提供重要依据.例如,文献[45]在对包含 1.8 万个蛋白质(节点)和 4.4 万种蛋白质之间的相互关系(边)的蛋白质交互作用网络进行研究的过程中,通过计算蛋白质交互作用网络的连通性、中心性以及匹配已知性质的结构,发现介数中心性(betweenness)大且连联通性小的蛋白质在人类基因中是冗余的蛋白质,这无疑为分析基因和蛋白质功能提供了重要的依据.

6 结论与展望

动态图作为目前广泛使用的数据模型,对动态图数据匹配方面的研究具有十分重要的理论意义和应用前景,引起了学术和工业界越来越多的关注.随着大数据时代的到来,数据规模的激增,数据更新越发频繁,数据之间的关系越发复杂,这无疑给动态图数据匹配技术的研究和应用提出了新的挑战,也产生了新的机遇,可以从以下几个方面开展研究.

- 面对图数据规模激增的现状,可以研究如何基于主流的分布式并行图处理框架系统,例如谷歌的 Pregel 和 Apache 的 Giraph,来实现图匹配的并行分布式处理,提高匹配技术应对大规模图数据的可扩展性.这些主流的分布式并行图处理框架系统通常采用以节点为中心的处理模式.以节点为中心的处理模式的基本计算单元是节点,其将数据图进行划分,由计算框架中的每个节点存储和处理一个分区的图数据.文献[29]还提出了一种动态图处理框架 BLADYG,提出用以块为中心的分布式并行计算模型来处理大规模数据图.中心块计算模型的基本计算单元是块,每一个块存储和处理的是数据图的一个连通子图.这两种分布式并行图处理模型各有优劣,研究人员可以根据应用需求灵活选择;
- 面对图数据更新越发频繁的现状,采用快照的匹配方法很难满足动态图匹配的实时性需求,因此需要研究高效的面向动态图特点的匹配方式和查询优化策略.当前,研究者可以从增量算法设计和候选集剪枝两方面进行研究,从而缩小查询范围,提高匹配效率;
- 基于连接的匹配技术是实现动态图模式匹配的有效手段,从模式图或数据图中提取优质的特征,是该技术面临的核心问题之一.为了避免发生数据图更新导致特征失效的情况,目前的研究工作一般选择单边子图或双边子图这类简单子图作为数据图的特征.然而,这类简单子图的可识别特性不强,导致中间匹配结果激增,加重了子图连接阶段的处理代价.因而,如何提取更加有识别度的特征,同时设计高效的特征维护方案,是当前研究的主要问题.另一方面,基于探索的动态图匹配方式有其特有的优势,但却面临着匹配结果准确率偏低的问题,可以研究基于探索的动态图匹配技术与基于连接的动态图匹配技术相结合的新型处理方案,重点考虑结合过程中如何发挥各自技术类型的优势,从而弥补各自技术类型的缺憾.基于模拟的匹配技术是当前研究的热点,它避免了复杂的子图同构匹配计算,对模拟匹配技术进行扩展研究,将是另一个研究关注点.另外,设计面向应用需求的图匹配相似度度量模型、开发更有效的图匹配近似算法,也是未来的研究方向;
- 目前,大部分研究涉及的动态图数据匹配问题都是针对模式图不变、数据图随时间动态变化的情况,然而在实际生活中,模式图发生变化的情况也非常普遍.例如:网络安全领域,病毒经常发生变种,网络攻击模式也在不断发生演变;在计算生物学领域,蛋白质变性、流感病毒变异也时常发生.文献[38]研究了数据图不变、模式图发生动态变化时的图模式匹配增量算法.研究者们一方面可以开拓研究这方面的增量处理技术;另一方面,也可以将模式图变化与数据图变化结合起来,拓展动态图数据匹配技术的应用领域.

