软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2015,26(1):167−177 [doi: 10.13328/j.cnki.jos.004692]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

图聚集技术的现状与挑战*

潘秋萍, 游进国, 张志朋, 董朋志, 胡宝丽

(昆明理工大学 信息工程与自动化学院,云南 昆明 650500)

通讯作者: 游进国, E-mail: jgyou@126.com

摘 要: 图聚集技术旨在获取能够涵盖原图大部分信息的简洁超图,用于提炼概要信息、解决存储消耗和社交隐私保护等问题.对当前的图聚集技术进行研究,综述了现有图聚集技术中的分组方法并对其进行分类,将分组标准划分为基于属性一致性、基于邻接分组一致性、基于关联强度一致性、基于邻接顶点一致性和基于零重建误差这 5 类;在高层次上将各分组标准概括为基于属性、基于结构和同时基于属性和结构的图聚集.较为全面地总结和分析了当前图聚集技术的研究现状和进展,并探讨了未来研究的方向.

关键词: 图数据;图聚集;分组标准;属性信息一致;结构信息一致

中图法分类号: TP311

中文引用格式: 潘秋萍,游进国,张志朋,董朋志,胡宝丽.图聚集技术的现状与挑战.软件学报,2015,26(1):167−177. http://www.jos.org.cn/1000-9825/4692.htm

英文引用格式: Pan QP, You JG, Zhang ZP, Dong PZ, Hu BL. Progress and challenges of graph aggregation and summarization techniques. Ruan Jian Xue Bao/Journal of Software, 2015,26(1):167−177 (in Chinese). http://www.jos.org.cn/1000-9825/4692.htm

Progress and Challenges of Graph Aggregation and Summarization Techniques

PAN Qiu-Ping, YOU Jin-Guo, ZHANG Zhi-Peng, DONG Peng-Zhi, HU Bao-Li

(Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China)

Abstract: Graph aggregation and summarization is to obtain a concise supergraph covering the most information of the underlying input graph, and it is used to extract summarization, solve storage consumption and protect privacy in social networks. This paper investigates current graph aggregation and summarization techniques and further reviews and classifies their partitioning/grouping methods. Based on the consistency of grouping information, five grouping criteria are specified: The consistency of attribute information, the consistency of neighborhood group, the consistency of connection strength, the consistency of neighborhood vertex and reconstruction zero error. From the top level view, graph aggregation and summarization techniques can be classified into three types, namely, attribute similarity, structure cohesiveness and the hybrid of both. This paper comprehensively summarizes the state of art of current research works, and explores the research directions in the future.

Key words: graph data; graph aggregation; grouping criteria; consistency of attribute information; consistency of structure information

图数据被广泛地应用于多种领域,用于描述事物之间的复杂关联,如各用户之间相互加为好友形成的社交网、各作者相互合作形成的学术研究网、各网页之间相互链接形成的 Web 网、各公路相互交错形成的公路网以及电力网、蛋白质交互网和语义网等[1−3].随着这些领域的发展和数据的增加,当前图数据应用面临如下挑战:

- 潜在信息很难获取

* 基金项目: 国家自然科学基金(61462050); 云南省自然科学基金(2013FZ020, 201303095); 云南省教育厅科学研究基金重点项目(2013Z125); 高等学校学科创新引智计划(111 计划)(B12028)

收稿时间: 2013-05-29; 修改时间: 2013-09-09; 定稿时间: 2014-07-01; jos 在线出版时间: 2014-08-19

CNKI 网络优先出版: 2014-08-19 14:21, http://www.cnki.net/kcms/doi/10.13328/j.cnki.jos.004692.html

社交网、通信网和 Web 网等应用中图数据规模不断扩大,一些大规模图已由上亿个顶点和千亿条边构成[4],很难从中直接获取需要的信息.这些应用所构成的图已经大到无法直接观察去提炼其潜在的有用信息,只能将其中属性或结构相似的顶点聚集在一起,从而在更高的层次去研究这些图,准确地把握其中有效的信息[5].

- 存储空间的局限性

在存储图数据时不仅要存储顶点,而且要存储顶点间的关系,因此图数据的存储更占用空间.有效地压缩图数据的存储空间,对于提高图数据挖掘与分析算法的可扩展性、改进大图处理效率具有较大的实际意义.

- 社交网的隐私保护

社交网络作为一项虚拟的信息服务,占据的却是所有用户的真实信息.因此,社交网的隐私保护已成为当前新的研究热点.由于社交网络中的顶点对应着具体的人,所以可以将顶点集合抽象为一个超点,在一定程度上隐藏底层图的顶点和结构信息,起到有效抵制黑客攻击的作用[6].

图聚集技术由于能够有效地处理上述问题,因而成为学术界和工业界关注的新的研究内容.图聚集将原图中的顶点和边集合抽象到一个更高的层次,从而获得一个简洁的超图.该超图能够涵盖原图中几乎所有的信息,其顶点称为超级顶点(简称超点),它代表原图中的一组顶点;其边称为超级边(简称超边),它代表其关联的两个超点之间的边.

一些文献存在与图聚集含义相同或相似的专业术语,如图压缩(graph compression)[7−10]、图概要(graph synopsis)[1]、图概括(graph summarization)[11−13]、图约化(graph simplification)[14]、网络抽象(network abstract)[15]等.以上术语当其输出为一个超图时,本文统一称为图聚集.但是有些术语与图聚集间存在较大的差异,如图聚类(graph clustering)[16]、图划分(graph partitioning)[17]等.

