ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.17, No.8, August 2006, pp.1717−1730
DOI: 10.1360/jos171717

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

P2P 数据管理*

余 敏+, 李战怀, 张龙波

(西北工业大学 计算机学院,陕西 西安 710072)

P2P Data Management

YU Min+, LI Zhan-Huai, ZHANG Long-Bo

(School of Computer, Northwestern Polytechnical University, Xi'an 710072, China)

+ Corresponding author: Phn: +86-29-88493772, E-mail: yum@mail.nwpu.edu.cn, http://www.nwpu.edu.cn

Yu M, Li ZH, Zhang LB. P2P data management. *Journal of Software*, 2006,17(8): 1717−1730. http://www.jos.org.cn/1000-9825/17/1717.htm

Abstract: P2P (peer-to-peer) is the key technology of reconstructing the future distributed architecture and has a good application perspective. As the issues in P2P systems mostly come down to data placement and retrieval, P2P data management has recently become an active topic in database community. In this paper, the advantages of P2P systems are first described. Then the goals of P2P data management researches are presented. Thirdly, research of P2P data management is described from three facets, i.e. P2P information retrieval, P2P database-style queries and P2P continuous queries. Particularly, the index construction methods, semantic coordination, query semantics, query processing strategies, types of queries supported, and query optimization of P2P database-style queries are discussed in detail. Finally, the issues to be further studied are proposed.

Key words: P2P (peer-to-peer); index; semantic coordination; query semantics; complex query processing; query optimization

摘 要: P2P(peer-to-peer)技术是未来重构分布式体系结构的关键技术,拥有广阔的应用前景.P2P 系统的大多数问题都可归结为数据放置和检索问题,因此,P2P 数据管理成为数据库领域活跃的研究课题.当前,P2P 数据管理主要有信息检索、数据库查询和连续查询 3 个子领域,取得了许多研究成果.在介绍 P2P 技术的优点后,指出了 P2P 数据管理研究的目标.然后针对上述 3 个方面,论述 P2P 数据管理研究的现状,着重讨论了 P2P 数据库查询的索引构造策略、语义异构的解决方法、查询语义、查询处理策略、查询类型和查询优化技术.通过比较,指出了现状与目标的差距,提出了需要进一步研究的问题.

关键词: P2P(peer-to-peer);索引;语义调和;查询语义;复杂查询处理;查询优化
中图法分类号: TP311 文献标识码: A

随着计算机技术的发展,在当今的软、硬件技术环境下,客户/服务器模型已不能满足需求,其单点故障和热点问题已经变得越来越不可接受.Peer-to-Peer 模型(又称 P2P 模型或对等计算模型)是一种新型的体系结构模型,其许多优势有待进一步发掘[1]:首先,P2P 系统的每个成员均可贡献数据和计算资源(例如,未用的 CPU 周期

* Supported by the National Natural Science Foundation of China under Grant No.60573096 (国家自然科学基金)
Received 2005-05-27; Accepted 2006-03-09

和存储资源),新成员的加入可能引入系统中原来缺乏的特殊数据或资源,随着系统成员的增加,系统的丰富性、多样性等各种有益的特性得以扩大;其次,P2P 系统具有分散性,系统的健壮性、可用性和性能可能随着 peer 的数量增加而有所扩展;另外,通过在许多 peer 间路由请求和复制内容,系统可以隐藏数据的提供者和消费者的身份,使个人的隐私得到保护[1].因此,P2P 被认为是未来重构分布式体系结构的关键技术[2],它在搜索引擎、数据流管理、语义网、协作信息过滤等领域具有广阔的应用前景.

