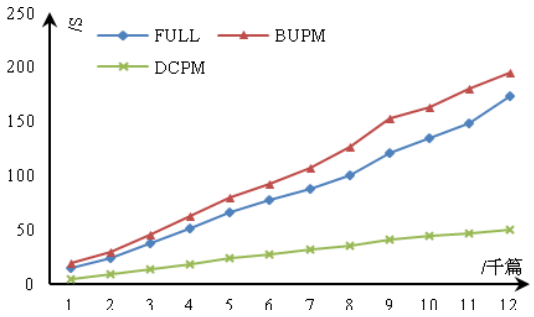

Fig.4 Significant edge based partial materialization time consuming of the InfoNetCuboid (ACM)

图 4 基于显著性边的信息网络单元物化时间比较(ACM)(δ=2)

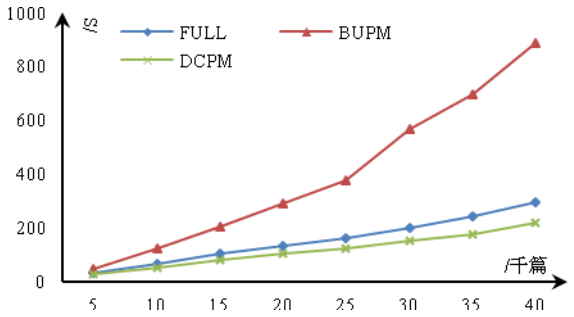

Fig.5 Significant edge based partial materialization time consuming of the InfoNetCuboid (MAG)

图 5 基于显著性边的信息网络单元物化时间比较(MAG)(δ=2)

图 4 和图 5 分别展示了两组数据集上基于显著性边度量(频繁的合作关系)通用度量模型的信息网络方体物化算法时间对比.实验结果表明,完全物化操作所需的计算时间远大于本文提出的基于透析计算的部分物化算法所需要时间.此外,从图中可以看出:随着数据量的增加,本文提出的部分物化策略所需的操作时间增长相对比较缓慢,因此,算法具有较好的扩展性,可以有效应对大规模场景下的应用需求.

考虑到存在部分用户可能对于合作者网络中特定合作模式感兴趣,本文设计了一组基于用户兴趣模型,即中心作者合作关系(如星形结构)的信息网络方体物化实验方案.实验结果如图 6、图 7 所示.

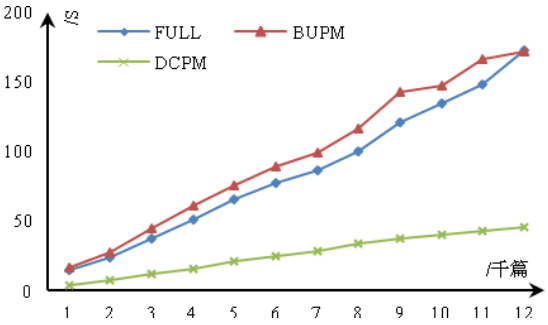

Fig.6 Staring central author based partial materialization time comparision (ACM)

图 6 基于星形中心作者的信息网络方体物化时间比较(ACM)

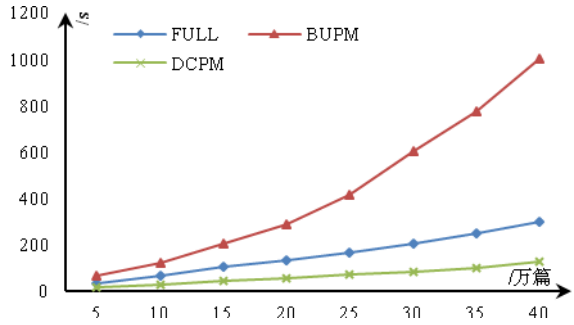

Fig.7 Staring central author based partial materialization time comparision (MAG)

图 7 基于星形中心作者的信息网络方体物化时间比较(MAG)

实验结果表明:在不同的用户兴趣度模型下,基于透析计算的部分物化策略可以有效降低信息网络方体的计算时间开销.分别对图 4~图 7 中的基于基本方体的完全物化策略运行时间和基于透析计算的部分物化策略运行时间的降低百分比取均值,可得出部分物化较基于基本方体的部分物化策略运行效率平均降低 75%,从而表明了本文提出的透析计算策略的有效性.

