类、图分类和图的频繁子图挖掘四类算法, 总结了目前这几类算法的特点和其中存在的问题。

1.2.1 图查询

图查询是在现有查询图数据库中找出与输入图模式(检索图)相同或相似的图模式, 针对大规模的图数据库, 图查询技术需要解决三个问题: 一是较大数据量情况下的图查询效率低; 二是图查询结果是否与输入图的模式图相同或满足一定阈值的相似度, 即是查询如何保证准确度; 三是查询的结果能够覆盖到整个查询图。为了对图查询进行有效的分析, 许多研究者提出了各种类型的查询, 主要包括两类:

(1) 通过对大图查询, 返回图中重要节点或节点间的特性, 查询的方式包括: 可达性查询, 用以判断两个节点间是否存在一条路径; 距离查询, 用以返回两个节点间的最短路径; 关键字查询, 用以发现节点之间关系与特殊关键字相关的群体。研究的方法主要分为索引构建方式<sup>[2-3]</sup>、基于大图可达性优化<sup>[4-5]</sup>和基于最短路径优化<sup>[6-7]</sup>为索引加速的方式。GIndex<sup>[3]</sup>中首次提出了差异频繁子图作为索引结构, 支持相似节点查询。算法用动态支持度和模式区分等方式对频繁子图进行过滤, 过滤后剩余的频繁子图集合作为数据特征。

(2) 图与图的查询, 包括图查询和模式匹配查询。其中图查询注重图与图数据结构上的匹配, 而模式匹配查询除结构上的匹配外, 还要求语义的匹配, 比图查

询更灵活。大多数图应用中最关键的问题是如何有效地进行图查询。L. Zou<sup>[8]</sup>等人针对模糊模式查询问题提出了 Distance-Join 算法框架; H. Shang<sup>[9]</sup>等人提出了 QuickSI (Quick Subgraph Isomorphism) 算法将子图查询问题转化为序列的验证问题。针对如何提高图查询效率问题, 从降低图空间复杂度方面, C. Chen<sup>[10]</sup>等人提出了 SUMMARIZE-MINE 算法对查询图空间的数据进行压缩; 从缩减查询时间方面, Z. Zeng<sup>[11]</sup>等人提出了 Comparing Star 算法在查询前先计算子图的距离。针对不确定图的查询问题, Z. Zou<sup>[12]</sup>等人用分支界定的算法找出图中  $k$ -top 的最大集合, 并通过修剪规则减小搜索节点数量和降低内存消耗。

总之, 相比关系查询, 图查询具有图数据种类繁多, 查询过程子图同构测试不能保证查询性能、可扩展性以及数据结构复杂, 操作困难等特点, 因此图查询是图数据操作的难点之一。

1.2.2 图聚类

在大图上的聚类主要是把图划分为若干子图, 使两个子图间的边的权值和尽可能小, 子图内的边的权值和尽可能大。图聚类在分析和可视化大型图应用中是非常有效的技术, 并且在社交网络社区侦测有着良好的应用前景。图聚类目的是基于不同的准则将顶点划分成不同的类别, 现有的大多数图聚类方法主要分为以下三种:

(1) 基于顶点拓扑结构的关联度聚类。这些聚类方法包括基于规范化分割<sup>[13]</sup>, 模块化<sup>[14]</sup>, 结构紧密度<sup>[15]</sup>或随机游走<sup>[16]</sup>等方式, 吴焯<sup>[17]</sup>等结合信息论中最小长度原则, 基于遗传算法, 提出一种高效的属性图聚类方法 GA-AGC, 并通过实验表明该算法具有较高的聚类质量和接近线性的复杂度。该类算法主要关注图的顶点拓扑结构而忽视了顶点的属性在聚类过程中的作用, 没有考虑顶点之间属性的相似性, 使同一类中的顶点的属性可能差别很大。

(2) 基于顶点相关的属性值的相似度聚类。属性值相同的节点聚集成一类, 其中 E. Tiakas 等<sup>[18]</sup>提出通过传递节点相似度来聚类的方法。该算法首先应用特征值来完成所有图的节点的排序, 其次选择特征节点并计算这些节点与其他所有节点的相似度, 并将相似度保留到相似列表中, 第三步从相似度列表中, 构建第一个有相似度达到一定阈值的集群, 重复二、三步直到所有节点聚类完成。该算法可用于加权或无加权的无向图, 且时间和空间复杂度都较低, 因此能适用于大图。但是该类方法由于在聚类过程中没有考虑图中顶点的拓扑结构, 使同一类中的顶点之间连接的边比较稀疏, 而有密集边顶点集合有可能会被划分为不同类中。