图划分与图聚集均划分图中顶点:

- 图聚集需保持同一组中顶点属性或者结构相似,且使用存储消耗、重建误差等度量函数评估结果图的性能;而图划分则保持每组中顶点数目近似一致,且交叉边总数最小;
- 通过交叉边数量及容量、负载均衡和数据冗余等衡量划分的好坏;
- 两者也应用于不同的领域,其中,图聚集一般用于图数据的分析与挖掘,而图划分则应用于大图的分布式存储.

图聚类与图聚集的相似点在于它们均将顶点进行分组.区别为图聚类侧重寻找局部稠密的结构;而图聚集从全局出发,侧重寻找属性或者结构相似的顶点,它不仅概括稠密的结构,还概括稀疏的结构.图聚类和图聚集之间详细的比较见表 1.

Table 1 Comparision between graph aggregation and graph clustering

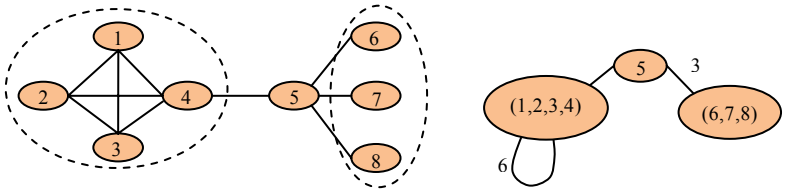
表 1 图聚集与图聚类之间的比较

	分组顶点连通性	是否概括稠密或稀疏结构	分组的视角	输出是否为图
图聚类	高	稠密	局部	不要求
图聚集	不要求	均可	整体	是

如图 1 所示,原图 $G(V,E)$中,V={1,2,3,4,5,6,7,8},E={(1,2),(1,3),(1,4),(2,3),(2,4),(3,4),(4,5),(5,6),(5,7),(5,8)},经过如图 1(a)所示的分组后,得到超图 $G_S(V_S,E_S)$(如图 1(b)所示).该超图含有 3 个超点 V_S={(1,2,3,4),5,(6,7,8)}以及 3 条超边,其中,超点(1,2,3,4)上的超边权值为 6,表示超点内部顶点间所有边的数目为 6;超点(5)和超点(6,7,8)之间的超边权值为 3,代表两超点间相连的边数为 3.

由图 1 也可分析图聚集和图聚类之间的差别.图聚集不仅概括了如(1,2,3,4)间的局部稠密结构,而且从全局出发,由顶点(6,7,8)的邻接点均为顶点 5,概括了(6,7,8)之间稀疏的结构;但图聚类只能概括(1,2,3,4)间稠密的结构.

本文将图聚集技术划分为基于属性的图聚集、基于结构的图聚集、基于属性/结构的图聚集这 3 种,图聚集技术的分类标准取决于分组的特征,在第 1 节将详细描述不同分组标准的限制以及它们之间的关系,并根据分组标准将图聚集技术进行分类.



(a) 原图 (b) 相应的超图

Fig.1 Example of supergraph

图 1 超图示例

1 图聚集分类

图聚集过程中最重要的环节就是分组,为了将输入图抽象到更高的层次,需要将图划分成若干组,进行聚集,形成超点和超边.在当前图聚集算法中,文献[10]设置了基于邻接分组一致性的分组标准,使得同一分组内部各顶点具有相同的邻接分组,但并未区分各顶点的关联强弱;文献[18−20]基于邻接顶点一致性,使得同一分组内部各顶点具有相同的邻接顶点,然而分组内部顶点间的关联却并不一致;文献[11]提出零重建误差分组标准,使得分组内部各顶点间存在相同的关联,且分组间的边涵盖两分组间可存在的所有的边.这些标准仅适用于不含属性的图,但实际上大多数的图均具有属性.因此,文献[21−26]将分组标准设置为基于属性一致性的划分,使得各超点内部的属性值一致,但该分组标准忽略了图中的结构信息;文献[27−29]使得各分组均满足属性以及邻接分组一致性,该分组标准同时包含了图中的属性和结构信息,但基于邻接分组一致性得到的分组,其内部各顶点的关联强弱无法区分;文献[13]提出了基于属性以及关联强度一致的分组标准,其中,关联强度一致实现了同一分组内部各顶点关联的邻接顶点数目相同.由于图中包含的顶点非常庞大,且其中的属性以及结构较为复杂,因此上述精确分组标准并不完全适用于实际应用,但这些算法均以精确的分组标准为基准,在可控制的误差范围内获取有损图聚集.以上所述的图均指的是同构图,即,图中各顶点以及边的类型均一致,但实质上,就图的类型而言,不仅包含同构图,也包含异构图.异构图中各顶点以及边的类型均不相同,相对于同构图而言包含的信息更多,本文则着重针对同构图进行分析.

对此,本文提出 5 类分组标准:基于属性信息一致性的划分、基于邻接分组一致性的划分、基于关联强度一致性的划分、基于邻接顶点一致性的划分、基于零重建误差的划分.这 5 类分组标准的形式化定义如下所示,具体举例如图 2 所示.其中,设原图为 $G(V,E)$, $neighborgroups_{V_i}(u)$ 表示分组 V_i 中顶点 u 的邻接顶点所在分组, $neighbornum_{V_i,V_j}(u)$ 表示分组 V_i 中顶点 u 关联的邻接分组 V_j 中顶点的数目, n_{V_j} 表示分组 V_j 中的顶点总数, $neighbornodes_{V_i}(u)$ 表示顶点 u 在分组 V_i 中的邻接顶点.