通过对算法 2 的分析可知:基于基本方体的部分物化策略(BUPM)是在完全物化操作基础上执行的剪枝操作,即,算法的运行时间一定比完全物化策略的运行时间长.本文的实验结果也反映了这一事实.

5.3.2 方体格物化时间

根据第 2 节设计的信息网络方体格体系结构,本文通过实验分别计算了不同计算策略下、不同拓扑维层级执行部分物化操作所需的时间开销,实验结果如图 8、图 9 所示.

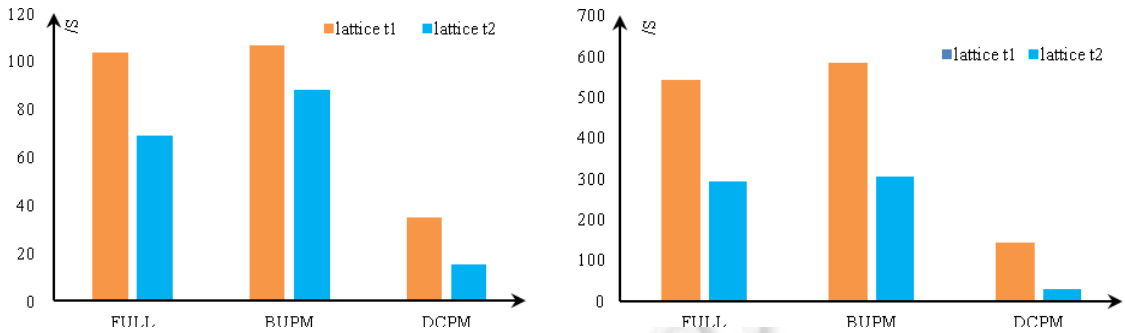

Fig.8 General measure based InfoNetLattice partial materialization time comparision

图 8 基于通用度量模型的信息网络方体格物化时间对比

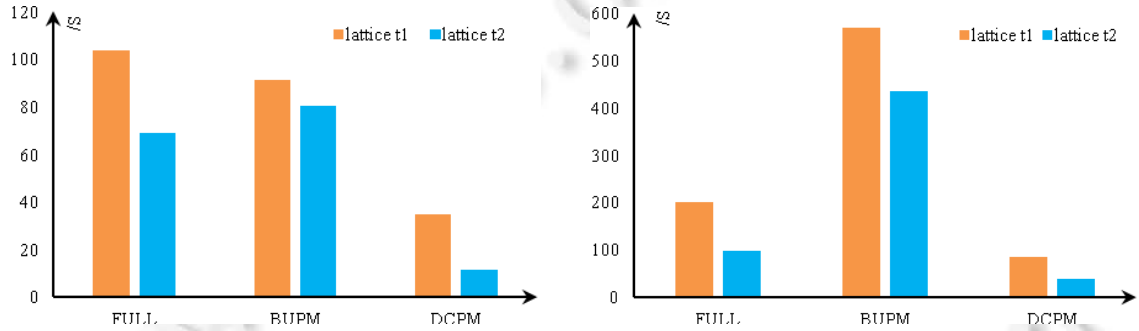

Fig.9 General measure based InfoNetLattice partial materialization time comparision

图 9 基于用户兴趣度模型的信息网络方体格物化时间对比

从图 8、图 9 可以看出:在通用度量模型和用户兴趣度模型中,基于透析计算的部分物化策略不同方体格均具有较高的计算效率,这得益于不同方体格之间和不同信息维层级之间提供的有效剪枝信息.

值得注意的是:在基于用户兴趣度的模型中,基于基本方体的部分物化策略的不同方体格之间的计算时间差距相对较小.这是因为网络中符合该带兴趣度度量的子图模式通常是不平凡的,即,在不同拓扑维层次上的表现出较强的相似性,因此,在不同方体格上的运行时间差距不如通用度量模型下显著.