Napster,Gnutella 和 KaZaA 等 P2P 系统的广泛应用以及 P2P 计算的潜在应用前景,使得 P2P 计算成为计算机科学极端活跃的研究课题,影响着网络、分布式系统、信息系统、算法和数据库等诸多领域.最初,P2P 计算研究围绕重叠网络(overlay network)构建进行,2002 年前后,一些研究人员[3]对其研究进展及 P2P 文件共享系统进行了综述,这个阶段的 P2P 系统缺乏语义支持,既不能很好地满足用户的需求,也不能有效地利用系统的资源.由于 P2P 系统的大多数问题都可归结为数据放置和检索问题,数据库研究人员加入到 P2P 计算研究的行列,出现了 PeerDB,Hyperion,Piazza 等 P2P 数据管理项目,引起 P2P 系统从文件共享向复杂查询处理的转化,促进了 P2P 计算向资源发现、网络监视、语义网构建及 AmI (ambient intelligence)中的协作信息过滤等领域发展.人们对 P2P 系统的地位有了全新的看法[4],如图 1 所示.

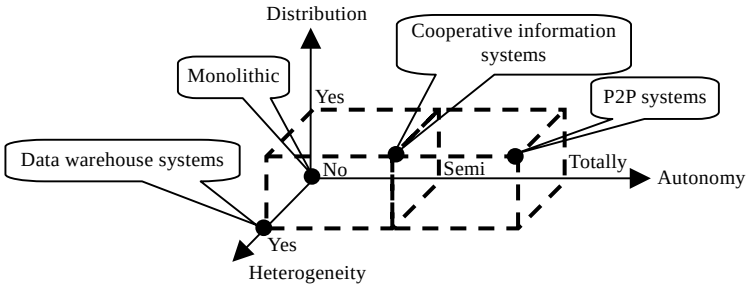

Fig.1 The position of P2P system

图 1 P2P 系统的地位

本文第 1 节指出 P2P 数据管理的研究目标.第 2 节分别从即兴查询支持和连续查询支持两个方面论述 P2P 数据管理的研究现状,涉及索引构造策略、语义异构调解策略、查询处理语义、查询处理方法等内容.第 3 节指出现状与目标的差距,讨论需要进一步研究的问题,展望未来的研究方向.

1 P2P 数据管理研究目标

目前,P2P 上的数据管理研究非常活跃,有许多研究项目:美国的 Peers[5],PEPPER[6],PIER(peer-to-peer information exchange and retrieval)[7],Piazza[8],RDFPeers[9] 等;加拿大的 Hyperion[10];德国的 Edutella[11],GridVine[12],P2P-DIET[13]等;新加坡的 CQ-Buddy[14]以及与中国合作的 PeerDB[15]等.尽管各项目的目的和侧重点不同,但是提供下面这些特征是 P2P 上的数据管理研究共同的目标:

1) 可扩展性:查询处理的性能和服务质量不因网络规模扩大而显著下降;
2) 自治性:查询处理算法应在一定程度上尊重 peer 的自治性;
3) 效率:以消耗的绝对资源(带宽、处理能力、存储等)度量效率,为了保证高吞吐量,必须有效使用资源,降低系统开销;
4) 服务质量:它是用户可感知的质量,可以用不同的指标(结果的数量、响应时间等)度量[16];
5) 健壮性:包括对故障健壮和对攻击健壮,在面临 peer 故障、离开、攻击时系统保持可用性、服务质量和效率;
6) 支持语义异构 peer:语义异构是 peer 自治的结果,各 peer 用自己的模式组织和呈现数据,为了互操作,需要适应 P2P 动态即兴环境的解决方案;

7) 查询处理能力:可以从其支持的被查询的数据类型和查询算子类型来体现,文献[17]指出所有的数据都可以视为关系数据,因此支持对关系数据的关系完备查询具有普遍性,同时适用于数据库查询和文本搜索[17].

这些目标很难同时达成,现在的研究项目分别从 P2P 网络的语义重叠网络和重叠网络两个方面改善现有 P2P 网络的性能和服务质量,它们或者解决语义互操作性而只提供有限的可扩展性(Hyperion, Edutella,PeerDB,Piazza 等),或者提供高可扩展性但不处理语义互操作性(PIER,RDFPeers 等).