References:

[1] Gong NZ, Xu W, He Y, Huang L, Mittal P, Stefanov E, Sekar V, Song D. Evolution of social-attribute networks: Measurements, modeling, and implications using Google+. In: Proc. of the 2012 ACM Internet Measurement Conf. Libraries: ACM Press, 2012. 131−144. [doi: 10.1145/2398776.2398792]

[2] Fan W, Wang X, Wu Y. Fan W, Wang X, Wu Y. ExpFinder: Finding experts by graph pattern matching. In: Proc. of the 29th Int'l Conf. on Data Engineering. Brisbane: IEEE Computer Society, 2013. 1316−1319. [doi: 10.1109/ICDE.2013.6544933]

[3] Choudhury S, Holder L, Chin G, Ray A, Beus S, Feo J. StreamWorks: A system for dynamic graph search. In: Proc. of the 2013 ACM SIGMOD Int'l Conf. on Management of Data. NewYork: ACM Press, 2013. 1101−1104. [doi: 10.1145/2463676.2463697]

[4] Yu J, Liu YB, Zhang Y, Liu MY, Tan JL, Guo L. Survey on large-scale graph pattern matching. Journal of Computer Research and Development, 2015,52(2):391−409 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2015.20140188]

[5] Lee J, Han WS, Kasperovics R, Lee JH. An in-depth comparison of subgraph isomorphism algorithms in graph databases. Proc. of the VLDB Endowment, 2013,6(2):133−144. [doi: 10.14778/2535568.2448946]

[6] Malewicz G, Austern MH, Bik AJC, Dehnert JC, Horn I, Leiser N. Pregel: A system for large-scale graph processing. In: Proc. of the 2010 ACM SIGMOD Int'l Conf. on Management of Data. Indianapolis: ACM Press, 2010,18(18):135−146. [doi: 10.1145/1807167.1807184]

[7] Shao B, Wang H, Li Y. Trinity: A distributed graph engine on a memory cloud. In: Proc. of the 2013 ACM SIGMOD Int'l Conf. on Management of Data. NewYork: ACM Press, 2013. 505−516. [doi: 10.1145/2463676.2467799]

[8] Apache girph. 2017. http://incubator.apache.org/giraph

[9] Yan D, Cheng J, Lu Y, Ng W. Blogel: A block-centric framework for distributed computation on real-world graphs. Proc. of the VLDB Endowment, 2014,7(14):1981−1992. [doi: 10.14778/2733085.2733103]

[10] Fan WF, Xu JB, Wu YH, Yu WY, Jiang JX, Zheng ZY, Zhang B, Cao Y, Tian C. Parallelizing sequential graph computations. In: Proc. of the 2017 ACM SIGMOD Int'l Conf. on Management of Data. Chicago: ACM Press, 2017. 495−510. [doi: 10.1145/3035918. 3035942]

[11] Postech Database Lab. NASA, Yeast, and Human datasets. 2014. http://dtp.nci.nih.gov/docs/aids/aids_data.html

[12] Wang C, Chen L. Continuous subgraph pattern search over graph streams. In: Proc. of the 25th Int'l Conf. on Data Engineering. Shanghai: IEEE Computer Society, 2009. 393−404. [doi: 10.1109/ICDE.2009.132]

[13] Fan WF, Li JZ, Luo JZ, Tan ZJ, Wang X, Wu YH. Incremental graph pattern matching. In: Proc. of the 2011 ACM SIGMOD Int'l Conf. on Management of Data. Athens: ACM Press, 2011,38(3):925−936. [doi: 10.1145/2489791]

[14] Choudhury S, Holder L, Chin G, Feo J,. In: Proc. of the Workshop on Dynamic Networks Management and Mining. NewYork: ACM Press, 2013. 1−8. [doi: 10.1145/2489247.2489251]

[15] Song C, Ge T, Chen C, Wang J. Event pattern matching over graph streams. Proc. of the VLDB Endowment, 2014,8(4):413−424. [doi: 10.14778/2735496.2735504]

[16] Zong B, Raghavendra R, Srivatsa M, Yan X, Singh AK, Lee KW. Cloud service placement via subgraph matching. In: Proc. of the 30th Int'l Conf. on Data Engineering. Chicago: IEEE Computer Society, 2014. 832−843. [doi: 10.1109/ICDE.2014.6816704]

[17] Mondal J, Deshpande A. Managing large dynamic graphs efficiently. In: Proc. of the 2012 ACM SIGMOD Int'l Conf. on Management of Data. Scottsdale: ACM Press, 2012. 145−156. [doi: 10.1145/2213836.2213854]

