大数据时代的图搜索技术

马 帅 李 佳 刘旭东 怀进鹏
北京航空航天大学软件开发环境国家重点实验室 北京 100191

摘 要 众所周知，互联网及其应用升级推动数据量几何级数的增长，促使我们进入“大数据”时代。图数据是一类重要的大数据，与关系表格和XML等数据结构相比，图具有更强的表达能力，可以表示相对复杂的结构，所以被广泛地应用于各个领域。图的应用促使图搜索技术——“大数据”时代一种新的搜索模式的产生，并已引起业界越来越多的关注。文章介绍图搜索应用及其相关技术，揭示在“大数据”时代图搜索的重要意义及其面临的问题与挑战。

关键词 大数据时代；图搜索；社会计算

引言

在当今社会计算时代，网络和应用升级推动数据量几何级数增长，数据变得愈为重要。据Internet World统计，2012年互联网用户突破24亿，在社交网络(如Facebook)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，这些统计数据说明数据的规模越来越大。在继人力、资本之后，数据成为一种新的非物质生产要素，成为支撑科学研究和各类应用服务不可或缺的战略资源，社会计算进入了“大数据”时代。如何从海量数据中归纳、过滤信息，从而进行快速、准确的决策已经成为用户最为迫切的需求。

图的表达能力强、应用广泛，在社交网络、生物数据分析、推荐系统、复杂对象识别和软件代码剽窃检测等领域都起着重要的作用[1-4]。社会计算一般认为需要考虑社会的结构、组织和活动等社会因素，而所有的社会活动构成了社会网络，其本质上是图的一种表现形式。在社会网络中，可以把用户看作图的顶点，用户之间的关系(如朋友关系)看作图的边。图的广泛应用自然而然使得图搜索(从图中搜索信息)成为了工业界和学术界的共同关注点。

基金项目：国家优秀青年自然科学基金资助项目(61322207)；国家973计划资助项目(2014CB340304)；国家863计划资助项目(2013AA01A213)；教育部留学回国人员科研启动基金资助项目；广东省引进创新科研团队计划资助项目(2011D005)；深圳市引进海外高层次人才“孔雀计划”资助项目(1105100030834361)

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？社会计算时代出现了以社会网络为代表的新兴事物，而历史上每次出现一种重要的事物都会导致计算模式的改变，我们需要从查询搜索的历史发展来辩证地分析为什么社会计算时代需要一种新型的搜索模式。如图1所示，计算机采用的搜索方式经历了文件系统搜索→数据库搜索→Web网络搜索→社会网络搜索，这样的发展历程。

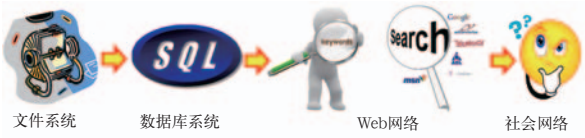

图1 搜索的演变

1) 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[5]，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。它提供简单的搜索功能，使用户可以搜索文件。

2) 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极大地提高了数据搜索的灵活性。用户可以通过SQL语言来进行各种复杂的搜索。

3) 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。

4) 社会网络。20世纪末至今，随着Web2.0和社会计算的兴起，社交网络系统开始大量应用，如Facebook、LinkedIn、人人网等。这里有一个值得指出的现象：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，图搜索是一种“大数据”时代适应社会计算的搜索方式。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，所以图搜索自然而然的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，比如知识图谱(knowledge graph)被引入Google搜索引擎来提高搜索结果的质量[6]。此外，Facebook于2013年初推出了Graph Search[7]，它的功能是让用户能搜索到社交链接上的信息，例如“我朋友都喜欢哪些纽约的餐馆”、“我去年和某某人的合照”、“我朋友去过的国家公园”等。这标志着图搜索在社会计算“大数据”时代将很有可能成为一种统一的面向社会网络的搜索模式。综上表明图搜索是“大数据”时代适应社会计算的重要搜索方式，并起到了极其重要的作用。

图搜索作为近年来逐渐兴起的一种新型搜索技术，为用户获取所需信息提供了一种方便快捷的搜索方式。图搜索的核心关键问题是建立满足新型应用需求的图搜索理论和模型，并提供高效的搜索查询技术，以提高搜索的效率和查询结果的准确性。大数据时代的图搜索理论和技术是目前国际上数据库领域的研究热点之一。本文将首先介绍图搜索及其应用，然后介绍几类具体的图搜索，进而揭示在“大数据”时代图搜索理论与技术研究的重要意义和挑战。