(3) 基于顶点拓扑结构关联度和属性相似度的聚类。考虑上述两种类型算法的优劣, M. Ester<sup>[19]</sup> 等提出了一种同时考虑关系数据和属性数据的 NetScan 算法。NetScan 方法试图将一个具有关系和属性的图划分成互相连接且具有相似属性的子图。该算法需要指定需要聚类的类的数目, 而在聚类前难以确定合适的类型数目。针对 NetScan 存在的问题, Moser 等<sup>[20]</sup> 提出了不需要用户指定聚类数量的 JointClust 方法, 该方法采用层次聚类算法, 比较依赖类中心的初始选择。

总体来看, 目前大多数图聚类算法中, 仅注重聚类的质量而不注重计算的效率, 算法复杂度较高, 因此在处理社交网络大规模数据的聚类中, 提高聚类算法计算效率就成了需要解决的问题。

### 1.2.3 图分类

图分类是挖掘图的特征子图来构建分类模型, 再通过分类模型来对图中的子图结构进行分类。根据图集中节点有无标签, 可将图分类划分为有监督分类<sup>[21]</sup>和无监督分类<sup>[22]</sup>。但在真实的图数据中可能同时存在有标签和无标签的节点, 或者遇到有标签的节点数据丢失的情况。针对以上问题, Jerome 提出了半监督分类<sup>[23]</sup>。文中从构建分类模型的方式将图分类方法分成两类:

(1) 用频繁子图方法构建<sup>[24]</sup>。先采用频繁子图挖掘算法在图集中挖掘出频繁子图, 并从根据频繁子图中的某些特征(子图的拓扑结构或属性特征)来构建分类模型。该方法的优点是能以结构特征对图数据分类, 进而获得整个图的结构特征, 能够适用于各种类型的图数据, 特别是对未知领域的图数据的挖掘。缺点是模式图的规模和支持度难以确定, 模式图规模较小或支持度较低时, 产生的分类类型较多, 分类模型难以构造, 并且挖掘时间较长, 分类的结果没有实际意义; 模式图规模较大或支持度较高时产生的分类类型较少, 构建的分类模型分类的结果难以确保准确率。

(2) 用图核函数构建。先设计一个图核函数, 再通过比较子图与核函数的相似度来进行分类。图核函数通常是以路径、子树或者子图为基础, 如 P. Mahé 等<sup>[25]</sup> 提出了以子树为图核的图分类方法。此外, T. Kudo 等<sup>[26]</sup> 提出一种基于 Boosting 的图分类算法, 同时 Saigo<sup>[21]</sup> 提出基于 Boosting 的改进算法 GBoost, 分类的效果要好于频繁子图挖掘。用图核函数来构建分类模型能够避免使用频繁子图不好控制模式图规模和支持度的问题, 在特定领域的图分类上效果会好于频繁子图。

但是对于多领域的图数据分类, 设计图核函数比较困难, 并且设计的图核函数可能跟需要分类的图数据联系不紧密, 分类产生的结果也没实际的意义。

### 1.2.4 图的频繁子图挖掘

图的频繁模式挖掘是从图数据集中寻找出现次数不少于最小支持度的子图结构。候选模式结构, 通过某种方式得到的图, 但还未检验其频数是否大于等于阈值。常见的算法按照遍历的策略主要分为三类: 基于贪心搜索的算法, 如 SUBDUE<sup>[27]</sup>、SubdueCL<sup>[28]</sup>、GBI<sup>[29]</sup> 等, 其中著名的 SUBDUE 基于最小描述长度原则(Minimum Description Length, MDL), 通过用一个顶点替换模式来找出那些可以有效压缩原始输入数据的模式; 基于广度优先遍历的算法, 其中包括 AGM(Apriori-based Graph Mining)<sup>[30]</sup>、FSG(Frequent Subgraph Graph)<sup>[31]</sup>、路径连接算法<sup>[32]</sup> 等; 基于深度优先遍历的算法, 包括 gSpan<sup>[33]</sup>、CloseGraph<sup>[34]</sup>、Gaston<sup>[35]</sup> 等。