[18] Gao J, Zhou C, Zhou J, Yu JX. Continuous pattern detection over billion-edge graph using distributed framework. In: Proc. of the 30th Int'l Conf. on Data Engineering. Chicago: IEEE Computer Society, 2014. 556−567. [doi: 10.1109/ICDE.2014.6816681]

[19] Sun Z, Wang H, Wang H, Shao B, Li J. Efficient subgraph matching on billion node graphs. Proc. of the VLDB Endowment, 2012, 5(9):788−799. [doi: 10.14778/2311906.2311907]

[20] Low Y, Gonzalez JE, Kyrola A, Bickson D, Guestrin C, Hellerstein J. GraphLab: A new framework for parallel machine learning. https://www.mendeley.com/research-papers/graphlab-new-parallel-framework-machine-learning/

[21] Aridhi S, Montresor A, Velegrakis Y. BLADYG: A graph processing framework for large dynamic graphs. Big Data Reaearch. 2017. [doi: 10.1016/j.bdr.2017.05.003]

[22] Valiant L. A bridging model for parallel computation. Communications of the ACM, 1990,33(8):103−111. [doi: 10.1145/79173.79181]

[23] Gao J, Zhou C, Yu JX. Toward continuous pattern detection over evolving large graph with snapshot isolation. The VLDB Journal, 2016,25(2):269−290. [doi: 10.1007/s00778-015-0416-z]

[24] Khurana U, Deshpande A. Efficient snapshot retrieval over historical graph data. In: Proc. of the 29th Int'l Conf. on Data Engineering. Brisbane: IEEE Computer Society, 2013. 997−1008. [doi: 10.1109/ICDE.2013.6544892]

[25] Choudhury S, Holder L, Chin G, Mackey P, Agarwal K, Feo J. Query optimization for dynamic graphs. Corrsionence & Technology Protection, 2014. http://xueshu.baidu.com/s?wd=paperuri%3A%285125477c2fae3f27651d70c5d9193b26%29&filter=

sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http%3A%2F%2Fde.arxiv.org%2Fpdf%2F1407.3745&ie=utf-8&sc_us=2466112609285908482

[26] Kao JS, Chou J. Distributed incremental pattern matching on streaming graphs. In: Proc. of the ACM Workshop on High Performance Graph Processing. Kyoto: ACM Press, 2016. 43−50.

[27] Yang J, Jin W. BR-Index: An indexing structure for subgraph matching in very large dynamic graphs. In: Proc. of the Scientific and Statistical Database Management. 2011. 322−331. [doi: 10.1007/978-3-642-22351-8_20]

[28] Ma S, Cao Y, Fan W, Huai J, Wo T. Capturing topology in graph pattern matching. Proc. of the VLDB Endowment, 2011,5(4): 310−321. [doi: 10.14778/2095686.2095690]

[29] Fan WF, Hu C, Tian C. Incremental graph computations: Doable and undoable. In: Proc. of the 2017 ACM SIGMOD Int'l Conf. on Management of Data. Chicago: ACM Press, 2017. 155−169. [doi: 10.1145/3035918.3035944]

[30] Yan X, Yu PS, Han J. Graph indexing: A frequent structure-based approach. In: Proc. of the 2004 ACM SIGMOD Int'l Conf. on Management of Data. Paris: ACM Press, 2004. 335−346. [doi: 10.1145/1007568.1007607]

[31] Giugno R, Shasha D. GraphGrep: A fast and universal method for querying graphs. In: Proc. of the IEEE Int'l Conf. on Pattern Recognition. Quebec: IEEE Computer Society, 2002. 112−115. [doi: 10.1109/ICPR.2002.1048250]

[32] Zhao P, Han J. On graph query optimization in large networks. Proc. of the VLDB Endowment, 2010,3(1-2):340−351. [doi: 10.14778/1920841.1920887]

[33] Shang H, Zhang Y, Lin X, Yu JX. Taming verification hardness: An efficient algorithm for testing subgraph isomorphism. Proc. of the VLDB Endowment, 2008,1(1):364−375. [doi: 10.14778/1453856.1452899]

[34] Yang J, Zhang S, Jin W. DELTA: Indexing and querying multi-labeled graphs. In: Proc. of the 20th ACM Int'l Conf. on Information and Knowledge Management. Glasgow: ACM Press, 2011. 1765−1774. [doi: 10.1145/2063576.2063832]