2 P2P 系统中的查询

2.1 即兴查询

2.1.1 信息检索

信息检索(information retrieval,简称 IR)过程由元数据搜索、文档位置搜索和文档下载组成[18].IR 包括布尔检索和分级检索,目前已舍弃布尔检索而更加倾向于分级检索[19].

传统搜索引擎不能提供最新结果,或不能搜索动态产生网页的站点的后台[20].与传统搜索引擎不同,P2P 网络规模更大,由成千上万高度动态且通常异构的节点组成,节点使用公共的资源描述语言,相关站点各自回答查询,没有传统搜索引擎中常见的负责选择文献集、重写查询和合并分级结果的集中式中介[21],没有集中的仓储,没有集中故障点和全局的管理器,各查询总是获得最当前结果.2004 年,Weikum[19]指出向搜索引擎中引入统计信息将改善信息检索的效率和质量,这些统计信息主要从用户的查询日志和点击操作流中获得,P2P 有助于消除大的 ISP(Internet service provider)对获得有用统计信息的垄断,因此,下一代 Web 搜索引擎是 P2P 搜索引擎,且在 P2P 搜索中应用统计信息是今后计算机科学的主要研究内容.

P2P 信息检索主要有贝叶斯模型(Bayesian model)、向量空间模型(vector space model,简称 VSM)和潜在语义索引模型(latent semantic indexing,简称 LSI)[18,22].文献[18]详细综述了当前 P2P 信息检索的研究进展,因此,本文只总结文献[18]以后的新进展.为了将语义丰富的信息检索技术与 P2P 系统集成,凌波等人[2]在 BestPeer 平台上开发出 PeerIS,认为采用基于语义的自配置机制,可以使 peer 能够根据信息偏好、行为和查询统计数据综合地确定和调整自己的重要节点,从而以最小的代价检索到所需数据.文献[2]用类似于文本分类法的方法确定并用 VSM 表达语义类别,用本地文件的语义类别分布来定义 peer 的兴趣偏好.Heng Tao Shen 等人[23]分别用 VSM 和 LSI 分两步构造概要,将大的文档表示、peer 表示和 super-peer 表示转化成小的高维点,扩展了 VA-file 索引技术来索引转换后的点,从而支持高效的 peer 和文档搜索.

总的来说,P2P IR 系统通过 3 种策略实现资源定位和路由:(1) 从查询 peer 向所有邻居 peer 迭代地洪泛查询,直到达到 TTL.这种方法简单且在 peer 加入和离开系统时健壮,但是可扩展性差,且只提供在查询 peer 的有限半径内的查询回答.(2) 使用分布式哈希表(distributed hash table,简称 DHT).当 peer 上的新数据到达或者新 peer 加入网络时,网络中所有的新内容都发布到对应 key 的节点.在 peer 离开时,相关内容必须取消发布.这样,当一个新文集被加入 peer 时,通常含有大量需要被索引的术语,由于由哈希函数决定每个术语发布到哪个 peer 上,故完全发布一个新文档需要寻址大量 peer[24,25].因此,当请求大量复制的项时,DHT 途径效率低下.并且,DHT 名字空间的虚拟性使得浏览导航或者预取等搜索过程复杂化[24].(3) 基于兴趣创建捷径的路由索引.这种索引没有 DHT 那样的发布和取消发布的开销,索引项的查全率相当高,但是需要处理好统计信息维护开销和查全率之间的折衷.

然而,全面的关键字搜索仍是悬而未决的问题[16].分级关键字搜索需要对文档整个集合的全局统计(如文档频率),目前还没有收集和维护统计信息的健壮、高效、分布式方法.考虑到在 P2P 网络动态环境下维护统计信息的难度,文献[24]指出应用启发式技术是很有前途的途径:第一,通过估计 peer 的文档集合的重叠程度,将查询发到能向回答集中添加新方面的 peer.该方法还需要继续研究如何提高估计的准确程度来改善结果的质量;第二,在提供相似类型文档的 P2P 系统中引入公共范畴分类学,在路由索引中包含范畴信息,且在查询中指定结果应隶属的范畴,从而仅将查询发到查询中指定范畴及其父范畴的 peer.此方法仍需进一步地研究如何有效设定

范畴分类系统.