6 相关工作

Chen 等人[17]首先提出了基于图的在线分析处理,对信息维、拓扑维等基本概念进行了定义,给出了相应的框架设计;该文的扩展版本[18,19]对该框架的技术路线进行详细阐述,但仍然是围绕 I-OLAP 和 T-OLAP 的操作来讨论,并未提出新的观点或相关技术.Qu 等人[20]对拓扑维的在线分析处理做了进一步的讨论,对 T-OLAP 操作中特定类型的度量做了重点分析,但未给出实际的算法与实现;同时,并未对广义的 T-OLAP 指明设计与实现方向.Zhao 等人[21]通过对真实网络进行抽象,提出了 Graph Cube 多维网络模型.但该模型本质上仍与传统数据立方体一致,针对拓扑维的相关操作无法与信息维的相关操作进行有效区分,无法满足信息网络在线分析处理的需求.Li 等人[14]首次提出了以 Graph 数据为中心度量的 OLAP 的概念,文献提出了基于图的数据立方概念和创建过程,但对于信息网络方体格的设计与实现未给出具体的解决方案.Jin 等人[22]提出了一种适用于在线图分析处理的 VisualCube 模型,但该模型只考虑了信息维,无法解决信息网络环境下拓扑维引入所带来的复杂问题.Nie 等人[23]利用维建模的方法对基于图的信息网络数据进行模型设计,实现了多维信息网络仓库模型,为信息网络的在线分析处理提供了必备的基础设施,但关注的焦点仍然集中在原始数据的底层存储结构,该存储结构解决的书物理存储问题,未涉及信息网络的逻辑存储结构进行上层分析处理所需要的技术路线.

Xu 等人[24]首次提出了面向信息网络的在线图处理模型,设计并实现了相应的基本操作算法,但对于基本方体的处理仍然是采用完全物化计算策略,在维度数量较高的情况下,难以应对计算过程中时间和空间方面剧增的问题.Morfonios 等人[25]通过对社交编著系统的分析,提出了一种针对实体聚集视图的查询搜索框架,并采用完全物化的策略来对所有可能的查询需求进行预计算,显然,这种策略的空间开销过于庞大.

Yin 等人[26]指出,Chen 的模型只适用于同构网络,提出了一种适用于异构信息网络建模的 HMGraphCube 模型,但异构网络的复杂度远高于同构网络,因此,如何对异构信息网络进行高效预计算是不可避免的问题.文献关注的焦点仍然是集中在对新操作的讨论,未提及物化的相关问题.Beheshti 等人[27]指出:当前的信息网络在线分析处理模型过分关注基于图的在线查询和分析处理,并不能很好地支持基于语义驱动的计算,因此提出了一种基于图的拓扑结构进行决策的 GOLAP 模型.但该模型的焦点仍然集中在操作层面,未考虑海量数据场景下的操作效率问题.Wang 等人[28]针对当前 GraphOLAP 模型在分析处理异构信息网络方面存在的效率不足问题,提出了一种面向大规模网络的多维分析处理框架.该框架关注焦点仍然集中在信息网络数据立方的建模,并在此基础上引入了两种新的操作.虽然文献提及了一种基于 2-源路径的部分物化策略,但作者未给出详细的算法描述和实施方案.此外,该方案本质上是为新操作服务的,应用场景相对比较局限.Wang 等人[29]提出了一种适用于大规模信息网络的并行图分析处理框架,虽然该模型可以高效地处理大规模网络图结构,但该模型面向的是基于拓扑维的处理场景,即:更多的关注点集中在网络拓扑结构的变化上,未考虑到信息维的相关信息,本质上仍属于 OLGP 较具体的一组应用场景.

Sabine 等人[30]通过对 OLAP 以及著作数据的调研,对信息网络在线分析处理的应用场景进行了分析和归纳,虽然指出了几个可能的研究方向,但是未对信息网络立方的物化工作方向进行预测.同时,近年来也鲜有相关工作涉及具体的部分物化技术和处理思路.

7 总结与展望

本文借鉴医学的透析原理,首次提出了基于透析计算的信息网络数据立方 InfoNetCube 剪枝策略和部分物化原理.本文贡献可总结为:

(1) 提出了信息网络方体格的概念,提出了 InfoNetLattice 外部体系结构和内部体系结构.提出了信息网络方体及信息网络方体单元在信息网络立方体中的结构特点和联系;

(2) 提出了基于透析计算的 InfoNetCube 部分物化策略,设计和实现了相应的算法.实验结果表明,该部分物化策略可以有效降低信息网络方体物化过程的计算时间和空间开销.

本文在基于用户兴趣度度量模型的实验中,算法按照传统的 DFS 策略执行时间相对较长.如何充分利用网络提供的信息来提高算法的子图模式匹配效率,是本文后续工作中需要解决的问题.正如相关工作中提到的,本文的研究场景仍然是基于同构网络,但是本文所设计的信息网络方体格体系结构具有很好地适用性,后续的研究工作将考虑将算法应用场景扩展到异构信息网络上.