- 标准 I. 基于属性信息一致性的划分:若两顶点在同一分组,则两顶点的任一对应属性值(用 a_i 表示)均相同.即:

$$\forall u,v\in V,V_j\subseteq V,\text{若 } u\in V_j \text{ 且 } v\in V_j,\text{则}\forall i,a_i(u)=a_i(v);$$

- 标准 II. 基于邻接分组一致性的划分:若分组中某一顶点与另一分组中某些顶点存在邻接关系,则分组中任一顶点均与另一分组存在邻接关系.即:$\forall u,v\in V_i,V_i,V_j\subseteq V$,若 $neighborgroups_{V_i}(u)=V_j$,则:

$$neighborgroups_{V_i}(u)=neighborgroups_{V_i}(v);$$

- 标准 III. 基于关联强度一致性的划分:若分组中某一顶点与另一分组中某些顶点存在邻接关系,则分组中任一顶点与另一顶点中相同数目的顶点存在邻接关系.即:

$\forall u,v\in V_i,V_i,V_j\subseteq V$,若 $neighborgroups_{V_i}(u)=V_j$ 且 $neighbornum_{V_i,V_j}(u)=k(k\leqslant n_{V_j})$,则:

$$neighborgroups_{V_i}(u)=neighborgroups_{V_i}(v)\text{且}neighbornum_{V_i,V_j}(u)=neighbornum_{V_i,V_j}(v);$$

- 标准 IV. 基于邻接顶点一致性的划分:分组中某一顶点与另一分组中某些顶点存在邻接关系,则分组

中的任一顶点均与另一分组中所有顶点存在邻接关系.即:

$\forall u,v\in V_i,V_i,V_j\subseteq V$,若 $neighborgroups_{V_i}(u)=V_j$,则:

$$neighborgroups_{V_i}(u)=neighborgroups_{V_i}(v),neighbornum_{V_i,V_j}(u)=neighbornum_{V_i,V_j}(v)=n_{V_j};$$

- 标准 V. 基于零重建误差的划分:若分组中某一顶点与另一分组中某些顶点存在邻接关系,则分组中的任一顶点均与另一分组中所有顶点存在邻接关系,且组内任意两顶点均相连,或者均不相连.即:

$\forall u,v\in V_i,V_i,V_j\subseteq V$,若 $neighborgroups_{V_i}(u)=V_j$,则:

$$neighborgroups_{V_i}(u)=neighborgroups_{V_i}(v),neighbornum_{V_i,V_j}(u)=neighbornum_{V_i,V_j}(v)=n_{V_j}$$且

$$neighborgroups_{V_i}(u)=V_i或neighborgroups_{V_i}(u)=\varnothing.$$

如图 2 所示,A_1,A_2,A_3 是基于属性一致性划分得到的 3 个不同分组,此时,A_1,A_2,A_3 满足标准 I,即属性信息一致性;对于超点 A_1,B 而言,A_1 中每个顶点均能在 B 中找到其邻接顶点,且 A_1 中顶点与 B 中对应的顶点关联的边数不一致,因此,A_1 相对于 B 仅满足标准 II,即邻接分组一致性;A_2 中每个顶点均能在 B 中找到其邻接顶点,且关联边数一致,因此,A_2 相对于 B 满足标准 III,即关联强度一致性;对于 A_3 与 C 而言, A_3 内的每个顶点均与 C 中所有顶点关联,因此,A_3 满足标准 IV,即邻接顶点一致性;但 A_3 内部两个顶点存在关联,导致超图不能完全重建成原图,而 A_2 与 C 不仅满足标准 IV,且 A_2,C 内部各顶点间关联一致,因此,A_2 相对于 C 满足标准 V.

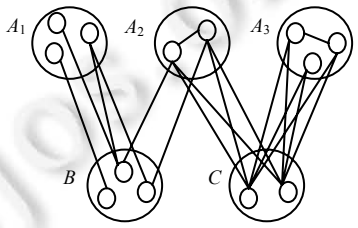


Fig.2 Example of grouping criteria

图 2 分组标准示例

以上分组标准可概括为:基于属性一致性分组标准,基于结构一致性分组标准.其中,标准 I 属于基于属性一致性分组标准;标准 II,III,IV,V 属于基于结构一致性分组标准,且相互之间存在包含的关系,即,II⊇III⊇IV⊇V(如图 3 所示),其中,U 表示所有可能的分组集合.根据图 3 中各分组标准对应的集合关系对图聚集进行以下分类:当图聚集仅满足标准 I 时,称为基于属性的图聚集;当仅满足 II,III,IV,V 时,称为基于结构的图聚集;当满足 I-II,I-III,I-IV, I-V 的交集时,称为同时基于属性和结构的图聚集.

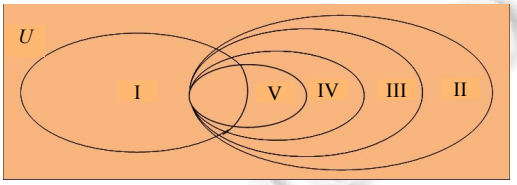


Fig.3 Relationships between grouping criteria

图 3 分组标准之间的联系

由上文可知,图聚集可分为基于属性的图聚集、基于结构的图聚集和基于属性/结构的图聚集.其中,基于属性和基于结构的图聚集均能实现属性或结构在同一分组一致、不同分组相异的特性.但由于基于属性的图聚集分组内部结构松散,而基于结构的图聚集分组内部属性值不一致,因此,当前更热衷于研究基于属性/结构一致性的图聚集.