此外, 在处理较大的图规模时, 频繁子图挖掘会产生较多的子图集, 并且子图集的规模较大, 处理的效率较低, 针对这个问题部分研究者提出了最大频繁子图挖掘算法。典型的算法有 Spin<sup>[36]</sup> 和 MARGIN<sup>[37]</sup>。对比 MARGIN 与 Spin 算法, 由于 MARGIN 在挖掘中需要存储子图模式, 在存储消耗上要比 Spin 多, 但在处理效率上, MARGIN 在网页数据的处理效率会更高。李继腾等<sup>[38]</sup> 提出一种最大频繁子图挖掘算法(MFME), 该算法将图中较为集中的边通过映射的方式形成对应的边表, 并在边表上面进行子图挖掘, 对于只考虑图结构的挖掘能够有效提高效率。

从上述频繁子图挖掘文献的实验结果来看, 广度优先策略比深度优先策略的要慢。在输入为单图模式下, SUBDUE<sup>[27]</sup> 相比其他算法较快, 而输入为合集时 Gaston<sup>[35]</sup> 会比较快。算法在一些应用中不能在有限时间内挖掘所有子图模式。

## 2 图形数据库研究

随着社交网络的兴起, 传统的关系数据库管理系统暴露出了一些问题, 主要是数据建模中的缺陷, 以及在处理海量数据和多服务器上水平伸缩的限制。其中, 图形数据库就是其中一种非关系数据库。

图形数据库<sup>[39]</sup>(graph database) 是使用图结构与节点、边、属性来表示和存储数据的数据库。节点代表实体, 如人、企业、账户或任何其他项目您可能想要跟踪; 属性是节点的相关信息, 如人是一个节点, 就有姓名、年龄、身高等属性; 边是连接两个节点的关联关系, 边上也有相关的属性。

当前, 图形数据库研究与应用已经得到了重视。维基百科<sup>[39]</sup> 列出了一些比较有名的图形数据库项目。在图形数据库与关系数据库对比方面的研究, Chad 等<sup>[40]</sup> 通过比较 MySQL 和 Neo4j 这两种数据库在处理 DAG(Directed Acyclic Graph) 数据结构的效率, 来说明



关系数据库和图形数据库之间的优劣,其中在字符串查询和结构查询上面图形数据库有着较高的响应速度;Florian 等<sup>[41]</sup>比较 Neo4j 图形数据库中的几种查询语言和在 MySQL 使用 JPA 查询的效率,比较之下尤其是在处理社交网页大量交互数据时,Neo4j 完全可以替代关系数据库。在图形数据库特性方面的研究,Renzo<sup>[42]</sup>通过比较六种图形数据库(Neo4j, Hyper-Graph-DB, DEX, InfoGrid, Sones, VertexDB),显示各数据库有支持不同的图结构、查询 API 接口、查询语言和基本的完整性约束;N. Mart 等<sup>[43]</sup>介绍了高效的图形数据库 DEX,数据库基于位图的结构,通过查询优化能够有效处理数十亿的节点的图数据。在提高图形数据库性能方面的研究,Yiping 等<sup>[44]</sup>提出了一种 CGS(Correlated Graph Search)算法,该算法采用皮尔森相关系数方法来加速图数据挖掘,提高了查询速度并降低了内存消耗,更重要的是算法在图查询过程中稳定性较高;Pri-ma 等<sup>[45]</sup>提出一种在线社交网络图数据分割模型,用于降低访问图形数据库的延迟。

目前开发和应用的图数据库管理系统主要有以下几种:

