图查询：社会计算时代的新型搜索

马 帅 李 佳 刘旭东 怀进鹏
北京航空航天大学

关键词：图查询 社会计算

图查询

图的表达能力非常强，近年来受到工业界和学术界的共同关注。本文将介绍当前图查询在工业界的应用和学术界的研究动态，进而揭示在社会计算时代图查询研究的重要意义。

图查询的应用

图查询在各个领域中的应用情况如下：

复杂对象识别 一个例子是脏数据清洗技术。有调查表明脏数据导致美国商业每年损失6000亿美元[1]。采用数据清洗技术可以减少因脏数据带来的损失。英国电信（BT）公司采用数据清洗技术挽回的整体商业价值超过6亿英镑[2]。数据清洗主要包括数据修复和对象识别两项技术[3]，而复杂对象的识别是对象识别中最难的问题，即在数据结构不规则的情况下，识别表示同一实体的复杂对象。一种解决方法是将复杂对象表示为图，采用图匹配技术来有效地发现和识别相同的复杂实体，如子图同构和扩展同态查询[4,5]等。

社交网络和Web网络 社交网络的飞速发展，对社会和个人的行为产生了深远的影响[6,7]。在社交网络中，如果用图表示，用户可以看作图的顶点，用户之间的关系（如朋友关系等）可以看作图的边；与社交网络类似，Web网络中的网页可以看作图的顶点，网页之间的链接关系可以看作图的边。图在社交网络中有着重要的应用[8,9,11]，如近邻查询和图压缩[10]等；图在Web网络中也有着广泛的应用，如网页的聚类可以看作图的分类问题[12]，镜像站点检测问题可以看作图的匹配问题[4]等。

生物数据分析 大量的生物数据可以表示成图数据。基于图的查询对分析生物数据有着重要的意义[13]。在蛋白质交互网络中，通过分析图，可以有效地分析基因和蛋白质的功能，并为器官的功能组织和演化提供重要的参考依据[14]。

软件代码剽窃检测 随着开源软件的流行，软件代码的剽窃变得相对容易。将程序中的数据和控制流程转变为程序依赖图，在其中执行子图同构查询可以有效地检测代码是否被剽窃[15]。感兴趣的读者可阅读文献[26]来看一些具体应用实例。