图聚集也可根据重建误差划分为有损图聚集和无损图聚集.无损图聚集能够由超图重建成完整的输入图,能够精确地回答基于输入图的查询.但有损图聚集更有实际意义,这是由于一致性标准对于实际应用中的图数

据并不完全适用.一些应用中的图数据无法避免噪音和不确定性,当对这些含有噪音和不确定性的图数据进行一致性分组时,得到的分组数目十分庞大,从而造成低效的图数据分析和挖掘.此外,在某些应用场景下,图查询不需要精确的回答,如:在并不精确的邻居顶点集合下,社交网仍能识别出相关的社区;Web 网在并不清楚地知道每个网页所有链接网页的条件下,也能够获得很好的 PageRank 值等.图 1(b)所示的超图丢失了(4,5)这条边的信息,因此只有 1/4 的概率准确地重建成原图,图 1 所示的图聚集是有损图聚集.

2 主要图聚集技术

2.1 基于属性的图聚集

基于属性一致性的图聚集均满足分组标准 I,一般与 OLAP 技术相结合,代表技术为 Graph OLAP/Cube.

OLAP 被广泛地应用于数据仓库,提供决策分析的功能.它基于多维数据模型,将数据看作数据立方体形式,通过计算每个维度组合的度量,提高在线分析性能.而在实际的多维网络中,暗含着的图结构数据却存在着管理、查询、聚集等问题.

文献[21−24]涉及的 Graph OLAP 技术利用网络快照构建图模型,基于图的属性将网络快照进行聚集,得到高层次的超图,并实现 OLAP 的基本操作,如:上卷、下钻、切片/切块等.文献[21]将属性分为两种,分别是 Info-Dims 和 Topo-Dims,其中,Info-Dims 基于不同的维度和粒度决定网络快照所在的分组;Topo-Dims 则操作单个网络内的顶点和边,将具有关联的顶点集合聚合成一个整体.基于 Info-Dims 和 Topo-Dims 构建的 OLAP 分别为 I-OLAP 和 T-OLAP.

对于 I-OLAP,利用 Info-Dims 覆盖多个网络快照,构建 I-aggregated graph.由于多个快照描述的是相同顶点间的不同关联,因此,I-aggregated graph 针对相同的顶点,将快照集合的顶点和边的属性进行聚集,不改变原图的拓扑结构.

对于 T-OLAP,利用 Topo-Dims 压缩单个网络的拓扑结构,得到 T-aggregated graph.任一顶点对应原图中 Topo-Dims 值相同的顶点分组,顶点属性覆盖对应分组中顶点属性集合,压缩原图的拓扑结构,隐藏图中的细节信息.

而 Graph Cube[25]类似于关系数据库构建 Cube,但不同于以往的度量值,图的度量是一个聚集网络,它代表的是对应顶点分组后,原图的聚集结果.而对于已构建的 Graph Cubes,OLAP 被分为 Cuboid Query 和 Crossboid Query.Cuboid Query 对应着传统的 OLAP,即,对于一个或者多个维度进行上卷、下钻、切片、切块等,它们均仅涉及一个 Cuboid.而 Crossboid Query 则是在多个 Cuboids 上进行交叉查询.

对于上述的图聚集技术,仅针对顶点的属性进行分组,无法回答基于原图结构的查询,如顶点 a,b 间相同的邻居顶点等;其次,相比于传统的 OLAP,网络更加庞大和复杂,所以在多维空间中必须要压缩多重聚集网络.

2.2 基于结构的图聚集

2.2.1 基于概率的图聚集

文献[11]的切入点是在超图上精确地回答基于原图的查询,针对这种查询,方法之一就是将超图重建成原图进行查询,此时,超图的重建误差越小,查询结果越精确.虽然该方法利用随机划分进行分组,但它以零重建误差为基准,因此最终结果的分组均满足分组标准 V.

基于 Random-Worlds 的 Indifference 原则,超图重建原图时,得到的各结果图的概率一致(包含原图的概率).将超图转换为矩阵,根据概率公式计算其期望邻接矩阵,概率公式如公式(1)~公式(3)所示.其中,E_i 表示超点 V_i 内部的边数目,E_{ij} 表示超点 V_i 和超点 V_j 之间的边数目,$|V_i|$表示超点 V_i 内部顶点数目.如图 4 所示.

1) 当 u,v 属于同一超点 V_i 中的不同顶点时:

$$\bar{A}=\frac{2E_i}{|V_i|(|V_i|-1)} \tag{1}$$

2) 当 u,v 属于不同超点 V_i,V_j 中的不同顶点时:

$$\bar{A}(u,v)=\frac{E_{ij}}{|V_i|\times|V_j|} \tag{2}$$

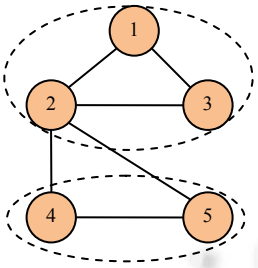
3) 当 u,v 属于相同顶点时:

$$\bar{A}(u,v)=0 \tag{3}$$

通过超图得到的期望邻接矩阵和原图的邻接矩阵间的差异构建重建误差,重建误差越小,得到的超图精确性越好,重建误差公式如公式(4)所示.

$$RE(A\,|\,\bar{A})=\frac{1}{|V|^2}\sum_{i=1}^{|V|}\sum_{j=1}^{|V|}|\,\bar{A}(i,j)-A(i,j)\,| \tag{4}$$

基于期望邻接矩阵的图聚集方法得到的超图能够精确地回答基于原图的查询,但仅限于不含属性信息的图.通过计算模型的描述消耗 $TB(\bar{A})$,能够有效地控制结果集的数目.但由于需要保证超图的信息完整性,它不能最小化结果集的数目.



