图数据挖掘研究 *

崔景洋
（河北地质大学 信息工程学院，河北 石家庄 050031） Test

〔摘要〕 随着近几年信息技术的发展，人们在生产生活的各个方面积累了大量复杂类型数据结构，图数据结构就是其中之一，对于图挖掘的研究也逐渐成为了科研领域的热点. 文章通过对于目前学术界关于图挖掘的研究成果的总结，介绍了几类基本的图算法. 同时，对于图数据研究的应用与面临的挑战也做了简要分析.

〔关键词〕 数据挖掘；图数据；频繁子图挖掘；图分类；图聚类
〔文章编号〕 1672-2027(2018)01-0038-03 〔中图分类号〕 TP391 〔文献标识码〕 A

0 引言

一般来说，图数据是指以图形为对象的形式化表示，包括点、线、面等属性，是一种常见的数据结构，主要用以表示事物之间的联系、结构等. 相较之传统数据结构而言，图数据是复杂的. 这种复杂表现在处理图数据过程中的各个方面. 存储方面，由于图中边的数量与点的数量成指数级的关系，再加上大数据时代的来临，顶点以及边的规模都变得越来越大，就为图数据的存储提出了新的挑战. 计算方面，传统分类聚类等数据挖掘算法主要是针对简单类型数据进行挖掘，图的特殊性使得普通算法难以应对如此错综复杂的图内顶点之间的关系. 这就使得对于图挖掘算法的需求变得越来越迫切.

在社交网络(social networks)中，其顶点表示人，边表示人与人之间的关系[1]. 对于社交网络，可以通过图聚类算法来进行交际圈的划分，以形成准确的用户画像，来辅助企业的战略决策[2]. 在生物学信息中，图数据可以用来表示蛋白质交互网络(protein-protein interaction，PPI)，通过挖掘实验得到的 PPI 网络[3]，可以使得生物科学家很清晰地认识某种蛋白质结构且不用耗费大量人力财力进行测定实验.

根据图数据的性质进行图数据挖掘的研究，无论是在科学研究上还是生产生活中，都有比较大的意义. 尤其在信息高速产出的今天，对于事物之间错综复杂关系的总结，有利于我们从宏观上对于事物进行更加细致的掌握.

1 图挖掘的主要研究内容

图挖掘这一领域很难用一个时间点来明确图挖掘研究的提出. 近几年，国内外学者对于图挖掘领域的研究成果层出不穷. 一些新概念、新技术例如机器学习、并行化计算等新方法也被应用到了图数据的分析过程之中. 下面介绍几类比较重要的图数据挖掘方法.

1.1 图分类

无论是对于图的研究还是在计算机领域的其他研究中，分类算法都占有比较大的比重. 图的分类是根据图结构的相似程度分别归类. 目前，图分类的方法主要包括基于图频繁结构及基于图核函数的分类方法.

基于图频繁结构的方法主要包含三个主要步骤：首先将图分类问题转化为频繁子图求解问题，通过挖掘得到频繁子图，第二步是选择频繁子图中的分类特征，最后通过第二步所选取的特征构建分类模型[4]. 值得注意的是，选取的分类特征不同，对于分类的结果也有一定的影响. 王海荣[5]考虑到了这个问题，提出了一种

* 收稿日期：2018-01-17
基金项目：国家自然科学基金(61503260)；河北省研究生创新资助项目(CXZZSS2017132).
作者简介：崔景洋(1992-)，男，河北固安人，河北地质大学在读硕士研究生，主要从事数据挖掘与机器学习研究.

通过加权法提取频繁子图的算法,并将之应用到了文本分类的过程中.

图核是通过核函数的方法,将图映射到高维空间.低维线性不可分的图在更高维度就会变得可分.不过,映射到高维空间后势必会增加计算量,但是核函数可以代替高维计算结果,往往可以减小计算量以降低计算时间.所以这种分类方法的难点在于图核函数的选取.主流的图核函数选取方法有两种:一种是基于游走的图核[6],另一种是基于循环的图核[7].王桂娟等[8]对图内在结构进行分析,通过机器学习的方法对图进行分类.

1.2　图聚类

图的聚类分析在目前的研究现状中,主要包含两个方向.第一个方向是在一个图内进行顶点或边的聚类,第二个方向是在多个图中进行图与图之间的相似性归纳.

图内顶点聚类是将图中联系密切的顶点及其相关边组成一个子图的过程[9],聚类的任务就是要从大图中找出那些相似的顶点.另一方面,对于若干小图来说,图之间的距离就可以根据结构相似性函数来衡量.高阳等[10]使用拉普拉斯矩阵和谱聚类方法对于图数据进行聚类,将 NP 难的问题转化为了多项式时间算法.