Neo4j 具有 ACID 事务管理,高可用性,可伸缩到亿级数量的节点和关系,高速遍历等特点,同时提供 Java、python、php 等多种语言开发接口;DEX 由 Java 和 C++ 语言开发,主要特点是在查询和检索大型网络数据时性能较高,能以较低的存储消耗来处理数十亿级的数据量;HyperGraphDB 的图形模型被称为直接式超图形,超图形允许其一条边线指向两个以上的节点,而 HyperGraphDB 在此基础上允许一条边线指向其他边线,相比其他图形类数据库能够处理更多复杂结构;Titan 是分布式图数据库,采用可插拔式的存储构架,可以在 Cassandra、HBase 或 BerkeleyDB 数据库上构建图数据库集群,具备良好的可扩展性和高效的处理能力;AllegroGraph 采用 RDF(即资源描述框架)三元组作为边线来处理,原本设计意图是创建以 RDF 为中心的语义网络应用程序,并支持 SPARQL、RDFS++ 以及 Prolog 等由包括 Java 程序在内的各类客户应用程序。还有众多图数据库,开发者可根据各自需求选择:处理属性图可以选择 Neo4j 和 DEX,其中 Neo4j 有较好的易用性和丰富的开发资源,而且 DEX 免费版本有处理节点限制;需要构建大规模图数据库集群并具备较好可扩展能力可选择 Titan;处理超图数据的可以选择 HyperGraphDB。

总之,图形数据库在处理大规模社交网络数据集中,能体现出较低的延迟和更高的效率,而且具备良好的扩展性,目前在语义网和 RDF、GIS、基因分析、社交网络数据建模、深度推荐算法等领域都有使用。

### 3 图数据挖掘在社交网络的应用

截至 2013 年 6 月底,我国网民规模达 5.91 亿人<sup>[46]</sup>,随着越来越多用户的加入,数据量呈指数增长,对社交网络的分析将面临很大的困难。Aggarwal 等在文献[47]中介绍了一些社交网络分析应用场景,以及在中心地位分析、角色分析、网络建模等方面研究中存在的问题。针对在社交网络分析面临数据增量较大和图形数据库更适于存储和处理社交网络关系的特点,R. Soussi 等<sup>[48]</sup>提出了一种从关系数据库数据转换成图形数据库数据的机制,并从图形数据库中抽取社交网络关系的方法。S. Kadge 等<sup>[49]</sup>提出了基于图预测社交网站,预测是基于有向加权的社交图,方法是构建用户行为的社交网络图,用图挖掘技术预测未来的社交行为,并与 Apriori 和 F-Tree 两种算法的预测效率进行了对比。J. Cao 等<sup>[50]</sup>分析和模拟企业社交网络中的用户交互行为,形成企业用户的组织图和社交图,用两类图构建用户交互模型,用于预测企业用户之间的交互行为。李孝伟等<sup>[51]</sup>提出了一种融合节点与链接属性的社交网络社区划分算法,该算法融合节点属性的相似度、节点间链接权值等链接属性信息,结合聚类算法实现了对社交网络的社区划分。

### 4 结束语

随着社交网络的发展,图数据挖掘技术作为一种数据结构在社交网络中建模变得越来越重要。为了进一步对图数据进行查询、分类、聚类 and 频繁模式挖掘,图数据挖掘技术的研究与应用已经成为一项重要的任务。文中综述了图数据理论研究的相关内容,并主要介绍了图形数据库研究现状和应用领域以及图数据挖掘在社交网络的应用。由于篇幅有限,不可能涵盖这个领域的所有内容,希望这篇综述能对图数据挖掘技术研究应用起到一定的参考作用。

#### 参考文献:

- [1] 丁悦,张阳,李战怀,等.图数据挖掘技术的研究与进展[J].计算机应用,2012,32(1):182-190.
- [2] Giugno R, Shasha D. GraphGrep: a fast and universal method for querying graphs[C]//Proc of 16th international conference on pattern recognition. [s. l.]:IEEE,2002:112-115.
- [3] Yan Xifeng, Yu P S, Han Jiawei. Graph indexing: a frequent structure-based approach[C]//Proc of the 2004 ACM SIGMOD international conference on management of data. New York, NY, USA: ACM,2004:335-346.
- [4] Agrawal R, Borgida A, Jagadish H V. Efficient management of transitive relationships in large data and knowledge bases [C]//Proc of the ACM SIGMOD international conference on management of data. [s. l.]:[s. n.],1989:253-262.