(a)

	1	2	3	4	5
1	0	1	1	0	0
2	0	0	1	1	1
3	1	1	0	0	0
4	0	1	0	0	1
5	0	1	0	1	0

(b) (a)策略的邻接矩阵

	1	2	3	4	5
1	0	1	1	1/3	1/3
2	1	0	1	1/3	1/3
3	1	1	0	1/3	1/3
4	1/3	1/3	1/3	0	1
5	1/3	1/3	1/3	1	0

(c) (a)策略的期望邻接矩阵

Fig.4 Example of graph aggregation based on probability

图 4 基于概率的图聚集举例

2.2.2 基于最大化压缩的图聚集

S-Node[10]提出了 Web 图的两层表示方式,顶层图由超点和超边组成,它们分别指向底层对应的有向图结构.其顶层图即是超图,利用分组标准 II,实现邻接分组一致.在 Web 图中体现为将域相同的顶点聚集形成初始分组,通过相同的 URL 前缀循环提炼原始分组.最后,基于 *K*-Means 等聚类方法进一步压缩,直至不能分裂.S-Node 的表示包含 4 个部分,分别是超点图(supernode graph)、超点内部关系图(intranode graph)、超点间存在的超边图(positive superedge graph)、超点间不存在的超边图(negative superedge graph).根据存在的超边和不存在的超边数目比较,选择数目较少的作为 S-Node 中的超边图.但该方法存在这样的问题:超图中的超边未考虑两超点间的关联强度,仅存在一条边的两超点间也会构建超边,这给理解图中原有信息带来了困难,

文献[18]对 S-Node 进行改进,通过 MDL 原则构建超边,使其在强关联的情况下才构建超边,弱关联的边保存在 corrections 集合中,保证了信息的完整性.其中,各超边均包含两邻接分组间的所有边集合,实现分组标准 IV,即邻接顶点一致,并且通过创建自连接边有效地将紧密的团结构合并成一个超点.S-Node 通过 URL 的信息仅获取 Web 图的无损压缩,但基于 MDL 的图聚集方法不仅可以进行无损压缩,也允许有损压缩.在实际应用中,由于图的数据量较大,此时,有损压缩的意义更大.

基于 MDL 的图聚集方法,利用信息论中的 MDL 原则获取度量函数消耗减少率,如公式(5)、公式(6)所示.循环获取 s 值最大的顶点对合并,直至 s 的值小于等于 0 为止.

$$s(u,v)=(c_u+c_v-c_w)/(c_u+c_v) \tag{5}$$

$$c_{u,x}=\min\{|\Pi_{u,x}|-|A_{u,x}|+1,|A_{u,x}|\} \tag{6}$$

其中,$s(u,v)$表示合并前后消耗减少率,c_u,c_v,c_w 分别表示顶点 u,v,w 的消耗值,$A_{u,x}$ 表示 u,x 之间实际存在的边集合,$\Pi_{u,x}$ 表示 u,x 之间所有的边集合.如图 5 所示.

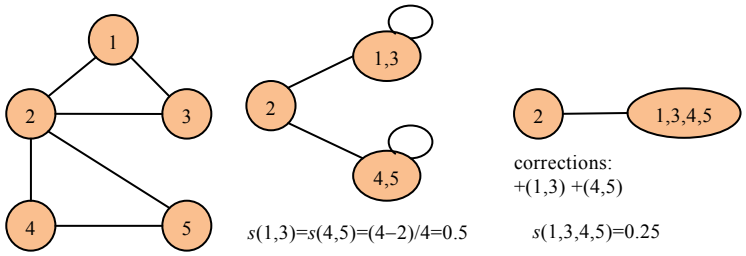


Fig.5 Example of graph aggregation based on MDL

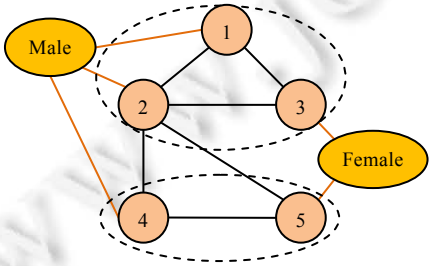
图 5 基于 MDL 的聚集方法举例

基于 MDL 的图聚集方法对应的 corrections 占据的空间消耗较大,几乎占所有消耗的 80%,可针对这个进行再次压缩;其次,它的计算量比较大,每次循环合并两顶点,均需重新计算与两顶点或者其邻居顶点相关的顶点对的 s 值;本质上,它仅考虑了图的结构信息,忽略了顶点属性、边的类型以及权重等其他信息.

2.3 基于属性和结构的图聚集

2.3.1 基于熵模型(entropy)的图聚集

文献[13]利用信息论中的熵模型衡量属性/结构的信息量,并通过创建属性顶点和属性边将属性信息转换为邻接信息,如图 6 所示.