References:

[1] Han JW, Yan XF, Yu PS. Scalable OLAP and mining of information networks. In: Proc. of the 12th Int'l Conf. on Extending Database Technology: Advances in Database Technology. ACM Press, 2009. [doi: 10.1145/1516360.1516505]

[2] Han JW, Sun Y, Yan X, Yu PS. Mining knowledge from databases: An information network analysis approach. In: Proc. of the 2010 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2010. [doi: 10.1145/1807167.1807333]

[3] Han JW. Mining heterogeneous information networks by exploring the power of links. In: Proc. of the Int'l Conf. on Discovery Science. Berlin, Heidelberg: Springer-Verlag, 2009. [doi: 10.1007/978-3-642-04747-3_2]

[4] Aggarwal CC, Wang HX, eds. Managing and Mining Graph Data. Vol.40. New York: Springer-Verlag, 2010.

[5] Newman MEJ. Networks: An Introduction. Oxford University Press, 2010.

[6] Gray J, Chaudhuri S, Bosworth a, Layman A, Reichart D, Venkatrao M, Pirahesh H, Pellow F. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. Data Mining and Knowledge Discovery, 1997,1(1):29−53. [doi: 10.1023/A:1009726021843]

[7] Chaudhuri S, Dayal U. An overview of data warehousing and OLAP technology. ACM Sigmod Record, 1997,26(1):65−74.

[8] Sarawagi S, Agrawal R, Megiddo N. Discovery-Driven exploration of OLAP data cubes. In: Proc. of the Int'l Conf. on Extending Database Technology. Berlin, Heidelberg: Springer-Verlag, 1998. 168−182. [doi: 10.1007/BFb0100984]

[9] Sun YZ, Wu TY, Yin ZJ, Cheng H, Han JW, Yin XX, Zhao PX. BibNetMiner: Mining bibliographic information networks. In: Proc. of the 2008 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2008. [doi: 10.1145/1376616.1376770]

[10] Burdick D, Doan A, Ramakrishnan R, Vaithyanathan S. OLAP over imprecise data with domain constraints. In: Proc. of the VLDB. 2007. 39−50.

[11] Morfonios K, Konakas S, Ioannidis Y, Kotsis N. ROLAP implementations of the data cube. ACM Computing Surveys (CSUR), 2007,39(4):12. [doi: 10.1145/1287620.1287623]

[12] Zhang N, Tian Y, Patel JM. Discovery-Driven graph summarization. In: Proc. of the ICDE. 2010. 880−891. [doi: 10.1109/ICDE.2010.5447830]

[13] LI C, Yu PS, Zhao L, Xie Y, Lin WQ. InfoNetOLAPer: Integrating InfoNetWarehouse and InfoNetCube with InfoNetOLAP. In: Proc. of the VLDB 2011. Demo, 2011.

[14] 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]

[15] Pei J, Chai W, Zhao C, Tang SW, Yang DQ. An algebra for online analytical processing data cube. Ruan Jian Xue Bao/Journal of Software, 1999,10(6):561−568 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/10/561.htm

[16] Yang Y. The principles and developments of the hemodialysis system. Chinese Journal of Medical Instruction, 2001,25(5):288−296 (in Chinese with English abstract).

[17] Chen C, Yan XF, Zhu FD, Han JW, Yu PS. Graph OLAP: Towards online analytical processing on graphs. In: Proc. of the Int'l Conf. on Data Mining (ICDM 2008). 2008. [doi: 10.1109/ICDM.2008.30]

[18] Chen C, Zhu F, Yan XF, Han JW, Yu P, Ramakrishnan R. InfoNetOLAP: OLAP and mining of information networks. In: Yu PS, Faloutsos C, Han JW, eds. Proc. of the Link Mining: Models, Algorithms and Applications. Springer-Verlag. [doi: 10.1007/978-1-4419-6515-8_16]

[19] Chen C, Yan XF, Zhu FD, Han JW. Graph OLAP: A multi-dimensional framework for graph data analysis. Knowledge and Information Systems (KAIS), 2009,21(1):41−63. [doi: 10.1007/s10115-009-0228-9]

[20] Qu Q, Zhu FD, Yan XF, Han JW, Yu PS, Li HY. Efficient topological OLAP on information networks. In: Proc. of the Int'l Conf. on Database Systems for Advanced Applications. Berlin, Heidelberg: Springer-Verlag, 2011. [doi: 10.1007/978-3-642-20149-3_29]

[21] Zhao PX, Aggarwal C, Wang M. gSketch: On query estimation in graph streams. In: Proc. of the 38th Int'l Conf. on Very Large Data Bases (VLDB 2012). Istanbul, 2012.