- [5] Trißl S, Leser U. Fast and practical indexing and querying of very large graphs[C]//Proc of the ACM SIGMOD international conference on management of data. [s. l.]: [s. n.], 2007: 845–856.
- [6] Samet H, Sankaranarayanan J, Alborzi H. Scalable network distance browsing in spatial databases[C]//Proc of the ACM SIGMOD international conference on management of data. Vancouver: [s. n.], 2008: 43–54.
- [7] Xiao Yanghua, Wu Wentao, Pei Jian, et al. Efficiently indexing shortest paths by exploiting symmetry in graphs[C]//Proc of 12th international conference on extending database technology. Saint Petersburg: [s. n.], 2009: 493–504.
- [8] Zou Lei, Chen Lei, Ozsu M T. Distance-join; pattern match query in a large graph database[C]//Proc of VLDB. [s. l.]: [s. n.], 2009: 886–897.
- [9] Shang Haichuan, Zhang Ying, Lin Xuemin, et al. Taming verification hardness; an efficient algorithm for testing sub-graph isomorphism[C]//Proc of VLDB. [s. l.]: [s. n.], 2008: 364–375.
- [10] Chen C, Lin C X, Fredrikson M, et al. Mining graph patterns efficiently via randomized summaries[C]//Proc of VLDB. [s. l.]: [s. n.], 2009: 742–753.
- [11] Zeng Z, Tung A K H, Wang J, et al. Comparing stars; on approximating graph edit distance[C]//Proc of VLDB. [s. l.]: [s. n.], 2009: 25–36.
- [12] Zou Zhaonian, Li Jianzhong, Cao Hong, et al. Finding top-k maximal cliques in an uncertain graph[C]//Proc of 26th international conference on data engineering. Long Beach, California: [s. n.], 2010: 649–652.
- [13] Shi Jianbo, Malik J. Normalized cuts and image segmentation [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888–905.
- [14] Newman M E J, Girvan M. Finding and evaluating community structure in networks [J]. Physical Review E, 2004, 69: 026113.
- [15] Xu Xiaowei, Yuruk N, Feng Zhidan, et al. SCAN; a structural clustering algorithm for networks[C]//Proc of the 13th ACM SIGKDD international conference on knowledge discovery and data mining. San Jose: [s. n.], 2007.
- [16] Satuluri V, Parthasarathy S. Scalable graph clustering using stochastic flows; applications to community discovery [C]//Proc of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. Paris: [s. n.], 2009: 737–745.
- [17] 吴 烨, 钟志农, 熊 伟, 等. 一种高效的属性图聚类方法 [J]. 计算机学报, 2013, 36(8): 1704–1713.
- [18] Tiakas E, Papadopoulos A N, Manolopoulos Y. Graph node clustering via transitive node similarity [C]//Proc of PCI. Tripoli: IEEE, 2010: 72–77.
- [19] Ester M, Ge R, Gao B J, et al. Joint cluster analysis of attribute data and relationship data; the connected k-center problem [J]. ACM Transactions on Knowledge Discovery from Data, 2008, 2(2): 1–35.
- [20] Moser F, Ge R, Ester M. Joint cluster analysis of attribute and relationship data without apriori specification of the number of clusters[C]//Proc of knowledge discovery in data. [s. l.]: [s. n.], 2007: 510–519.
- [21] Saigo H, Nowozin S, Kadowaki T, et al. GBoost; A mathematical programming approach to graph classification and regression [J]. Machine Learning, 2009, 75(1): 69–89.
- [22] Tsuda K, Kudo T. Clustering graphs by weighted substructure mining[C]//Proc of the 23rd international conference on machine learning. Pittsburgh: [s. n.], 2006: 953–960.
- [23] Callut J, Francoise K, Saerens M, et al. Semi-supervised classification from discriminative random walks [M]//Machine learning and knowledge discovery in databases. [s. l.]: [s. n.], 2008: 162–177.
- [24] Deshpande M, Kuramochi M, Karypis G. Frequent sub-structure based approaches for classifying chemical compounds [C]//Proc of 3rd IEEE international conference on data mining. [s. l.]: [s. n.], 2005: 1036–1050.
- [25] Mahé P, Vert J P. Graph kernels based on tree-patterns for molecules [J]. Machine Learning, 2009, 75(1): 3–35.
- [26] Kudo T, Maeda E, Matsumoto Y. An application of boosting to graph classification [C]//Proc of Advances in neural information processing systems. Vancouver: [s. n.], 2005: 729–736.
- [27] Cook D J, Holder L B. Graph-based data mining [J]. IEEE Intelligent Systems and Their Applications, 2000, 15(2): 32–41.
- [28] Gonzalez J, Holder L, Cook D. Application of graph based concept learning to the predictive toxicology domain [C]//Proc of the predictive toxicology challenge workshop. [s. l.]: [s. n.], 2001.
- [29] Yoshida K, Motoda H, Indurkha N. Graph-based induction as a unified learning framework [J]. Journal of Applied Intelligence, 1994, 4: 297–328.
- [30] Inokuchi A, Washio T, Okada T, et al. Applying the apriori-based graph mining method to muta-genesis data analysis [J]. Journal of Computer Aided Chemistry, 2001, 2(1): 87–92.
- [31] Kuramochi M, Karypis G. Frequent subgraph discovery [C]//Proc of IEEE international conference on data mining. San Jose: IEEE, 2001: 313–320.
- [32] Vanetik N, Gudes E, Shimony S E. Computing frequent graph patterns from semi-structured data [C]//Proc of IEEE international conference on data mining. [s. l.]: [s. n.], 2002: 458–465.
- [33] Yan Xifeng, Han Jiawei. gSpan; graph-based substructure pattern mining [C]//Proc of IEEE international conference on data mining. [s. l.]: [s. n.], 2002: 721–724.
- [34] Yan Xifeng, Han Jiawei. Close-graph; mining closed frequent graph patterns [C]//Proc of the ninth ACM SIGKOD interna-