	Node	Male	Female
S_1	1	1	0
	2	1	0
	3	0	1
S_2	4	1	0
	5	0	1

基于熵的属性一致性:

$Entropy(S_1)=-0.66\times\log 0.66-0.33\log 0.33=0.9$

$Entropy(S_2)=-0.5\log 0.5-0.5\log 0.5=1$

Fig.6 Example of graph aggregation based on entropy

图 6 基于熵的图聚集方法举例

该方法不仅实现了分组标准 I 和 II,保证了属性和邻接分组的一致性,而且实现了分组标准 III,获取了关联强度的一致性.设置参数 k 控制各分组中顶点数目,但为了均衡属性和结构的信息,在一致性分组的基础上放宽了三者的限制.基于熵模型的度量函数 $R(P)$的公式如公式(7)所示.

$$R(P)=\sum_{S_i\in P}|S_i|\times WeightedEntropy(S_i) \tag{7}$$

通过权值λ和$(1-\lambda)$,自定义属性和结构在图聚集过程中所占的比重.由公式(8)、公式(9)可得到超点 S_i 分别关联权值为λ的属性顶点 a_j 和权值为$(1-\lambda)$的其他超点得到的熵值:

$$WeightedEntropy(S_i)=\lambda\sum_{j=1}^{l}H(p(a_j^m=1))+(1-\lambda)\sum_{j=1}^{k}H(p(b_j^m=1)) \tag{8}$$

$$H(p(x))=-\sum_{i=1}^{d}\sum_{x_i=0}^{1}p(x_i)\log_2 p(x_i) \tag{9}$$

基于熵模型的图聚集实现了属性/结构的相似一致性,对于含有属性的图而言,该方法能够有效地度量属性和结构的信息量.但其并未计算出属性和结构的相应权重,需要用户自己定义,且不能自动调节所有边的权重.

文献[30]通过分组标准 I 和 III,确保各分组的属性以及关联强度一致,并基于直观上较好的超图所满足的特

性,推导出包含多样性、简洁性、覆盖性和实用性在内的超图质量函数 $Quality(G_\phi)$,相应的公式如下所示:

$$Quality(G_\phi)=\frac{Diversity(G_\phi)}{Coverage(G_\phi)\times Conciseness(G_\phi)\times Utility(G_\phi)} \tag{10}$$

基于 Jaccard 相似度,利用 LSH(locality sensitive hashing)划分技术,使同一组内顶点的属性集尽量相似,并降低时间复杂度至 $O(|V|\log|V|+E)$.实现了多样性和简洁性,基于熵度量关联一致性,使用 *K*-Means 算法进一步分组,尽量实现关联一致性满足覆盖性和实用性.但整体上看,质量函数和计算过程并不紧密关联,并未体现出 4 种特性在同一情况下的整体度量.

2.3.2 基于多粒度的图聚集

文献[27]提出的 SNAP 算法由用户选定属性和关联类型,其超图中各分组虽然满足分组标准 I 和 II,实现属性和邻接分组一致.但由于实际数据的噪音和不确定关联,导致其效果并不好.由此提出改进算法 *k*-SNAP,该方法放宽了邻接分组的一致性,并给定分组数目 *k* 控制超图的大小.通过 Top-Down 方法,由属性一致性分组分裂到属性和邻接分组相似一致的 *k* 个分组,实现了 OLAP 中的下钻过程.也可通过 Bottom-Up 方法由属性和邻接分组一致的分组合并到属性和邻接分组相似一致的 *k* 个分组,实现了 OLAP 中的上卷过程.但 *k*-SNAP 中仍然存在如下一些问题[29,31]:

1) *k*-SNAP 仅适用于属性域较小的情况,这是因为 *k*-SNAP 中要求各分组的属性值需保持一致,当选定的属性的值域较大时,如链接的引用次数,利用度量 CT 基于邻接分组一致性分裂成 *k* 个分组时,分组的数据在属性上存在严重的倾斜;

2) *k*-SNAP 能够实现多粒度的图聚集,但是用户往往需要计算出很多超图,才能得到自己感兴趣的结果.

文献[29]针对上述问题进行了一些改进:针对问题 1),它通过算法 CANAL 自动划分域较大的属性的值,将分组的属性值一致转换为属性值域一致.该算法将属性值相邻的任意两分组,基于两分组的邻接分组一致性度量 SimLink 进行合并后,通过合并前后Δ的转变率的度量 u_p,获取合并后使得超图的性能最差的两邻接分组间的属性值边界线,进而将原属性域划分成 *C*(期望分组数)个区间;针对问题 2),通过给出的度量 *Interestingness*(*S*),粗略地衡量了用户感兴趣的因素,包括超图信息的多样性(diversity)、覆盖性(coverage)以及简洁性(conciseness),使得超图在具有实际意义的同时含有较多的信息.

表 2 总结了上文多种图聚集方法的特点:

- 基于属性的图聚集技术研究的多层次多维度的方法,包含 Graph OLAP 和 Graph Cube,仅满足分组标准 I;
- 基于结构的图聚集技术主要针对超图还原的概率和最大化压缩原图两方面进行研究;
- 基于属性/结构的图聚集则利用熵模型和 OLAP 方式,在属性一致性上均满足分组标准 I,但对于结构一致的标准最多能满足标准 III;
- 部分算法能够满足分组标准 V,实现无损图聚集.