此外，图在虚拟机映射[16]、公路网络[17]、模式

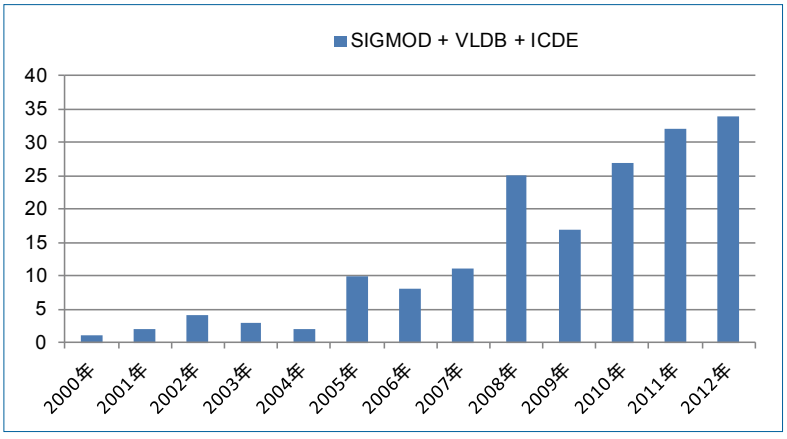

图1 论文统计

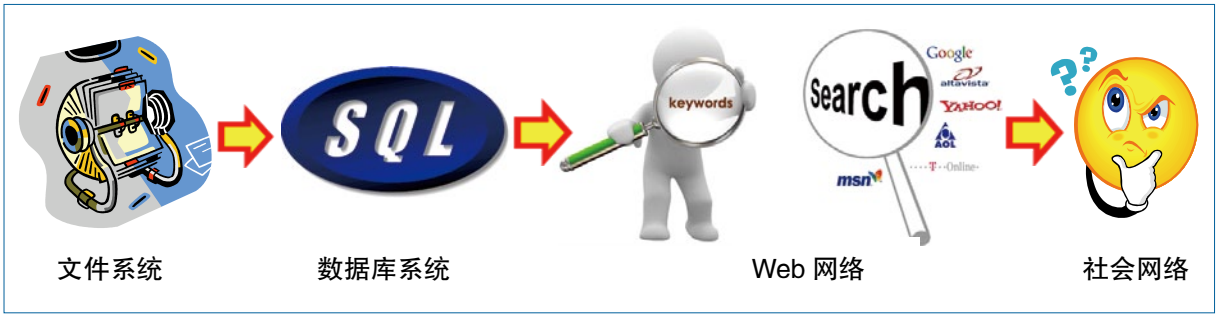

图2 查询的演变

识别[18]和超大规模集成电路（very large scale integration，VLSI）设计[19]等方面也有重要的应用。

我们通过对数据库领域三大系统会议（SIGMOD、VLDB和ICDE）发表的关于图的研究论文的数量进行统计分析（见图1），来研究图在数据库领域学术界的研究发展动态。结果如下：

1 从2000年前后，数据库研究人员开始关注图的研究；

2 从2000年至今，图的论文数量总体上处于上升趋势；

3 从2008年开始，研究图的工作开始显著增多成为数据库领域的研究热点之一。

以上事实说明，图查询目前在工业界和学术界的关注度明显增加。

图查询的重要性

为什么近年来工业界和学术界共同关注图的理论与技术呢？我们需要从查询搜索的历史发展来看待这种现象（见图2）。计算机采用的查询方式经历了文件系统查询→数据库查询→web→Web→网络社会网络查询，这样的发展历程。

文件系统 从20世纪60年代开始，计算机开始装配了具有现代意义的操作系统[20]，而文件系统是操作系统提供的一种存储和组织计算机文件的方法。它提供简单的查询功能，使用户可以搜索文件。

数据库系统 20世纪60年代中期，数据库系统开始在商业中得以应用。20世纪70年代，关系数据模型成为数据库管理系统的主流。70年代后期发明的结构化查询语言SQL极大地提高了数据查询的灵活性。用户可以通过SQL语言来进行各种复杂的查询。

Web网络 从20世纪90年代开始，随着万维网（World Wide Web）的兴起，Web搜索引擎（Google、Bing和Yahoo!等）广泛应用。它们通过提供“关键词搜索”这一简单而实用的功能，使得几乎所有的用户都可以方便地搜索万维网数据。

社会网络 20世纪末至今，随着Web 2.0和社会计算的兴起，社交网络系统开始大量应用，如Facebook（脸书，亦有译作“脸谱”）、LinkedIn、人人网等。结合图1，读者可以发现一个有趣的现象：数据库领域的研究人员几乎在同一时期开始关注图的理论与技术。

现在，人们需要一种什么样的搜索技术呢？我们认为，图查询是适合社会计算时代的搜索方式。虽然对于社会计算目前还没有明确公认的定义，但是大家普遍承认社会计算一般需要考虑社会的结构、组织和活动等社会因素。所有的社会活动构成了社会网络，本质上这是图的一种表现形式，所以图搜索（即从图中查询信息）自然而然地就成了工业界和学术界的共同关注点。社会网络也对传统的搜索引擎起到了积极的推动作用，比如知识图（knowledge graph）被引入Google搜索引擎来提高搜索结果的质量[21]。

以上种种表明图查询是社会计算时代搜索的一种重要的特征，并起到了极其重要的作用，因此得到了工业界和学术界的共同关注。下面将介绍子图查询及其面临的问题与挑战。

子图查询

子图查询指的是图的查询语言，即给定一个输入图，输出为给定图的子图。我们将介绍三类子图查询：凝聚子图查询、关键词图查询和图模式匹配查询。

凝聚子图查询

凝聚子群原指社交网络整体用户中成员之间满足某种“紧密关系”的子集用户群。根据应用需求的不同，会有不同的“紧密关系”，从而产生不同的凝聚子群。社会网络可以用图来表示，所以也称凝聚子群为凝聚子图，相应地将从图中查询凝聚子图的查询称为凝聚子图查询（cohesive subgraph search）。