1 图搜索定义

图搜索指的是一类概念上非常广泛的图查询语言。图搜索中我们较常用的是子图搜索技术，即给定一个输入图，输出为给定图的子图。

下面我们首先给出图搜索一个简单的形式化定义[8]。给定一个模式图(pattern graph)Q和一个数据图(data graph)G：判断Q是否“匹配”G；或者从G中找出所有跟Q“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：图搜索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Q通常比较小，仅仅包含几个或者几十个顶点；而数据图G通常较大，甚至包含以“亿”为数量级的顶点和边。

根据图搜索的定义，很多大家常见熟知的图查询都属于图搜索，比如最短路径[9]、邻接搜索[10]、图同态及其扩展搜索[11]、子图同构搜索[12]、图模拟[13]及其扩展强模拟搜索[14]等。这类图的搜索没有明确规定查询语言的语法，比如Ad-hoc，常用于完成图中的某单项特定搜索任务。此外，模式图和“匹配”语义的不同会形成不同的图搜索语言。按照模式图结构的不同可分为：点搜索(如可达性搜索、邻居节点搜索)、路径搜索(如最短路径搜索)以及子图结构搜索(如子图同构、图模拟、强模拟)等。

下面我们介绍图搜索在多个领域中的应用情况[15]。

1) 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[16-17]。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[18-19]，如近邻搜索和图压缩[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[1]，镜像站点检测问题可以看作图的匹配问题[13]等。

2）复杂对象识别。复杂对象识别是一种脏数据清洗技术。有调查表明脏数据导致美国商业每年损失6 000亿美元[20]。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[21]。数据清洗主要包括数据修复和对象识别两项技术[2]，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搜索[12-13]等。

3) 生物数据分析。大量的生物数据可以被表示成图数据，基于图的搜索对分析生物数据有着重要的意义[3]。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，并对器官的功能组织和演化行为提供重要的参考依据[22]。

4) 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搜索可以有效地检测代码是否被剽窃[4]。感兴趣的读者可阅读文献[15]来看一些具体应用实例。

此外，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以及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，图搜索都有重要应用。

2 四类图搜索

本节我们重点介绍路径搜索、凝聚子图搜索、关键词图搜索和图匹配搜索这四类图搜索。

2.1 路径搜索

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搜索的应用领域之一。下面我们通过交通路线搜索中的一个应用实例来介绍路径搜索[23]。留美学生小明要从美国加州的Irvine前往Riverside。他的主要任务是选择合适的路线，当然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Riverside，则该问题可以用最短路径搜索来表达。即模式图Q表达的约束是从Irvine到Riverside的最短路径，数据图G就是整个美国公路路线图。用现有关于最短路径搜索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号洲际公路。

2）如果小明需要用大型卡车将一批物资尽快从Irvine送到Riverside，出于对公共健康和安全的考虑，许多桥梁和铁路交汇处是不允许这类车辆通行的。这时我们能够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。

2.2 凝聚子图搜索

凝聚子群原指社交网络整体用户中的一个子集用户群，并且该子集用户群内用户之间满足某种“紧密关系”。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，其中图的顶点表示用户，边表示用户之间的关系，比如朋友关系等。这样，我们也称凝聚子群为凝聚子图，相应的我们将从图中搜索凝聚子图的搜索称为凝聚子图搜索。

下面我们介绍几种常见的凝聚子图，并且结合著名的“Padgett's Florentine家族网络”(见图2)来解释[24]。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，并且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。家族间通过婚姻结合和商业交易结成并巩固家族间政治经济同盟关系，我们利用凝聚子图搜索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贸关系，从而更好地了解当时佛罗伦萨的历史状况[24]。

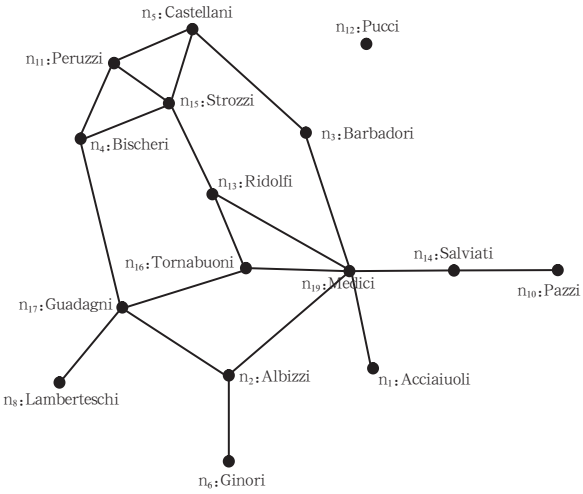

图2 Padgett's Florentine家族网络图

1) 极大团(maximal clique)。如图2所示，图中的极大团是图的完全子图，即该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。在图2中存在三个极大团：其中由Bischeri、Peruzzi、Strozzi三个家族组成含有三个顶点的极大团，任意两个家族间均存在婚姻关系，且不存在其他家族与这三个家族同时存在婚姻关系；其余两个极大团也类似。