1.3　频繁子图挖掘

频繁子图问题就是找出在某一图结构中频繁出现的子图结构,并将这些频繁子图结构用到分类、聚类、搜索等方面.传统查找频繁项集的方法例如 Apriori,FP-Growth 算法等在频繁子图求解时也是适用的.单晓欢等[11]提出了一种标签约束的频繁子图 Top-k 查询方法.另外,针对当前形势下的大数据量的问题,通常的解决方式是将频繁子图挖掘问题进行并行化以加快处理速度[12,13].

1.4　图关键词查询

图中数据关键字查询的主要难点在于图的遍历和检索[14],查询的可以是图的结构,也可以是顶点或者边的属性.按照其返回结果的不同,可以分成两类:一类返回的是关键词位置,另外一类返回的是关键词的结构.一般的处理过程是根据遍历索引的方法定位关键词所在的位置,并记录搜索结果后,根据一定的排序策略将结果展现给用户[15].

1.5　图匹配

图的一个匹配就是一个图中若干没有端点的边的集合,最大匹配就是求这个边集最大有多少条边.图匹配技术可以作为图数据查询的一种重要技术手段,一般分为精确图匹配以及非精确图匹配[16].大部分情况下需要一种可以接受错误和忍耐噪声的一种算法来解决实际生活中遇到的问题.所以可以通过定义一种距离,来评价图与图之间的相似程度[17].

2　图挖掘应用的发展趋势

2.1　社交网络挖掘

近几年,互联网发展迅猛.越来越多的人们选择将互联网作为一种社交手段.在此背景下,对于社交网络的研究层出不穷.对社交网络进行分析,可以看做是上述图数据挖掘研究内容的一个融合性应用.陈克寒等[18]提出了一种基于用户聚类的社交网络推荐方法.根据用户与用户之间的关系将其进行聚类分析,并跟据聚类的结果构架了一个基于用户主题的推荐系统.周方[19]将图分类技术应用于社交网络,通过标记社交网络中节点的形式,来进行社交网络的社会学分析.张青[20]设计了一种基于信息熵传传播模型的子图查询方法,该方法在大数据量时有着更高的效率.

2.2　学术网络挖掘

对于学术网络的研究相较之社交网络而言,无论是从数量上还是质量上都是远远达不到的.但是 2008 年开始,我国授予博士学位的人数就超过了美国,成为了世界上每年授予博士学位人数最多的国家.同时,我国学者仅在 2015 年一年发表的论文数量就高达百万篇[21].学术网络可以加强学者之间的交流,对于知识共享、思维方式等方面的推动有着不小的作用.刘萍等[22]根据文献间的引用关系对于学者学术影响力进行了研究,通过学者文献影响力测度新指标为图书情报学(LIS)领域的初学者提供了一个清晰明确的学习方向.

2.3　图数据库

许多新兴计算框架的出现,使得传统图数据的计算不得不考虑移植问题,开始有专门针对图数据优化的

数据库产生,例如:Neo4j[23],Graphchi[24]等.也有针对图的计算框架产生,例如Spark平台的图计算包——GraphX.邱胜海等[25]利用Neo4j提出了一种云计算环境中的图数据库建模方法.

3 总结

当今时代是大数据时代,数据量的急剧增加与图结构的复杂相结合将导致对于图的计算将变得越来越复杂,这就为图数据挖掘提出了如下要求:

1)图数据存储.传统方式方法对于解决大数据的问题已经显得力不从心,那么是否能够出现一种新型的数据结构来存储图数据,以方便计算以及与其他设备或者节点间的通信.

2)图数据计算.对于图数据的计算都需要反复迭代,这将耗费巨大的计算资源.目前主流算法都是针对规模较小的图数据所设计的内存算法,但是在现实生活中,图数据往往都是海量的储存在硬盘或者集群之中的数据.

3)实时性.在线实时处理业务的迅猛成长为大数据量的图数据流的解决方法提供了一个广阔的发展前景.例如在社交网络中的节点往往是动态变化的,而且由于这些数据不能离线存储,那么就要求我们去寻求针对图数据流的处理方法.

由于数据的不确定性是普遍存在的,而目前的主流算法解决的都是确定图的数据挖掘,如何将图的不确定性与已经解决的确定图问题建立联系也为研究人员提供了一个新的发展方向.