下面结合著名的帕吉特佛罗伦萨家族网络（Padgett's Florentine）（见图3）来解释[22]几种常见的凝聚子图。这个家族网络包括了15世纪早期意大利佛罗伦萨的16个大家族的婚姻关系网络。其中图的顶点表示家族，用家族的姓氏加以标注；边表示一个家族的某个成员和另一个家族的某个成员有着婚姻关系。家族间通过婚姻和商业交易结成巩固的政治经济同盟。我们可以利用凝聚子图查询技术从家族关系数据中找到这些家族同盟体，并研究不同家族同盟间的政治经济关系，从而更好地了解当时佛罗伦萨的历史[22]。

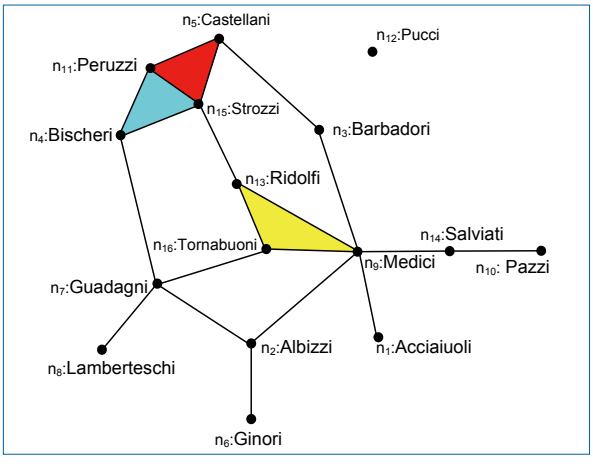

图3 帕吉特佛罗伦萨家族网络图

极大团（maximal clique） 图3中的极大团是图的完全子图，完全子图中任意两个顶点均相邻，并且图中不存在其他顶点与该子图中的所有顶点都相邻的现象。在帕吉特佛罗伦萨家族网络图中存在三个极大团：

- n_4:Bischer、n_{11}:Peruzzi和n_{15}:Strozzi，
- n_5:Castellani、n_{11}:Peruzz和n_{15}:Strozzi，
- n_9:Medici、n_{13}:Ridolfi和n_{16}:Tornabuoni。

其中第一个极大团由Bischeri、Peruzzi、Strozzi三个家族组成，含有三个顶点，团中任意两个家族间均存在婚姻关系，而且图中不存在与这三个家族同时存在婚姻关系的其他家族。其余两个极大团也类似。

由于团含有独特的数学性质和紧密的结构特征，是凝聚子群问题研究的基础。然而团对结构的要求非常严格，因此图中实际存在的团通常较少、较小，如帕吉特佛罗伦萨家族网络图中只存在三个极大团，且三个极大团各自分别仅含三个顶点。实际应用中，通过减少或者减弱对团的约束关系形成应用更广泛的凝聚子图，如n-极大团、n-宗派和k-极大核等。

n–极大团（n–clique） 一个图中的n-极大团是该图的极大子图，它仅要求其任意两顶点是可达的，并且最短距离不大于n即可，而不需要任意两点之间必须存在一条边。如，在帕吉特佛罗伦萨家族网络图中存在13个2-极大团。由于只对顶点间距离做限制，n-极大团的直径（团中任意顶点之间最短路径的最大值）有可能大于n。如由n_2:Albizzi、n_4:Bischeri、n_7:Guadagni、n_{13}:Ridolfi、n_{16}:Tornabuoni五个家族组成的2-极大团的直径为3（大于2）。其根本原因在于n-极大团中的任意两顶点间距离所在的最短路径上的顶点不一定都属于该n-极大团。

n–宗派（n–clan） 一个图中的n-宗派是该图中一个n-极大团，并且该n-极大团的直径不大于n。可以看出，为了使凝聚子群结构更具紧密性，n-宗派进一步增加了n-极大团的约束条件，从而使得

n-宗派中的任意两顶点间距离所在的最短路径上的顶点一定属于该n-宗派。

k-极大核（k-core） 一个图中的k-极大核为该图中顶点度均不小于k的极大子图。在帕吉特佛罗伦萨家族网络图中，由n_2:Albizzi、n_3:Barbadori、n_4:Bischeri、n_5:Castellani、n_7:Guadagni、n_9:Medici、n_{11}:Peruzzi、n_{13}:Ridolfi、n_{15}:Strozzi和n_{16}:Tornabuoni十个家族组成的子图就是一个2-极大核。

