社会网络与图匹配查询

马 帅 曹 洋 沃天宇 怀进鹏
北京航空航天大学

关键词：社会网络 图匹配查询

背景

无线互联网和3G等新兴技术的发展为用户之间通过各种社会网络，包括BBS论坛、在线社区（如Facebook、人人网和开心网等）、微博（如Twitter、新浪微博和腾讯微博等）进行交流带来了便利，网民也越来越多。据中国互联网络信息中心（CNNIC）发布的互联网络发展状况统计报告[1]显示，截至2011年12月底，中国网民规模已达到5.13亿人，社交网站使用率为47.6%，微博使用率由2010年的13.8%猛涨至2011年的48.7%。

社会网络的飞速发展，对个人和社会群体的行为产生了深远影响。以脸谱（Facebook）为例进行分析，从中可以发现：（1）用户群规模大，全球每13个人中就有1个人使用，并且超过一半的用户每天都登陆Facebook；（2）使用频繁，年龄段在10～34岁之间的用户有48%每天醒来（甚至有28%的用户在起床前）就查看自己的账户信息；（3）朋友圈较大，平均每个用户有130个朋友；（4）所有用户每月在线时间达7000亿分钟；（5）48%的年轻人通过Facebook获取新闻消息[2]。

在社会网络中，可以把用户看作为图的顶点，用户之间的关系（如朋友关系）看作图的边。图的相关理论和技术在社会网络中有着重要的应用，是目前学术界和业界关注的热点之一。

国内外知名研究机构和公司对图的研究与应用都非常重视。例如，微软研究院的Trinity项目[3]和用于数据中心的“Querying Large Distributed Graphs”项目[4]；谷歌的大图处理系统Pregel[5]和MapReduce[6]；雅虎研究院的“Graph Partitioning”项目[7]；Neo4j公司的开源图数据库[8]；美国加州大学圣巴巴拉分校（University of California Santa Barbara）的“Massive Graphs in Clusters”项目[9]；英国爱丁堡大学的模式匹配项目[10]以及北京航空航天大学和国内各大院校的关于图的研究。

图匹配查询

图匹配查询指的是一种概念上非常广泛的图的查询语言，在社交网络中有着广泛的应用。下面我们首先给出一个简单的形式化定义。

图匹配：给定一个模式图（pattern graph）G_1和一个数据图（data graph）G_2：

（1）判断G_1是否“匹配”G_2；或者

（2）从G_2中找出所有跟G_1“匹配”的子图。

注1：这里图是由顶点集和边集组合而成，而顶点和边上通常会有标签标注相关信息。

注2：图匹配的定义包含了两类查询，第一类查询是布尔查询，即需要回答“是”或者“否”的查询；第二类查询返回结果时需要利用第一类查询，两者之间有着紧密的关系。此外，模式图G_1通常比较小，仅仅包含几个或者几十个顶点；而数据图G_2通常较大，甚至包含以“亿”为数量级的顶点和边。

应用实例

下面通过社会关系查找、角色分析、推荐系统和交通路线选择4个应用实例来进一步介绍图匹配

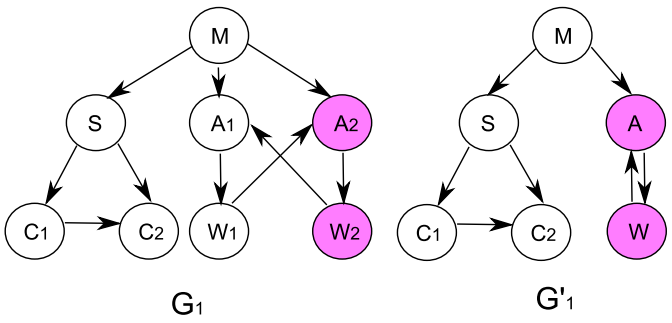

图1 员工角色分析

查询在社会网络中的应用情况。

实例一：社会关系查找

图匹配查询在社会关系查找中有着广泛的应用。下面给出一个查询远房亲属关系的应用案例[18]。华人社交网络记录了华人之间的社会关系，其中网络的顶点是人，且顶点上带有属性值，用来记录人的姓名；边是人与人之间的各种人际关系，且边上带有属性值，用来记录所连接的两人之间的相应关系，例如父子（女）、母子（女）、兄弟（妹）、姐弟（妹）、上下级、师生关系等。