[22] Jin X, Han JW, Cao LL, Luo JB, Ding BL, Lin CX. Visual cube and on-line analytical processing of images. In: Proc. of the 19th ACM Int'l Conf. on Information and Knowledge Management. ACM Press, 2010. [doi: 10.1145/1871437.1871546]

[23] Nie ZY, Li C, Tang CJ, Xu HY, Zhang YH, Yang N. Design of multi-dimensional information network data warehouse model for online graph processing. Journal of Frontiers of Computer Science and Technology, 2014,8(1):51−60 (in Chinese with English abstract).

[24] Xu HY, Li C, Tang CJ, Li YT, Dai SC, Yang N. On-Line graphic processing: Information network oriented on-line analytical processing. Journal of Frontiers of Computer Science and Technology, 2012,6(9):797−809 (in Chinese with English abstract).

[25] Morfonios K, Koutrika G. OLAP cubes for social searches: Standing on the shoulders of giants? In: Proc. of the WebDB. 2008.

[26] Yin M, Wu B, Zeng ZF. HMGraph OLAP: A novel framework for multi-dimensional heterogeneous network analysis. In: Proc. of the 15th Int'l Workshop on Data Warehousing and OLAP. ACM Press, 2012. [doi: 10.1145/2390045.2390067]

[27] Beheshti SMR, Benatallah B, Motahari-Nezhad HR, Allahbakhsh M. A framework and a language for on-line analytical processing on graphs. In: Proc. of the Int'l Conf. on Web Information Systems Engineering. Berlin, Heidelberg: Springer-Verlag, 2012. [doi: 10.1007/978-3-642-35063-4_16]

[28] Wang PS, Wu B, Wang B. TSMH graph cube: A novel framework for large scale multi-dimensional network analysis. In: Proc. of the IEEE Int'l Conf. on Data Science and Advanced Analytics (DSAA 2015). Vol.36678. IEEE, 2015. [doi: 10.1109/DSAA.2015.7344826]

[29] Wang ZK, Fan Q, Wang HJ, Tan KL, Agrawal D, Abbadi AE. Pagrol: Parallel graph olap over large-scale attributed graphs. In: Proc. of the 2014 IEEE 30th Int'l Conf. on Data Engineering. IEEE, 2014. [doi: 10.1109/ICDE.2014.6816676]

[30] Loudcher S, Jakawat W, Morales EPS, Favre C. Combining OLAP and information networks for bibliographic data analysis: A survey. Scientometrics, 2015,103(2):471−487. [doi: 10.1007/s11192-015-1539-0]

附中文参考文献:

[14] 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]

[15] 裴健,柴玮,赵畅,唐世渭,杨冬青.联机分析处理数据立方体代数.软件学报,1999,10(6):561−569. http://www.jos.org.cn/1000-9825/10/561.htm

[16] 杨焱.血液透析系统的基本原理及发展.中国医疗器械杂志,2001,25(5):288−291.

[23] 聂章艳,等.面向 OLGP 的多维信息网络数据仓库模型设计.计算机科学与探索,2014,8(1):51−60.

[24] 徐洪宇,李川,唐常杰,徐洪宇,张永辉,杨宁.在线图处理:面向信息网络的在线分析处理.计算机科学与探索,2012,6(9):797−809.

刘光明(1989−),男,山东临沂人,硕士生,主要研究领域为数据库,数据挖掘,信息网络.

任艳(1983−),女,工程师,主要研究领域为电子产品数据资源建设与应用.

李川(1977−),男,博士,副教授,CCF 专业会员,主要研究领域为数据库,数据挖掘,信息网络数据分析,社会网络分析,生物信息学.

杨宁(1974−),男,博士,讲师,CCF 专业会员,主要研究领域为时空数据挖掘,时态序列挖掘,异构信息网络挖掘,网络上的信息处理.

唐常杰(1946−),男,教授,博士生导师,CCF 杰出会员,主要研究领域为数据库系统,数据挖掘和数据仓库,知识工程,计算机安全.