参考文献:

[1] 李桃陶,周 斌,王忠振.基于社交网络的图数据挖掘应用研究[J].计算机技术与发展,2014(10):6-11

[2] 马 超.基于主题模型的社交网络用户画像分析方法[D].合肥:中国科学技术大学,2017

[3] 雷秀娟,黄 旭,吴 爽,等.基于连接强度的PPI网络蚁群优化聚类算法[J].电子学报,2012,40(4):695-702

[4] 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

[5] 王海荣.基于加权频繁子图挖掘的图模型在文本分类中的应用[J].科学技术与工程,2014,14(22):80-85

[6] GÄRTNER T,FLACH P,WROBEL S. On graph kernels:hardness results and efficient alternatives[C]// Computational Learning Theory and Kernel Machines,Conference on Computational Learning Theory and,Kernel Workshop,Colt/kernel 2003,Washington,Dc,Usa,August 24-27,2003,Proceedings. DBLP,2003:129-143

[7] 李宇峰,郭天佑,周志华.用于图分类的组合维核方法[J].计算机学报,2009,32(5):946-952

[8] 王桂娟,印 鉴,詹卫许.一种新的基于嵌入集的图分类方法[J].计算机研究与发展,2012,49(11):2311-2319

[9] 邹佩钢,陈 军.基于CombBLAS的同辈压力图聚类并行算法的设计与实现[J].计算机工程与科学,2017,39(3):424-429

[10] 高 阳,李昌华,李智杰,等.基于谱特征和图分割的图聚类算法[J].计算机工程与应用,2017(15):222-226

[11] 单晓欢,王广香,宋宝燕,等.大规模动态图中标签约束的频繁子图Top-K查询[J/OL].计算机科学与探索,2017-01-08

[12] 严玉良,董一鸿,何贤芒,等.FSMBUS:一种基于Spark的大规模频繁子图挖掘算法[J].计算机研究与发展,2015,52(8):1768-1783

[13] 孙鹤立,陈 强,刘 玮,等.利用MapReduce平台实现高效并行的频繁子图挖掘[J].计算机科学与探索,2014,8(7):790-801

[14] PAN Y,WU Y. ROU:advanced keyword search on graph[C]// ACM International Conference on Conference on Information & Knowledge Management. ACM,2013:1625-1630

[15] 杨书新,徐丽萍,夏小云,等.图数据关键词查询研究进展[J].电子学报,2014,42(11):2260-2267

[16] 于 静,刘燕兵,张 宇,等.大规模图数据匹配技术综述[J].计算机研究与发展,2015,52(2):391-409

[17] WOLVERTON M,BERRY P,HARRISON I,et al. LAW:a workbench for approximate pattern matching in relational data [C]// In Proceedings of the Fifteenth Innovative Applications of Artificial Intelligence Conference,2013:143-150

[18] 陈克寒,韩盼盼,吴 健.基于用户聚类的异构社交网络推荐算法[J].计算机学报,2013,36(2):349-359

[19] 周 方.社交网络节点分类技术研究[D].沈阳:辽宁大学,2015

[20] 张 青.社交网络中的子图查询研究[D].徐州:中国矿业大学,2014

[21] 马丙超.基于引文网络的文献在线推荐系统研究和实现[D].大连:大连理工大学,2016

[22] 刘 萍,杨 宁,邹德安.基于文献引文网络的学者学术影响力测度研究[J].情报理论与实践,2017,40(3):35-41

[23] VUKOTIC A,WATT N,ABEDRABBO T,et al. Neo4j in action[M]. New York:Manning Publications Co. 2014

[24] KYROLA A,BLELLOCH G,GUESTRIN C. GraphChi:large-scale graph computation on just a PC[C]// Usenix Conference on Operating Systems Design and Implementation. 2012:31-46

[25] 邱胜海,王云霞,樊树海,等.云环境下图数据库建模技术及其应用研究[J].计算机应用研究,2016,33(3):794-797

(下转第46页)

Based on the above applications, this article designed educational administration system, which is introduced from the student level, the administrator level and the teacher level respectively. Administrator interface to achieve student information management, teacher information management, departmental information management, subject information management, test results management and administrator information maintenance. Into the student level, you can see the students corresponding to the subject's performance, you can also change the login password, etc. Teacher information management can achieve teacher information to add and manage; department information management can achieve college, professional and class information to add and manage. Subject information management to achieve the addition and management of subject information, subject status management, the designated teacher and course designated test. Test results management can generate test scores and add results; into the teacher level, you can achieve the teacher password changes, teacher information changes, as well as more complex subjects like increased, increased professional, increased classes.

The system design is based on ASP. NET application development technology and SQL Server data, and adopts C/S mode structure. This system uses the educational system as a reference to introduce the whole process of website system construction. To some extent, the system is Web site development provides a reference.

〔**Key words**〕 the teaching management; department management; teacher management; student management; grade management

(**上接第** 40 **页**)

Research on Graph Data Mining

CUI Jingyang

(College of Information Engineering, Hebei GEO University, Shijiazhuang Hebei 050031, China)

〔**Abstract**〕 With the development of information technology in recent years, human being has accumulate a large number of complex types of data structures in production and living in all aspects. The data structure of graph is one of them, and the research on graph mining has gradually become the hotspot in scientific research field. Based on the definition graphs, this paper introduces several basic graph algorithms by summarizing the research results of graph mining in academic circles. Simultaneously, the applications of graph data research and the challenges are also briefly analyzed.

〔**Key words**〕 data mining; graph data; frequent graph; graph classification; graph clustering