Table 2 Collection of graph aggregation and summarization

表 2 图聚集技术汇总

类别		模型	分组标准					特点
			I	II	III	IV	V	
属性一致性		Graph OLAP	√					快照网络作为图模型; I-OLAP 和 T-OLAP;多层次,多维度分析
		Graph Cube	√					构建 Graph Cube, Cuboid Query 和 Crossboid Query
结构一致性	概率	Expected-Adjacency matrix					√	较准确地在超图上回答基于原图的查询
	最大化压缩	S-Node		√				考虑了聚集图的分组数目
		MDL	√					加强构建超边的条件;可实现有损压缩; 利用自连接边压缩团结构

Table 2 Collection of graph aggregation and summarization (continued)

表 2 图聚集技术汇总(续)

类别		模型	分组标准					特点
			I	II	III	IV	V	
结构/属性一致性	熵模型	Entropy	√		√			构建属性顶点和边,使用熵统一度量属性和结构信息
		Jaccard	√	√	√			$O(\|V\|\log\|V\|+E)$的时间复杂度,聚集图满足多样性、覆盖性、简洁性、实用性
	多粒度	SNAP/*k*-SNAP	√	√				用户选定分组数目;实现多粒度的图聚集,实现 OLAP 的上卷和下钻
		CANAL						给值域较广的属性划分区间;使用度量函数粗略筛选用户感兴趣的超图

3 结 论

本文研究和分析了当前图聚集的技术及算法,给出了 5 种分组标准,并根据各分组标准对应的关系将图聚集划分成基于属性一致性、基于结构一致性和同时基于属性和结构一致性的图聚集这 3 类.接着,概括性地描述了当前主要的图聚集算法.

随着图数据的不断快速增长以及图数据应用的多样化,在图聚集和可视化领域将面临越来越多的挑战:

- 可扩展的图聚集

图聚集技术虽然是用于解决大图中可扩展的图挖掘问题,但是如今针对高度可扩展的图聚集算法仍然存在很多问题.大多数工作还停留在基于内存的图聚集算法上,因此有必要研究基于磁盘的图聚集算法,或者参考并行处理模型,如 MapReduce 等.

- 不确定图的图聚集技术

实际应用中,有些图数据存在关联不确定的问题,在处理此类图聚集问题时,可考虑引入概率这一概念.对图中顶点间不确定的关系使用概率计算其期望值,得到最接近实际情况的关联值.

- 属性和结构相融合的图聚集

大多数基于属性和结构一致性的图聚集技术均通过信息论中熵的模型来实现,但是熵并不能完全解决属性和结构相矛盾的问题.属性和结构相融合的图聚集之所以困难,就在于无法确定两者之间的权重,很难将它们放在同一个平台上去考虑.所以,同时基于属性和结构一致性的图聚集仍然需要细致而深入地加以研究.

- 多图及异构图的聚集和概括

现有图聚集技术的研究对象均是单个图,实际上,针对多个图进行研究也是一个新的研究点.当前,已出现了诸如获取包含相同查询子图的多个图聚集[32]以及使用单个图概括多个图的结构信息[33]等.异构图由于顶点以及边的类型都可能不相同,对其进行聚集和概括将是一个挑战.

致谢 衷心感谢审稿专家对本文提出的宝贵意见和建议.

References:

[1] Aggarwal CC, Wang H. Managing and Mining Graph Data. Springer-Verlag, 2010. [doi: 10.1007/978-1-4419-6045-0]

[2] Chakrabarti D, Faloutsos C. Graph mining: Laws, generators, and algorithms. ACM Computing Surveys, 2006,38(1):2. [doi: 10.1145/1132952.1132954]

[3] Leskovec J, Kleinberg J, Faloutsos C. Graphs over time: Densification laws, shrinking diameters and possible explanations. In: Proc. of the 11th ACM SIGKDD (KDD 2005). New York: ACM Press, 2005. 177−187. [doi: 10.1145/1081870.1081893]

[4] Feng GD, Xiao YH. Distributed storage of big graphs. Communications of the CCF, 2012,8(11):12−15 (in Chinese with English abstract).

[5] Rodrigues JF, Traina AJM, Faloutsos C, Traina Jr C. SuperGraph visualization. In: Proc. of the 8th IEEE Int'l Symp. on Multimedia. 2006. 227−234. [doi: 10.1109/ISM.2006.143]

[6] Hay M, Miklau G, Jensen D, Weis P. Resisting structural re-identification in anonymized social networks. In: Proc. of the VLDB. 2008. [doi: 10.1007/s00778-010-0210-x]

[7] Adler M, Mitzenmacher M. Towards compressing Web graphs. In: Proc. of the Data Compression Conf. 2001. 203−212. [doi: 10.1109/DCC.2001.917151]

[8] Boldi P, Vigna S. The Web graph framework I: Compression techniques. In: Proc. of the WWW. 2004. 595−602. [doi: 10.1145/988672.988752]

[9] Suel T, Yuan J. Compressing the graph structure of the Web. In: Proc. of the Data Compression Conf. 2001. 213−222. [doi: 10.1109/DCC.2001.917152]

[10] Raghavan S, Garcia-Molina H. Representing Web graphs. In: Proc. of the ICDE. 2003. 405−416. [doi: 10.1109/ICDE.2003.1260809]

[11] LeFevre K, Terzi E. GraSS: Graph structure summarization. In: Proc. of the SDM 2010. 2010. 454−465. [doi: 10.1137/1.9781611972801.40]

[12] Liu Z, Yu JX. On summarizing graph homogeneously. In: Proc. of the Database Systems for Adanced Applications. 2011. 299−310. [doi: 10.1007/978-3-642-20244-5_29]

[13] Liu Z, Yu JX, Cheng H. Approximate homogeneous graph summarization. Information Processing Society of Japan, 2011,20(1): 77−88. [doi: 10.11185/imt.7.32]

[14] Toivonen H, Zhou F, Hartikainen A, Hinkka A. Compression of weighted graphs. In: Proc. of the 17th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2011. [doi: 10.1145/2020408.2020566]

[15] Zhou F. Methods for Network Abstraction. University of Helsinki, 2012.