有几类常见的社会关系查询：（1）查找给定的张三和李四两人是否是远房亲戚；（2）查找张三和李四是不是三代以内的近亲；（3）查找张三所有三代以内具有血缘关系的亲属；（4）查找并输出所有与张三有三代以内血缘关系的亲戚且是张三某个兄弟的老板。

事实上，上面所有查询都可用图模式匹配查询中的可达性查询及其扩展[18]来表达并找到结果。总体来讲，查询（1）和（2）只需输出“是”或“否”，查询（3）和（4）需要输出具体的数据（即符合要求的人）。注意到如果张三和李四是具有血缘关系的远房亲戚，则他们在网络中一定存在一条路径（即可达的），并且连接他们的路径中所有边的属性只可能是父子（女）、母子（女）、兄弟（妹）、姐弟（妹）四种具有血缘联系的社会关系。而对于判定三代以内的血亲关系，则可以通过限制关系网络中连接两人的符合边属性约束的路径长度来（即需要通过不超过三“跳”的路径连接张三和李四）实现。因此，以上4种查询应用都可以用带边属性约束的可达性查询完成。其中，查询（3）和（4）需要输出所有满足模式图约束的节点（人）。查询（4）可以由两步完成：首先查出所有与张三是三代以内有血亲关系的亲戚，然后再查出这些亲戚里哪些是张三的兄弟们（“兄弟”关系）的上司（“上下级”关系）。

实例二：角色分析

角色分析在社会网络中起到很重要的作用，如大型公司对其职工的重要性进行评估，并据此制定相应措施，以激励员工或者在经济不景气裁员时能够做出正确的决定。为了分析员工对公司的重要性，公司需要了解哪些人的工作是可以相互替换的，即分析角色的等效性，而这是图匹配查询的一个经典应用[12,13]。

公司的所有职员组成一个社会网络，他们之间的工作关系用有向边来表示。图1中G_1为某网络服务公司的成员关系图。其中M是部门负责人，负责对业务进行决策；S是秘书，负责M对业务需求传达给负责与客户沟通的销售人员C；A是主管助理，负责M的日常行政事务，并将具体安排给其助手W。这里“边”表示领导关系，不同脚标用于区别同一职位的不同员工（如C_1、C_2）。

在G_1中，我们可以通过基于图模拟[16]的图匹配查询来找出任意两个员工之间的是否可被“模拟”的关系。如果角色X能被Y所模拟，则证明X可以被Y替代。在此例中，A_1和A_2可以相互模拟；W_1和W_2可以相互模拟；C_1和C_2不能相互替代。事实上，A_1和A_2的工作内容都是接受主管M命令，然后部署传达给其助手，并且接受他们的反馈；W_1和W_2的情况类似；C_1需要向C_2传达业务信息，相反C_2不能如此，因此C_1不能被C_2代替。最后，如果公司通过对过去的工作绩效评比发现A_2优于A_1，W_2优于W_1，则最终选择裁员A_1和W_1，使得精简后的部门职员工作关系如图1中G'_1所示。