由于团含有较好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。然而团对结构要求非常严格，从而导致图中实际存在的团通常较小，如图2中只存在三个极大团，且三个极大团分别仅含三个顶点。实际应用中，通过对团减少或者减弱约束关系形成更具广泛应用的凝聚子图，如n-极大团、n-宗派和k-极大核等。

2) N-极大团(n-clique)。图中的n-极大团是图的极大子图，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。如图2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n2:Albizzi、n4:Bischeri、n7:Guadagni、n13:Ridolfi、n16: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。

3) N-宗派(n-clan)。图中的n-宗派是图中一个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。

4) K-极大核(k-core)。图中的k-极大核为图中顶点度均不小于k的极大子图。在图2中，由n2:Albizzi、n3:Barbadori、n4:Bischeri、n5:Castellani、n7:Guadagni、n9:Medici、n11:Peruzzi、n13:Ridolfi、n15:Strozzi和n16: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较少)，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，这实际上是把搜索结果的准确性和搜索效率两个因素综合考虑的一种折中方案。最后，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。

2.3 关键词图搜索

随着各行各业互联网的覆盖和信息化技术的普及，数据量的增长速度显著加快。怎样在如此大量的信息资源中获取出需要的数据信息，成为当今的一项研究热点。“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搜索界面，它已经成为事实上的互联网数据信息检索通用机制。

关键词图搜索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，图上的关键词搜索同时考虑顶点之间结构和包含的内容两类信息，这里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

搜索的基本要求是找到的子图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搜索方法和技术。下面我们介绍三类关键词图搜索。

1) 最小树语义。目前大多数关键词搜索采用最小树语义。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[9]。

2) R半径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Steiner顶点。实际上Steiner顶点就是与K中的关键词直接或者间接相关的顶点。以Steiner顶点及其相关边构成的G的子图就称为r半径Steiner图[25]。采用这种语义的关键词图搜索输出结果是r半径Steiner图。

3) R-极大团语义。采用这种语义的关键词图搜索输出结果是前面介绍的r-极大团，该方法在图中搜索得到含有关键词的顶点，而且r-极大团顶点集合包含了所有的输入关键词，并且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[26]。

实际上，由于关键词图搜索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，从而形成了各种语义。并且，由于对用户期望的搜索结果进行猜想，搜索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。

2.4 图匹配查询

图匹配查询中，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[12]、图模拟[13]和强模拟[14]等不同的图匹配查询语言。由于图模拟和强模拟的语义相对复杂，下面我们仅仅介绍子图同构(感兴趣的读者请阅读相关文献[13,14,27])。

1) 图同构(Graph Isomorphism)。要介绍子图同构，我们首先介绍图同构(Graph Isomorphism)。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VQ与G的顶点集VG之间存在一个双射关系f:VQ→VG，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。

2) 子图同构(Subgraph Isomorphism)。给定一个数据图G和一个查询图Q，则Q与G子图同构当且仅当在G中存在一个子图Gs与图Q同构。

我们通过推荐系统中的一个应用示例来介绍图匹配，如图3所示。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3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HR1推荐SE1，DM1推荐Bio3。项目负责人希望找到满足模式图Q(图3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。