从凝聚子图的定义可以看出，随着凝聚关系要求的放宽（n-极大团和n-宗派相比极大团限制条件更少），更多凝聚子图被抽取出来，且每个子图中所含顶点个数显著增多。此外，对于团来讲，从图中寻找一个最大团是NP难解问题，因此凝聚子图的极大团、n-极大团和n-宗派没有采用最大团的语义，这实际上是在查询结果的准确性和查询效率两个因素之间综合考虑的一种折中方案。对于k-极大核来讲，采用极大子图和最大子图语义是等同的，并且是多项式可解的。

关键词图查询

怎样在快速增长的信息资源中获取需要的数据信息，是当今一项研究热点。关键词搜索（keyword search）为用户从数据集中获取相关信息提供了有效的技术支持。由于关键词搜索极其友好的查询界面，它已经成为事实上的互联网数据信息检索通用机制。

关键词图查询（keyword search on graphs）指的是给定一组关键词，从图中查找“满足”该组关键词的子图，并且子图中的顶点满足“一定”的结构约束关系。这样，图上的关键词查询同时考虑顶点之间结构和包含的内容两类信息，通常输入图上的每个顶点都被标示了一组关键词。关键词图查询的基本要求是找到的子图的顶点中包含所有的输入关键词，而结构约束关系的不同导致了不同的查询方法和技术。下面我们介绍三类图关键词查询。

最小树语义 目前大多数关键词查询采用最小树语义方法。查找到的结果是树，所有输入的关键词一定出现在该树的某个顶点中，并且该树的所有边的权重之和最小[23]。

r半径斯坦纳（Steiner）图语义 给定一个半径小于等于r的图G和一组关键词K，如果G中两个顶点u、v均包含输入K中某个关键词，那么u和v之间路径上的点（包含u、v）称为斯坦纳顶点。实际上斯坦纳顶点就是与K中关键词直接或者间接相关的顶点。以斯坦纳顶点及其相关边构成的G的子图就称为r半径斯坦纳图[24]。采用这种语义的关键词图查询输出结果是r半径斯坦纳图。

r-极大团语义 采用这种语义的关键词图查询输出结果是r-极大团。该方法在图结构数据集中搜索得到含有关键词，而且r-极大团顶点集合包含了所有的输入关键词，同时任意两个顶点间距离都不大于r，这样就对搜索结果间关系的紧密程度做了限制[25]。

实际上，由于关键词图查询中缺少输入关键词之间的结构约束关系，因此需要通过“猜想”关键词之间的拓扑结构，形成各种语义。并且由于对用户期望的查询结果进行了猜想，查询的结果就需要结合排序（ranking）。因此所有的关键词搜索（包括经典的关键词搜索）都需要结合排序技术。

图模式匹配

图模式匹配（graph pattern matching）在社交网络中发挥着重要的作用，是指给定一个模式图（pattern graph）和一个数据图（data graph），从数据图中找出与模式图“匹配”的所有子图。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。此外，模式图通常比较小，仅仅包含几个或者几十个顶点；而数据图通常较大，甚至包含以“亿”为数量级的顶点和边[26]。尽管模式图结构都一样，由于匹配语义的不同，形成了子图同构[5]、图模拟[9]和强模拟[27]等不同的图匹配查询语言。

图模拟和强模拟的语义相对复杂，感兴趣的读者可阅读文献[9, 27]。我们在此将仅介绍子图同构（subgraph isomorphism）。在介绍子图同构定义

前，我们首先介绍图同构（graph isomorphism）的定义。

图同构 给定一个数据图G和一个查询图Q，则Q与G同构当且仅当Q顶点集V_Q与G的顶点集V_G之间存在一个双射关系$f: V_Q \rightarrow V_G$，使得若图Q中任意两个顶点u和v之间有一条边当且仅当在图G中相应顶点f(u)和f(v)之间有一条边。