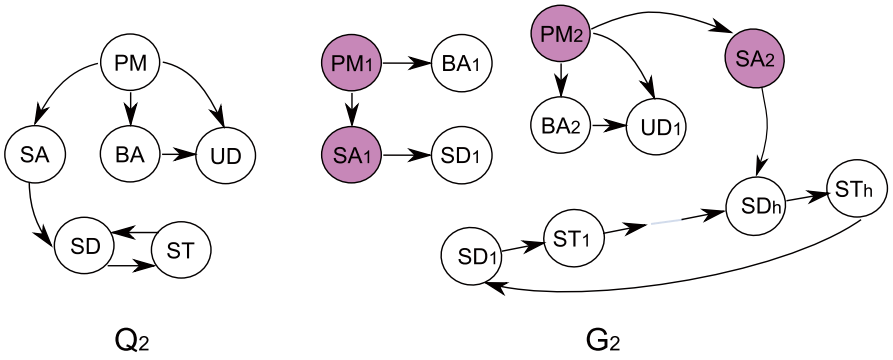

图2 专家推荐

实例三：推荐系统

在推荐系统中也常常会用到图匹配查询，如对新兴特定应用的高级推荐功能[21]。

图2中G_2是一类面向领域专家的求职社会网络，其中顶点表示专家，顶点标签是专业领域（脚标用来区分相同领域的不同专家），边表示良好的领导合作关系。现在需要从该网络中找出一个团队来研发一项软件产品。一个理想的软件开发团队包括以下角色：项目管理员（PM）、业务分析师（BA）、软件架构师（SA）、用户界面设计师（UD）、软件开发员（SD）以及软件测试员（ST）。由于团队成员的配合程度对软件开发至关重要，我们需要所选择的软件开发团队成员满足模式图Q_2所示的合作关系，例如，若存在从项目管理员到软件架构师，软件开发员到软件测试员以及软件测试员到软件开发员的边，则说明团队招聘的项目管理员曾经领导过软件架构师，招聘的软件开发员所开发的软件曾经被软件测试员测试过，并且软件测试员曾经将测试结果反馈给软件开发员以帮助其提高软件开发质量，并且以前的项目合作非常成功。至此，招聘合适专家构建软件开发团队就转化成了图模式匹配问题了，即如何从图G_2所示的领域专家社交网络中找出子图与模式图Q_2匹配，以满足软件开发团队的各种要求。

我们可以通过不同的匹配语义，找出满足不同程度约束的结果，如子图同构[15]、图模拟[16]和强模拟[17]等。感兴趣的读者请阅读相关论文。

实例四：交通路线选择

当前应用广泛的定位服务（location based services）使得交通网络领域也成为图匹配查询的应用领域。例如以下实例[18]。留美学生小明要从美国加州的欧文市（Irvine）前往河边市（Riverside），他的主要任务是选择合适的路线，路线的合适与否取决于小明对行程的要求和约束。

1. 如果小明需要自己开车以最短的时间到达河边市，则该问题可以用最短路径查询来表达。即模式图表达的约束是从欧文市到河边市的最短路径，数据图就是整个美国公路路线图。用现有关于最短路径查询的相关方法可以帮小明找到耗时最短的自驾行车路线是261号州际公路。

2. 如果小明需要用大型卡车将一批物资尽快从欧文市送到河边市，出于考虑公共健康和安全，许多桥梁和铁路交汇处是不允许这类车辆通行的。这时，可通过含有特定路线约束（如正则表达式）的模式图来查找最优交通路线。

上述实例说明图匹配查询在社会网络中有着广泛的应用。此外，在社交网络的好友和群（兴趣团体）的推荐、在线商城中的商品排序和自动推荐、视频分享网站上的视频归类、犯罪团伙预测、信息传播、系统冗余备份设计、网络重复结构检测等许多方面，图匹配查询都有重要应用。

相关技术

图匹配查询的相关技术包括模式图预处理、数据图预处理、索引技术和外存查询。

模式图预处理

通过对模式图预处理可以提高查询效率，常见

的模式图预处理有模式图最小化和模式图相似变换等[19~20]。模式图最小化是指在保证查询结果不变的前提下，对模式图中的冗余点和边进行最大化的去除；模式图相似变换则是指将模式图简化为相似的模式图，并在一定程度上保持查询结果的正确性，也包括将模式图表示为若干已知查询结果的模式图的组合，并利用已有查询结果进行相关处理，得到最终查询结果。这通常需要结合图匹配查询语言的查询语义的特点来具体分析处理[20]。

数据图预处理

数据图通常很大，在执行图匹配查询之前，往往需要对数据图进行预处理来提高效率。其主要采用的技术包括取样、压缩和划分。

数据图取样 为了便于处理，通常会对数据图进行取样处理，并在取样后的数据图上执行匹配或分析任务[21]。在取样过程中，会综合考虑数据图的各种特征，使得取样所得数据图能尽可能多地体现原始数据图的特征和信息，并在允许一定精度损失的情况下，通过对数据图信息进行筛选，来减小应用所需处理数据图的大小。由于取样一般存在数据丢失的问题，因此如何根据应用约束和需求选取合适的取样方法、确定取样的大小，如何保证取样结果能反应原始图的特征，以及如何在应用需要时快速增加取样大小等问题是进行数据图取样时需要考虑的核心问题。