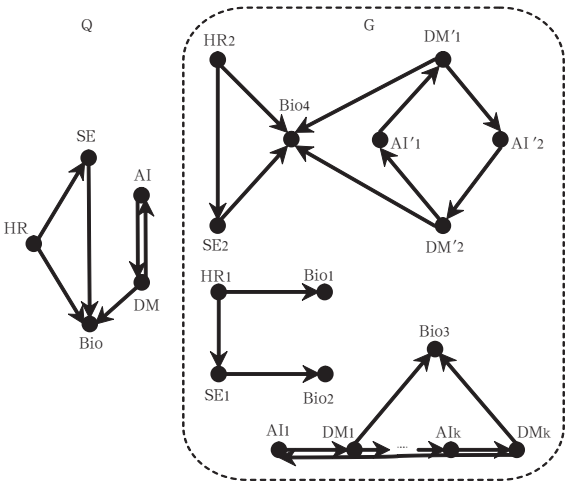

图3 图匹配查询实例

那么，当输入模式图和数据图分别为图3中Q和G时，在数据图G中执行基于子图同构的图搜索时，即“匹配”语义定义为子图同构时，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，因此搜索效率较低，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，所以通常很难在数据库中找到与查询图同构的数据子图，这些因素限制了子图同构的应用范围。因此，最近研究通过减少或者减弱约束条件来提高图匹配的实用性。目前主要有两种方法：一是引入或者提出新的图匹配模型，如图模拟和强模拟；二是近似图匹配[9]。在图3示例中，我们也可以采用图模拟和强模拟的语义来定义“匹配”，进而找出满足不同程度约束的图匹配结果，感兴趣的读者请阅读相关文献[13, 14]。

3 图搜索面临的挑战

在前面我们介绍了四类图搜索技术：路径搜索、凝聚子图搜索、关键词图搜索和图匹配搜索。我们可以发现传统的“关键词图搜索”主要用来搜索Web网络上的信息，而这些信息通常是相互孤立或者是“物-物弱关联”的，是关注“历史和存在”的信息。社会计算需要考虑社会的结构、组织和活动等社会因素，这样社会搜索中“关系”变得非常重要，而社会数据通常是“人-人强关联”或者是“人-物强关联”的，是关注“关系”和“未来”的信息。这使得传统的“关键词图搜索”不能满足社会计算时代社交网络和微博等社交媒体对社会搜索技术提出的普遍挑战。

社会网络是一种天然的图，从图数据的角度，结合传统的图搜索技术，为解决这个问题提供了新思路，如Facebook最近推出的社会搜索“Graph Search”也佐证了这一点。我们通过调查发现图搜索理论及相关技术在各个领域中有着重要的商业价值。而大数据时代图数据所具有的特点也对传统的图搜索理论和技术提出了挑战。

首先，图数据的处理要比XML数据困难的多，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，大数据时代的新应用对图搜索技术提出了新的挑战。比如在社交网络等新的应用中，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11亿[28]，每秒新增用户7.9个，每天新增用户超过600K[29]；Twitter用户超过5亿，每天新增用户超过100K[30]；人人网的用户数量总计已经超过2亿[31]；新浪微博用户已经超过5亿[32-33]。这些统计数据说明了：1)数据的规模越来越大，达到了前所未有的以亿为数量级的规模[34]；2)更新非常频繁，每时每刻都在发生更新，并且每天更新的规模也达到了以10万为数量级的规模[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。即，图数据具有大规模性、动态性和不确定性三个主要特点。第一个特点要求图搜索的效率要高，并且要充分权衡搜索效率和存储空间大小之间的关系；第二个特点要求图搜索要充分考虑动态变化因素和时序特征；而第三个特点则要求图搜索要解决图数据质量问题。

4 结语

本文介绍了大数据时代的新型搜索模式：图搜索。从分析图搜索在当前工业界的应用和学术界的研究发展动态，以及搜索的历史发展两个方面，我们看到了当前大数据时代图搜索的重要性，同时，我们也看到了图搜索所面临的挑战。可以看出，大数据时代图的搜索理论及相关技术是一个亟待研究和解决的内容，具有重要的科学意义和应用价值。

参考文献

[1] Adam Schenker,Mark Last,Horst Bunke,et al.Classification of Web Documents Using Graph Matching[C]//IJPRAI Conference,2004

[2] Fan Wenfei,Li Jianzhong,Ma Shuai,et al.Interaction between Record Matching and Data Repairing[C]// SIGMOD Conference,2011

[3] Patrick Durand,Laurent Labarre,Alain Meil,et al.GenoLink:a graph-based querying and browsing system for investigating the function of genes and proteins[J].BMC Bioinformatics,2006(7):21

[4] Liu Chao,Chen Chen,Han Jiawei,et al.GPLAG:detection

of software plagiarism by program dependence graph analysis[C]//KDD Conference,2006

[5] Per Brinch Hansen.Classic Operating Systems[M].New York:Springer-Verlag,2001

[6] 知识图[EB/OL].[2013-11-15].http://www.google.com/insidesearch/features/search/knowledge.html

[7] Facebook Graph Search[EB/OL].[2013-11-16].http://en.wikipedia.org/wiki/Facebook_Graph_Search