[35] Jiang H, Wang H, Yu PS, Zhou S. GString: A novel approach for efficient search in graph databases. In: Proc. of the 23rd Int'l Conf. on Data Engineering. Istanbul: IEEE Computer Society, 2007. 566−575. [doi: 10.1109/ICDE.2007.367902]

[36] Cheng J, Ke Y, Ng W, Lu A. FG-Index: Towards verification-free query processing on graph databases. In: Proc. of the 2007 ACM SIGMOD Int'l Conf. on Management of Data. Beijing: ACM Press, 2007. 857−872. [doi: 10.1145/1247480.1247874]

[37] Choudhury S, Holder L, Chin G, Agarwal K, Feo J. A selectivity based approach to continuous pattern detection in streaming graphs. Computer Science, 2015,93(8):939−945. [doi: 10.5441/002/edbt.2015.15]

[38] Zhang LX, Wang WP, Gao JL, Wang JX. Pattern graph change oriented incremental graph pattern matching. Ruan Jian Xue Bao/ Journal of Software, 2015,26(11):2964−2980 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4891.htm [doi: 10.13328/j.cnki.jos.004891]

[39] Wickramaarachchi C, Kannan R, Chelmis C, Prasanna VK. Distributed exact subgraph matching in small diameter dynamic graphs. In: Proc. of the IEEE Int'l Conf. on Big Data. Washington: IEEE Computer Society, 2017. 3360−3369. [doi: 10.1109/BigData.2016.7840996]

[40] Raghavendra R, Lobo J, Lee KW. Dynamic graph query primitives for SDN-based cloudnetwork management. In: Proc. of the 1st Workshop on Hot Topics in Software Defined Networks. Helsinki: ACM Press, 2012. 97−102. [doi: 10.1145/2342441.2342461]

[41] Mondal J, Deshpande A. CASQD: Continuous detection of activity-based subgraph pattern queries on dynamic graphs. In: Proc. of the 10th ACM Int'l Conf. on Distributed and Event-Based Systems. Irvine: ACM Press, 2016. 226−237. [10.1145/2933267.2933316]

[42] Reality mining dataset. 2017. http://reality.media.mit.edu

[43] Kuramochi M, Karypis G. Frequent subgraph discovery. In: Proc. of the IEEE Int'l Conf. on Data Mining. San Jose: IEEE Computer Society, 2001. 313−320. [doi: 10.1109/ICDM.2001.989534]

[44] Cordella LP, Foggia P, Sansone C, Vento M. A (sub)graph isomorphism algorithm for matching large graphs. IEEE Trans. on Pattern Analysis & Machine Intelligence, 2004,26(10):1367−1372. [doi: 10.1109/TPAMI.2004.75]

[45] Bader DA, Madduri K. A graph-theoretic analysis of the human protein-interaction network using multicore parallel algorithms. Parallel Computing, 2008,34(11):627−639. [doi: 10.1016/j.parco.2008.04.002]

附中文参考文献:

[4] 于静,刘燕兵,张宇,刘梦雅,谭建龙,郭莉.大规模图数据匹配技术综述.计算机研究与发展,2015,52(2):391−409. [doi: 10.7544/issn 1000-1239.2015.20140188]

[38] 张丽霞,王伟平,高建良,王建新.面向模式图变化的增量图模式匹配.软件学报,2015,26(11):2964−2980. http://www.jos.org.cn/1000-9825/4891.htm [doi: 10.13328/j.cnki.jos.004891]

许嘉(1984－),女,山东荣成人,博士,副教授,CCF 专业会员,主要研究领域为数据库理论与技术,图数据分析,数据隐私保护,大数据分布式并行计算.

张千桢(1992－),男,硕士生,CCF 学生会员,主要研究领域为图数据管理.

赵翔(1986－),男,博士,讲师,CCF 专业会员,主要研究领域为图数据管理与挖掘,基于大数据的情报智能.

吕品(1983－),男,博士,副研究员,CCF 专业会员,主要研究领域为无线网络与移动计算,物联网,网络虚拟化,网络数据分析.

李陶深(1957－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为无线 Mesh 网络,云计算与大数据,网络计算与信息安全,分布式工程数据库.