- tional conference on knowledge discovery and data mining. Washington, DC, USA: [ s. n. ], 2003: 286–295.
- [35] Zhou Xiaofeng, Gao Lin, Dong Anguo. An algorithm for finding frequent patterns in a large-parse graph [ C ] // Proc of international multi-conference of engineers and computer scientists. [ s. l. ] : [ s. n. ], 2007: 290–294.
- [36] Huan Jun, Wang Wei, Prins J, et al. Spin: mining maximal frequent sub-graphs from graph databases [ C ] // Proc of KDD. [ s. l. ] : [ s. n. ], 2004: 581–586.
- [37] Thomas L, Valluri S, Karlapalem K. MARGIN: maximal frequent sub-graph mining [ C ] // Proc of international conference on data mining. [ s. l. ] : [ s. n. ], 2006: 1097–1101.
- [38] 李继腾, 骆志刚, 丁凡, 等. 最大频繁子图挖掘算法研究 [ J ]. 计算机工程与科学, 2009, 31 ( 12 ) : 67–70.
- [39] 维基百科 [ EB/OL ]. 2001. [http://en.wikipedia.org/wiki/Graph\\_database](http://en.wikipedia.org/wiki/Graph_database).
- [40] Vicknair C, Macias M, Zhao Zhendong, et al. A comparison of a graph database and a relational database [ C ] // Proc of the 48th annual southeast regional conference. Oxford: [ s. n. ], 2010.
- [41] Holzschuher F, Peinl R. Performance of graph query languages: comparison of cypher, gremlin and native access in Neo4j [ C ] // Proc of international conference on extending database technology. [ s. l. ] : [ s. n. ], 2013: 195–204.
- [42] Angles R. A comparison of current graph database models [ C ] // Proc of IEEE 28th international conference on data engineering. Arlington: IEEE, 2012: 171–177.
- [43] Martinez-Bazan N, Gomez-Villamor S, Escalé-Claveras F. DEX: a high-performance graph database management system [ C ] // Proc of 27th international conference on data engineering. Hannover: IEEE, 2011: 124–127.
- [44] Ke Yiping, Cheng J, Ng W. Correlation search in graph databases [ C ] // Proc of 13th ACM SIMKOD international conference on knowledge discovery and data mining. San Jose: [ s. n. ], 2007: 390–399.
- [45] Chairunnanda P, Forsyth S, Daudjee K. Graph data partition models for online social networks [ C ] // Proc of 23rd ACM conference on hypertext and social media. [ s. l. ] : [ s. n. ], 2012: 175–179.
- [46] 第 32 次中国互联网络发展状况统计报告 [ R ]. 出版地不详: 中国互联网信息中心, 2013.
- [47] Aggarwal C, Wang H X. Managing and mining graph data [ M ]. Berlin: Springer-Verlag, 2010.
- [48] Soussi R, Aufaure M, Baazaoui H. Towards social network extraction using a graph database [ C ] // Proc of second international conference on advances in databases, knowledge, and data application. [ s. l. ] : [ s. n. ], 2010: 28–34.