[16] Zhou Y, Cheng H, Yu JX. Graph clustering based on structural/attribute similarities. Proc. of the VLDB Endowment, 2009,2(1). [doi: 10.14778/1687627.1687709]

[17] Abou-Rjeili A, Karypis G. Multilevel algorithms for partitioning power-law graphs. In: Proc. of the IPDPS 2006. 2006. [doi: 10.1109/IPDPS.2006.1639360]

[18] Navlakha S, Rastogi R, Shrivastava N. Graph summarization with bounded error. In: Proc. of the 2008 ACM-SIGMOD Int'l Conf. on Management of Data (SIGMOD 2008). Vancouver, 2008. 419−432. [doi: 10.1145/1376616.1376661]

[19] Toivonen H, Zhou F, Hartikainen A, Hinkka A. Network compression by node and edge mergers. In: Proc. of the Bisociative Knowledge Discovery. Berlin, Heidelberg: Springer-Verlag, 2012. 199−217. [doi: 10.1007/978-3-642-31830-6_14]

[20] Álvarez S, Brisaboa NR, Ladra S, Pedreira Ó. A compact representation of graph databases. In: Proc. of the 8th Workshop on Mining and Learning with Graphs. ACM Press, 2012. 18−25. [doi: 10.1145/1830252.1830255]

[21] Chen C, Yan X, Zhu F, Han J, Yu PS. Graph OLAP: Towards online analytical processing on graphs. In: Proc. of the ICDM. 2008. 103−112. [doi: 10.1109/ICDM.2008.30]

[22] Li C, Yu PS, Lin W. InfoNetOLAPer: Integrating InfoNetWarehouse and InfoNetCube with InfoNetOLAP. PVLDB, 2011,4(12): 1422−1425. [doi: 10.1109/CSSE.2008.667]

[23] Qu Q, Zhu F, Yan X, Han J, Yu PS, Li H. Efficient topological OLAP on information networks. In: Proc. of the DASFAA 2011. Hong Kong, 2011. 389−403. [doi: 10.1007/978-3-642-20149-3_29]

[24] Li C, Zhao L, Tang CJ, Chen Y, Li J, Zhao XM, Liu XL. Modeling, design and implementation of graph OLAPing. Ruan Jian Xue Bao/Journal of Software, 2011,22(2):258−268 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3771.htm [doi: 10.3724/SP.J.1001.2011.03771]

[25] Zhao P, Li X, Xin D, Han J. Graph cube: On warehousing and OLAP multidimensional networks. In: Proc. of the SIGMOD 2011. 2011. 12−16. [doi: 10.1145/1989323.1989413]

[26] Hassanlou N. Probabilistic graph summarization. University of Victoria, 2012. [doi: 10.1007/978-3-642-38562-9_55]

[27] Tian Y, Hankins RA, Patel JM. Efficient aggregation for graph summarization. In: Proc. of the 2008 ACM-SIGMOD Int'l Conf. on Management of Data (SIGMOD 2008). Vancouver, 2008. 567−580. [doi: 10.1145/1376616.1376675]

[28] Tian Y, Patel JM. Interactive Graph summarization. In: Proc. of the Link Mining: Models, Algorithms, and Applications. 2010. 389−409. [doi: 10.1007/978-1-4419-6515-8_15]

[29] Zhang N, Tian Y, Patel JM. Discovery-Driven graph summarization. In: Proc. of the 2010 IEEE 26th Int'l Conf. on Data Engineering (ICDE). IEEE, 2010. [doi: 10.1109/ICDE.2010.5447830]

[30] Yin D, Gao H, Zou Z. A novel efficient graph aggregation algorithm. Journal of Computer Research and Development, 2011,48(10): 1831−1841 (in Chinese with English abstract).

[31] Louati A, Aufaure MA, Lechevallier Y. Graph aggregation: Application to social networks. In: Advances in Theory and Applications of High Dimensional and Symbolic Data Analysis. Hermann, 2013. 157−177.

[32] Zou L, Chen L, Zhang HM, Lu YS, Lou Q. Summarization graph indexing: Beyond frequent structure-based approach. In: Proc. of the Database Systems for Advanced Applications. Berlin, Heidelberg: Springer-Verlag, 2008. 141−155. [doi: 10.1007/978-3-540-78568-2_13]

[33] Endriss U, Grandi U. Graph aggregation. In: Proc. of the 4th Int'l Workshop on Computational Social Choice (COMSOC 2012). 2012.

附中文参考文献:

[4] 冯国栋,肖仰华.大图的分布式存储.中国计算机学会通讯,2012,8(11):12−15.

[24] 李川,赵磊,唐常杰,陈瑜,李靓,赵小明,刘小玲.Graph OLAPing 的建模、设计与实现.软件学报,2011,22(2):258−268. http://www.jos.org.cn/1000-9825/3771.htm [doi: 10.3724/SP.J.1001.2011.03771]

[30] 尹丹,高宏,邹兆年.一种新的高效图聚集算法.计算机研究与发展,2011,48(10):1831−1841.

潘秋萍(1989－),女,江苏江都人,硕士,主要研究领域为数据挖掘.

董朋志(1989－),男,硕士,主要研究领域为数据仓库,并行计算.

游进国(1977－),男,博士,副教授,CCF 会员,主要研究领域为数据仓库,数据挖掘,并行计算.

胡宝丽(1989－),女,硕士,主要研究领域为数据挖掘.

张志朋(1988－),男,硕士,主要研究领域为数据仓库,并行计算.