[8] 马帅,李佳,刘旭东,等.图查询:社会计算时代的新型搜索[J].中国计算机学会通讯,2012,8(11):26-31

[9] Charu C,Aggarwal,Haixun Wang.Managing and Mining Graph Data[M].New York:Springer-Verlag,2010

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

[11] Wenfei Fan,Jianzhong Li,Shuai Ma,et al.Graph Homomorphism Revisited for Graph Matching[C]//VLDB Conference,2010

[12] Brian Gallaghe.Matching structure and semantics:A survey on graph-based pattern matching.AAAI FS,2006

[13] Monika Rauch Henzinger,Thomas A Henzinger,Peter W Kopke.Computing Simulations on Finite and Infinite Graphs.C//Proceedings of the 1995 IEEE 36th Annual Symposium on Foundations of Computer Science, 1995:453-462

[14] Ma Shuai,Cao Yang,Fan Wenfei,et al.Capturing Topology in Graph Pattern Matching[C]// VLDB Conference,2010

[15] 马帅,曹洋,沃天宇,等.社会网络与图匹配查询[J].中国计算机学会通讯,2012,8(4):20-24

[16] Facebook Statistics,Stats & Facts For 2011[EB/OL].[2013-11-10].http://www.digitalbuzzblog.com/facebook-statistics-stats-facts-2011/

[17] The Social Impact of the Internet on Our Society[EB/OL].[2013-11-11].http://www-users.math.umd.edu/~bnk/CAR/project.htm

[18] Tian Yuanyuan,Patel Jignesh M.TALE:A Tool for Approximate Large Graph Matching[C]//ICDE Conference,2008

[19] Barceló Pablo,Libkin Leonid,Wood Peter T,et al.Expressive languages for path queries over graph-structured data[C]//PODS Conference,2010

[20] Eckerson W.Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High Quality Data.The Data Warehousing Institute,2002

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

[22] Bader David A,Madduri Kamesh.A graph-theoretic analysis of the human protein-interaction network using multicore parallel algorithms[J].Parallel Computing(PC),2008,34(11):627-639

[23] Rice Michael N,Tsotras Vassilis J.Graph Indexing of Road Networks for Shortest Path Queries with Label Restrictions[C]//VLDB Conference,2010

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

[25] Li Guoliang,Beng Chin Ooi,Feng Jianhua,et al.Ease:Efficient and adaptive keyword search on unstructured,semi-structured and structured data[C]//SIGMOD Conference,2008

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

[27] Fan Wenfei,Li Jianzhong,Ma Shuai,et al.Graph Pattern Matching:From Intractable to Polynomial Time[C]//VLDB Conference,2010

[28] Facebook 2010 Growth Stats:Infographic[EB/OL].[2013-11-25].http://www.digitalbuzzblog.com/facebook-2010-growth-stats-infographic/

[29] Facebook[EB/OL].[2013-11-25].http://en.wikipedia.org/wiki/Facebook

[30] Twitter has 105,779,710 Registered Users, Adding 300K A Day[EB/OL].[2013-11-25].http://techcrunch.com/2010/04/14/twitter-has-105779710-registered-users-adding-300 k-a-day/

[31] 人人公司公布2013年第三季度未经审计财务报告[EB/OL].[2013-11-25].http://www.renren-inc.com/zh/news/121.html

[32] 新浪微博注册用户已超5亿[EB/OL].[2013-11-25].http://tech.sina.com.cn/i/2013-02-25/09348086534.shtml

[33] http://weibo.com

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

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

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

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

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

作者简历

马　帅

CCF会员，博士，北京航空航天大学计算机学院教授，博士生导师。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。

李　佳

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。

刘旭东

CCF会员，博士，北京航空航天大学计算机学院教授，博士生导师。主要研究方向为可信网络计算技术和中间件技术等。

怀进鹏

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

Graph Search in the Big Data Era

Ma Shuai
Li Jia
Liu Xudong
Huai Jinpeng

SKLSDE Lab, Beihang University, Beijing 100191, China

Abstract As we know, we have entered into the “Big Data” era. Compared with RDB and XML, graphs have more expressive power, and can represent more complex structures. Hence, graphs are widely applied in many fields of computer science and beyond. The wide use of graphs has brought about the emergence of “graph search”, a new searching paradigm, which has drawn more and more attentions from both industrial and academic communities. In this article, we focus on the applications and techniques of graph search, along with the challenges to be solved.

Keywords Big Data; Graph Search; Social Computing