子图同构 给定一个数据图G和一个查询图Q，则当且仅当在G中存在一个子图G_s与图Q同构时，Q与G子图同构。

下面我们通过一个例子介绍图模式匹配（图4）。假定一位项目负责人想要找到一位生物学家（Bio）来帮助团队中的几位软件工程师（SEs）分析基因数据。他利用专家推荐网络图G（图4右侧）来搜寻满足条件的生物学家。图中每个顶点代表一个专家，顶点上标签表示其专业方向，顶点之间的边代表两人间的推荐关系，比如HR_1推荐SE_1，DM_1推荐Bio_3。项目负责人希望找到满足模式图Q（图4左侧）所示条件的生物学家：（1）由一位HR推荐；（2）由一位SE推荐，也就是希望这个生物学家有过与软件工程师合作的经验；（3）由一位数据挖掘专家（DM）推荐，因为这份工作需要数据挖掘相关技术知识；（4）且SE也是由HR推荐的；（5）存在一位人工智能专家（AI）推荐DM且被DM推荐。

那么，当输入模式图和数据图分别为图4中Q和G时，在数据图G中执行基于子图同构的图匹配查询时，结果显示数据图G与查询图Q不是子图同构的，即G中不存在任何子图与Q具有完全相同的拓扑结构。

子图同构是NP-完全问题，因此查询效率较低，并且由于子图同构要求匹配图中存在子图与查询图具有完全相同的拓扑结构，所以通常命中率很低，这些因素限制了子图同构的应用范围。因此，最近有些研究试图通过减少或者减弱约束条件来提高图模式匹配的实用性。目前主要有两种方法：一是引入或者提出新的图模式匹配模型，如图模拟[9]和强模拟[27]；二是近似图匹配[23]。

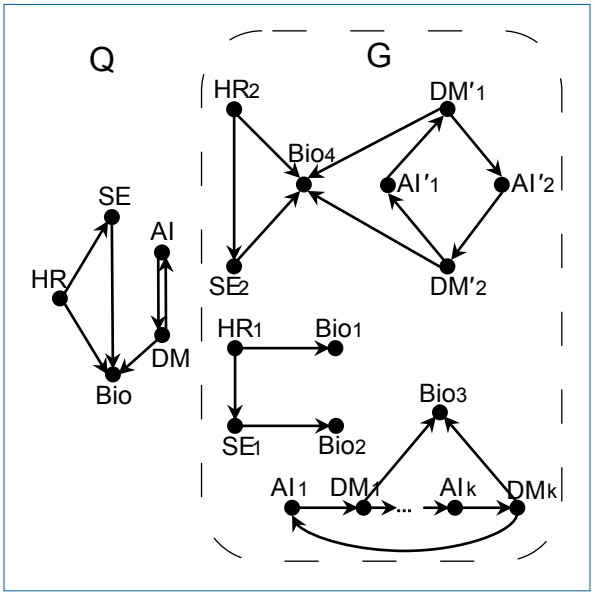

图4 图模式匹配实例

问题与挑战

在前面我们介绍了三类子图查询：凝聚子图查询、关键词图查询和图模式匹配。我们首先从查询的友好性和结果的准确性分析这三类查询存在的主要问题，然后介绍查询所面临的问题和挑战。

查询的友好性 凝聚子图查询不需要用户输入任何信息。关键词图查询一般来说只需要用户输入关键词，其查询结果的拓扑关系是人为假定的。而图模式匹配需要用户输入整个模式图，显然友好性最差。

结果的准确性 凝聚子图查询是查询特定的子图结构，因此不需要讨论结果的准确性。关键词图查询用户没有输入结构的约束条件，所以查询结果的准确性没法保证，从而需要对查询结果进行排序。图模式匹配通过输入的模式图来约束查询结果的拓扑结构，因此查询结果更为准确，且通常不需要对查询结果进行排序。

问题 目前图上的关键词查询的研究大多关注于查询效率、查询结果的拓扑关系以及查询结果多样化，而很少关心查询语义的合理性。其主要原因是很难为关键词查询定义一种通用的查询语义。

比较而言，虽然图模式匹配的查询结果较为准确，但由于用户需要输入模式图，因此使用起来极为不便。那么是否能够引入一种介于“关键词图查询”和“图模式匹配”的查询方法来充分结合二者之间的优点，同时克服各自的缺点呢？这依然是一个值得进一步研究的未解问题。