数据图压缩 与取样类似，压缩也是通过减小数据图的规模来提高查询效率的。与数据图取样不同的是，数据图压缩往往会针对某一类特殊的应用（查询）所需要的信息对原始数据图进行处理[22]，而并不需要一般性（通用性的）地保持原始数据图的各种属性或整体特征。更为重要的是，压缩通常对输出的数据图的格式不存在任何限制，而取样则一般会使用原始数据图的格式。

目前，已有许多针对可达性查询、邻接关系查询的研究工作[23]，并且信息编码和图文法理论和技术也被用来进行数据压缩。但是针对各类结构查询类的图模式语言的数据图压缩工作还较少。

数据图划分 分布式计算也是在大规模数据图上进行匹配查询的有效方法之一。通过将数据图划分成若干小图，并将其分布在多个计算节点上，可以在一定程度上提高匹配和分析速度。理想的图划分方法能够将图划分成相等大小的若干部分，以提高分布式处理的加速比。然而，这种图的等分划分问题的固有复杂度很高（NP难问题），因此目前大规模数据处理系统为了满足实际应用系统的要求，常采用哈希函数进行随机划分以达到快速划分。

索引技术

在数据图规模较大的情况下，为图匹配查询语言在数据图上建立索引，可以有效地提高查询效率。目前已存在针对（各类）图匹配查询的索引。而衡量这些索引好坏的标准主要有三个：索引大小、索引查询时间和索引构建时间。索引越小表示所设计的索引格式对数据存储的额外负担越小；索引查询时间表示查询语言在该索引上执行的时间，该时间相对于在原始图上执行的时间越小越好，是衡量索引性能的重要指标；索引创建时间是指从数据图创建相应索引的耗时。这三个指标往往相互矛盾。另外，当数据图出现变动时，索引更新的快慢决定了其对数据图动态特性支持的好坏。

外存查询

当数据图规模很大而无法完整地导入内存时，在其上进行图匹配查询就是外存查询。目前已有针对可达性、邻居关系、最短路等简单查询语言的外存查询研究。算法与外存的I/O互交次数是外存查询需要考虑的重要指标。通过设计合适的数据图的存储数据结构、利用局部查询等手段可以提高外存查询的效率。

结语

根据图匹配查询的定义，很多大家常见熟知的图查询都属于图匹配查询，比如最短路径[21]、邻接查询[23]、图同态及其扩展查询[24]、子图同构查询[26]、

图模拟[16]及其扩展强模拟查询[17]和关键字查询[21]等。这类图的查询没有明确的规定查询语言的语法，比较Ad-hoc，常用于完成图中的某单项特定查询任务。将这些图查询归为图匹配带来的好处是为这些自组织网的查询语言提供了统一的逻辑框架。此外，模式图和“匹配”语义的不同会形成不同的图匹配查询语言。按照模式图结构的不同可分为：点查询（如可达性查询、邻居节点查询等）、路经查询（如最短路查询等）以及子图结构查询（如子图同构、图模拟、强模拟等）。尽管模式图结构都相同，但由于约束语义的不同，因此形成了子图同构、图模拟和强模拟等不同的图匹配查询语言。

我们通过调查发现图匹配查询理论及相关技术在社会网络中有着重要的商业价值。而社会网络中的图所具有的动态性、不确定性和规模大等特点也对传统的图匹配查询理论和技术提出了挑战。我们认为今后需要重点开展以下三个方面的研究。

动态查询 由于数据图经常动态变化，其结构（顶点和边）和属性都可能出现变动。动态查询研究的是当数据图的顶点和边增加或减少后，如何在利用原有的匹配查询结果来快速得到数据图变化后的结果，而不需要从“零”开始重新进行计算。目前这方面的研究还处于起步阶段。

非确定查询 造成图的不确定性的原因有两个：由于数据采样或者数据丢失造成的数据本身的质量问题，以及数据自身的内在的动态变化特性，采用合理的模型描述图的不确定性，并设计有效的非确定性图匹配查询，是未来重要的研究方向之一。

分布式查询 当数据图规模大到无法在单台机器上处理时，就需要将数据图分布在多台机器上，以便进行分布式查询。衡量分布式查询好坏的指标包括查询过程中的节点访问次数、数据传输量和查询完成时间[27]。当三者相互矛盾时，如何综合考虑进行合理的取舍，以及设计分布式查询算法也是未来重要的研究方向之一。