- [49] Kadge S, Bhatia G. Graph based forecasting for social networking site [ C ] // Proc of international conference on communication, information and computing technology. [ s. l. ] : [ s. n. ], 2011.
- [50] Cao Jin, Gao Hongyu, Li L E, et al. Enterprise social network analysis and modeling: a tale of two graphs [ C ] // Proc of INFOCOM. Turin: IEEE, 2013: 2382–2390.
- [51] 李孝伟, 陈福才, 刘力雄. 一种融合节点与链接属性的社交网络社区划分算法 [ J ]. 计算机应用研究, 2013, 30 ( 5 ) : 1477–1480.
- +++++
- (上接第 5 页)
- [3] Huang Jinbo, Chavira M, Darwiche A. Solving MAP exactly by searching on compiled arithmetic circuits [ C ] // Proceedings of the twenty-first national conference on artificial intelligence and the eighteenth innovative applications of artificial intelligence conference. Boston: IEEE, 2006: 1143–1148.
- [4] Wu Dan, Wu Libing. Hierarchical junction trees as the secondary structure for inference in Bayesian networks [ C ] // Proc of eighth ACIS international conference on software engineering, artificial intelligence, networking, and parallel/distributed computing. Qingdao: IEEE, 2007: 706–712.
- [5] Xia Yinglong, Prasanna V K. Parallel exact inference on the cell broadband engine processor [ C ] // Proc of the 2008 ACM/IEEE conference on supercomputing. Austin, Texas: IEEE, 2008: 1–12.
- [6] Zheng Lu, Mengshoel O J, Chong J. Belief propagation by message passing in junction trees: computing each message faster using GPU parallelization [ C ] // Proc of the 27th conference on uncertainty in artificial intelligence. [ s. l. ] : [ s. n. ], 2011.
- [7] 李刚. 贝叶斯网络在制动系统故障中的应用及系统开发 [ D ]. 沈阳: 东北大学, 2007.
- [8] 宫义山, 高媛媛. 基于信息融合的诊断贝叶斯网络研究 [ J ]. 计算机技术与发展, 2009, 19 ( 6 ) : 106–108.
- [9] 黄锦增, 陈虎, 赖路双. 异构 GPU 集群的任务调度方法研究及实现 [ J ]. 计算机技术与发展, 2012, 22 ( 5 ) : 32–36.
- [10] 卢风顺, 宋君强, 银福康, 等. CPU/GPU 协同并行计算研究综述 [ J ]. 计算机科学, 2011, 38 ( 3 ) : 5–9.
- [11] 姚平. CUDA 平台上的 CPU/GPU 异步计算模式 [ D ]. 合肥: 中国科学技术大学, 2010.
- [12] 黄友平. 贝叶斯网络研究 [ D ]. 北京: 中国科学院研究生院, 2005.
- [13] 张舒. GPU 高性能运算之 CUDA [ M ]. 北京: 中国水利水电出版社, 2009.
- [14] 崔晨. 基于 GPU 的 H. 264 编码器关键模块的并行算法设计与实现 [ D ]. 大连: 大连理工大学, 2012.
- [15] NVIDIA. NVIDIA CUDA toolkit version [ EB/OL ]. 2013. <http://developer.nvidia.com/cuda-toolkit-32-downloads>.

# 基于社交网络的图数据挖掘应用研究

作者：[李桃陶](#)，[周斌](#)，[王忠振](#)，[LI Tao-tao](#)，[ZHOU Bin](#)，[WANG Zhong-zhen](#)  
作者单位：[国防科学技术大学 计算机学院, 湖南 长沙, 410073](#)  
刊名：[计算机技术与发展](#)  
英文刊名：[Computer Technology and Development](#)  
年，卷(期)：2014(10)

引用本文格式：[李桃陶](#). [周斌](#). [王忠振](#). [LI Tao-tao](#). [ZHOU Bin](#). [WANG Zhong-zhen](#) [基于社交网络的图数据挖掘应用研究](#)[期刊论文]-[计算机技术与发展](#) 2014(10)