面临挑战 图数据的处理要比XML数据困难得多，因为XML只是一种特殊的简单图（树），更比处理传统的关系数据难得多。此外，新应用对图的查询与存储技术提出了新的挑战。比如在社交网络等新应用中，用户群的总数量达到了数亿的量级，而每天新增用户达到了数十万的量级[28~32]。据统计：Facebook用户超过8亿[30]，每秒新增用户7.9个，每天新增用户超过60万[29]；Twitter（推特）用户超过1亿，每天新增用户超过30万[30]；人人网的用户数量总计已经超过1.2亿[32]；新浪微博用户已经超过2亿[31,33]。这些统计数据说明了：（1）数据的规模越来越大，达到了前所未有的数亿量级[34]；（2）更新非常频繁，每时每刻都在更新，并且每天更新的规模达到了10万数量级[35]。（3）同传统的关系数据一样[36]，在这些新应用中，也存在数据的不确定性[37]和数据丢失[38]等数据质量问题。总之，当前各种应用面对的图数据具有大规模性、动态性和不确定性三个主要特点[26]。

综上所述，图数据的第一个特点要求图的查询效率要高，并且要充分权衡查询效率和存储空间大小之间的关系；第二个特点要求图的查询要充分考虑动态变化因素和时序特征；第三个特点则要求图的查询要解决图数据质量问题。要解决这些问题就需要对其有进一步深入的理解，还需要借助于分布式处理[39,40]、增量处理[9,41]和数据清洗[42]等理论和技术[26]。

结语

本文介绍了社会计算时代的新型搜索模式——图查询。从分析图查询在当前工业界的应用和学术界研究搜索的历史和动态，我们看到了当前社会计算时代图查询的重要性。我们还重点介绍了凝聚子图查询、关键词图查询和图模式匹配三类子图查询，并对其进行了分析。

图查询作为社会计算时代一种新型搜索，得到了工业界和学术界的共同关注，目前还有很多问题亟待解决，这也是对我们的挑战。■

马 帅

CCF会员、大数据专家委员会委员。北京航空航天大学教授。主要研究方向为数据库理论与系统，图数据管理和数据质量等。mashuai@buaa.edu.cn

李 佳

北京航空航天大学博士生。主要研究方向为图匹配和社交推荐等。
lijia@act.buaa.edu.cn

刘旭东

CCF会员。北京航空航天大学教授。主要研究方向为可信网络计算技术和中间件技术等。
liuxd@buaa.edu.cn

怀进鹏

CCF名誉副理事长、会士。中国科学院院士，北京航空航天大学教授。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] W. W. Eckerson. Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High Quality Data. In The Data Warehousing Institute, 2002

[2] B. Otto and K. Weber. From health checks to the seven sisters: The data quality journey at BT, Sept. 2009. BT TR-BE HSG/CC CDQ/8

[3] Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, and Weiyuan Yu, Interaction between Record Matching and Data Repairing. In SIGMOD Conference, 2011

[4] Wenfei, Jianzhong Li, Shuai Ma, Hongzhi Wang and Yinghui Wu, Graph Homomorphism Revisited for Graph Matching. In VLDB Conference, 2010

[5] Ullmann, J. R., An Algorithm for Subgraph Isomorphism. Journal of ACM, 1976

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

[7] http://www-users.math.umd.edu/~bnk/CAR/project.htm

[8] Yuanyuan Tian and Jignesh M. Patel, TALE: A Tool for Approximate Large Graph Matching. In ICDE Conference, 2008

[9] Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, Yinghui Wu, and Yunpeng Wu, Graph Pattern Matching: From Intractable to Polynomial Time. In VLDB Conference, 2010

[10] Hossein Maserrat and Jian Pei, Neighbor query friendly compression of social networks. In KDD Conference, 2010

[11] P. Barcelo, C. A. Hurtado, L. Libkin, and P. T. Wood. Expressive languages for path queries over graph-structured data. In PODS Conference, 2010

[12] Adam Schenker, Mark Last, Horst Bunke and Abraham Kandel, Classification of Web Documents Using Graph Matching. In IJPRAI Conference, 2004