总之，大数据时代图的匹配查询理论及相关技术是一个亟待研究和解决的内容，尽管目前尚缺乏系统的研究，但它具有重要的科学意义和应用价值。■

马 帅

CCF会员。北京航空航天大学计算机学院教授。主要研究方向为数据库理论与系统，图匹配和数据质量等。

mashuai@buaa.edu.cn

曹 洋

北京航空航天大学计算机学院博士生。主要研究方向为图匹配和社交推荐等。

caoyang@act.buaa.edu.cn

沃天宇

CCF会员。北京航空航天大学计算机学院讲师。主要研究方向为分布式系统和网络计算等。

woty@act.buaa.edu.cn

怀进鹏

CCF名誉副理事长。中国科学院院士。北京航空航天大学计算机学院教授。主要研究方向为网络化软件技术和系统研究工作等。

huaijp@buaa.edu.cn

参考文献

[1] 中国互联网络信息中心.中国互联网络发展状况统计报告. 2011. http://www.cnnic.cn/dtygg/dtgg/201201/W020120116337628870651.pdf

[2] http://www.digitalbuzzblog.com/facebook-statistics-stats-facts-2011/

[3] http://research.microsoft.com/en-us/projects/trinity/

[4] http://research.microsoft.com/en-us/projects/ldg/default.aspx

[5] Grzegorz Malewicz, Matthew H. Austern, Aart J. C. Bik, James C. Dehnert, Ilan Horn, Naty Leiser and Grzegorz Czajkowski, Pregel: a system for large-scale graph processing. SIGMOD 2010

[6] Jeffrey Dean and Sanjay Ghemawat, MapReduce: Simplified Data Processing on Large Clusters. OSDI 2004

[7] http://research.yahoo.com/project/2368

[8] Neo4j: the graph database, http://neo4j.org/
[9] UCSB Massive Graphs in Clusters, http://graphs.cs.ucsb.edu/magic/index.html
[10] http://www.lfcs.inf.ed.ac.uk/research/database/Projects.html
[11] Jin, R. and Hong, H. and Wang, H. and Ruan, N. and Xiang, Y., Computing label-constraint reachability in graph databases, ICDE 2010
[12] Brynielsson, J. and Hogberg, J. and Kaati, L. and Martenson, C. and Svenson, P., Detecting social positions using simulation, ASONAM 2010
[13] Wasserman, Stanley and Faust, Katherine. Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press. 1994
[14] L. Terveen and D. McDonald. Social matching: A framework and research agenda. ACM Trans. Comput.-Hum. Interact., 12(3), 2005
[15] J. R. Ullmann. An algorithm for subgraph isomorphism. J.ACM, 23(1), 1976
[16] M. R. Henzinger, T. Henzinger and P. Kopke, Computing simulations on finite and infinite graphs. FOCS 1995
[17] Shuai Ma, Yang Cao, Wenfei Fan, Jinpeng Huai, and Tianyu Wo, Capturing Topology in Graph Pattern Matching. VLDB 2012
[18] Rice, M. and Tsotras, V.J., Graph indexing of road networks for shortest path queries with label restrictions, VLDB 2010
[19] D. Bustan and O. Grumberg. Simulation-based minimization. TOCL, 4(2), 2003
[20] Bordino, I. and Castillo, C. and Donato, D. and Gionis, A., Query similarity by projecting the query-flow graph, SIGIR 2010
[21] Elmagarmid, Ahmed K. and Aggarwal, Charu C. and Wang, Haixun, Managing and Mining Graph Data, Advances in Database Systems, Springer 2010
[22] Wenfei Fan Jianzhong Li, Xin Wang, and Yinghui Wu. Query Preserving Graph Compression. SIGMOD 2012
[23] Hossein Maserrat and Jian Pei, Neighbor query friendly compression of social networks. KDD 2010
[24] G. Ramalingam and Thomas Reps, An incremental algorithm for a generalization of the shortest-path problem. Journal of Algorithms, 1996
[25] Wenfei, Jianzhong Li, Shuai Ma, Hongzhi Wang and Yinghui Wu, Graph Homomorphism Revisited for Graph Matching. VLDB 2010
[26] Brian Gallaghe, Matching structure and semantics: A survey on graph-based pattern matching. AAAI FS. 2006
[27] Shuai Ma, Yang Cao, Jinpeng Huai, Tianyu Wo, Distributed Graph Pattern Matching, WWW 2010