[13] Patrick Durand, Laurent Labarre, Alain Meil, Jean-Louis Divol, Yves Vandenbrouck, Alain Viari and Jerome Wojcik, GenoLink: a graph-based querying and browsing system for investigating the function of genes and proteins. BMC Bioinformatics 2006

[14] David A. Bader and Kamesh Madduri, A graph-theoretic analysis of the human protein-interaction network using multicore parallel algorithms. Parallel Computing 2008

[15] Chao Liu, Chen Chen, Jiawei Han and Philip S. Yu, GPLAG: detection of software plagiarism by program dependence graph analysis. In KDD Conference, 2006

[16] N. M. Mosharaf Kabir Chowdhury, Muntasir Raihan Rahman and Raouf Boutaba, Virtual Network Embedding with Coordinated Node and Link Mapping. In INFOCOM Conference, 2009

[17] Zaiben Chen, Heng Tao Shen, Xiaofang Zhou and Jeffrey Xu Yu, Monitoring path nearest neighbor in road networks. In SIGMOD Conference, 2009

[18] Donatello Conte, Pasquale Foggia, Carlo Sansone and Mario Vento, Thirty Years Of Graph Matching In Pattern Recognition. In IJPRAI Conference, 2004

[19] George Karypis, Rajat Aggarwal, Vipin Kumar and Shashi Shekhar, Multilevel hypergraph partitioning: applications in VLSI domain. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 1999

[20] Hansen, Per Brinch, Classic Operating Systems. Springer, 2001

[21] http://www.google.com/insidesearch/features/search/knowledge.html

[22] Wasserman, Stanley and Faust, Katherine, Social Network Analysis: Methods and Applications, Cambridge University Press, 1994

[23] Charu C. Aggarwal and Haixun Wang, Managing and Mining Graph Data, Springer, 2010

[24] G. Li, B. C. Ooi, J. Feng, J. Wang, and L. Zhou. Ease: EASE: Ease: Efficient and adaptive keyword search on unstructured, semi-structured and structured data. In SIGMOD Conference, 2008

[25] Mehdi Kargar, Aijun An: Keyword Search in Graphs: Finding r-cliques. In VLDB Conference, 2011

[26] 马帅, 曹洋, 沃天宇, 怀进鹏, 社会网络与图匹配查询, 《中国计算机学会通讯》, 2012年第8卷第4期

[27] Shuai Ma, Yang Cao, Wenfei Fan, Jinpeng Huai, Tianyu Wo: Capturing Topology in Graph Pattern Matching, In VLDB Conference, 2012

[28] http://www.digitalbuzzblog.com/facebook-2010-growth-stats-infographic/

[29] http://techcrunch.com/2010/04/14/twitter-has-105779710-registered-users-adding-300 k-a-day/

[30] http://en.wikipedia.org/wiki/Facebook

[31] http://tech.sina.com.cn/i/2011-08-18/17095948514.shtml

[32] http://news.cnlist.com/CnlistNewsDetail.aspx?tablename=gsbd&GUID={99DE1B2E-0F17-4CB4- 8B49 - FAE881D02D59}

[33] http://weibo.com

[34] Maria Giatsoglou, Symeon Papadopoulos and Athena Vakali, Massive Graph Management for the Web and Web 2.0, New Directions in Web Data Management 1, Springer, 2011

[35] Mark Newman, Albert-László Barabási, and Duncan J. Watts, The Structure and Dynamics of Networks, Princeton University Press, Princeton, 2006

[36] Erhard Rahm and Hong Hai Do. Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin, 23(4): 2000, 3~13

[37] Eytan Adar and Christopher Re, Managing Uncertainty in Social Networks, IEEE Data Eng. Bull., 30(2), 2007, 15~22

[38] Gueorgi Kossinets, Effects of missing data in social networks. Social Networks 28: 2006, 247~268

[39] N. A. Lynch. Distributed Algorithms. Morgan Kaufmann, 1996

[40] Shuai Ma, Yang Cao, Jinpeng Huai, Tianyu Wo: Distributed graph pattern matching. WWW 2012

[41] Daniel Peng, Frank Dabek: Large-scale Incremental Processing Using Distributed Transactions and Notifications. OSDI 2010

[42] Wenfei Fan, Floris Geerts: Foundations of Data Quality Management Morgan & Claypool Publishers 2012