2.1.2 数据库查询

P2P 系统的大多数问题都可归结为数据放置和检索问题[26],但 P2P 网络本身缺乏对语义、数据转换和数据联系的支持,而处理这类问题是数据库研究团体的强项.因此,2001 年 Gribble 等人[1]提出将 P2P 技术与数据库系统进行联合研究[27].

Harren,Hellerstein 等人[17]指出,要求终端用户将数据加载到数据库中生成 P2P 数据库,会增加软件的复杂性,引起数据库存储相关的管理需求.因此,他们认为可以单独处理“查询处理”,将它和存储、事务管理分离,利用现存数据自然体现的属性构造查询引擎,智能地查询数据而提供用户习惯的存储语义.因此,P2P 上提供数据库查询操作符的系统并不一定由数据库组成,故统称为“P2P 数据管理系统”.

数据库团体对 P2P 系统的主要贡献是:引入了模式的概念[26],出现了基于模式的 P2P 数据管理系统(peer data management system,简称 PDMS),各 peer 用各自的模式表示自身的兴趣域,在 peer 对或 peer 小集合间局部提供语义映射,通过遍历映射路径,用一个 peer 模式提出的查询可检索到网络中可达的所有 peer 上的相关数据[8].从功能上看,基于模式的 PDMS 通常由描述存储的数据的模式语言、检索数据的查询语言、路由查询的网络拓扑及集成各不同 peer 的信息的技术组成[11].与简单的 P2P 网络相比,基于模式的 PDMS 具有许多优点:支持任意元数据模式和本体,允许对资源复杂的可扩展的描述,在原本固定有限的描述中引入了动态性,可以支持复杂的查询而不只是简单的基于关键字的搜索,为现存信息源增值留下了广阔的空间[28],是充当连接异构信息源的基础设施的理想之选[11].

基于模式的 PDMS 与分布式数据库系统(distributed database system,简称 DDBS)有相似之处,但 P2P 网络巨大的规模和 peer 的不可靠本质使得两者之间存在如下差异:

1) 在 DDBS 中,节点通常是稳定的,以受控的方式加入网络和退出网络;而在 P2P 系统中,节点随时可以加入和离开网络[15],因此成员关系是即兴的、动态的,很难预言或推理资源的位置和质量,当系统中数据变化率较高时,随着 peer 数量的增加,维护全局可访问的索引就不再可能.
2) 在 DDBS 中,节点通常知道一个共享的模式;而在 P2P 系统中,节点间通常没有预定的全局模式[15].
3) 在 DDBS 中,可检索到满足查询的全部回答;而在 P2P 系统中,节点不含有全部的数据,节点可能脱机,因此,通常不能检索到满足查询的全部回答[15],查询结果的正确性和完整性概念不能用传统数据库中的纯粹含义,P2P 查询结果极大地依赖于瞬间网络和已建立的语义映射[29].
4) 在 DDBS 中,通常能够确切知道可回答查询的节点的位置;而在 P2P 系统中,不存在协调全部 peer 的万能节点,节点通过将查询转发到邻居,逐步定位内容[15],实现高效的查询处理更困难,网络中存在很多冗余信息,会带来数据和计算冗余,数据冗余可通过严格控制数据放置解决,计算冗余主要通过查询优化和查询处理节点间交换信息来解决[7],系统分散性要求分布式优化.
5) P2P 系统中要解决的扩展性和 DDBS 领域的扩展性不同,DDBS 中规模指标主要是存储的字节数,而 P2P 系统的规模指标中参加的主机数比存储的字节数更重要.

因此,DDBS 的许多技术不能用于 PDMS,研究人员需要针对 PDMS 的上述特点研究新的方法.

在这一节的其余部分,我们将从索引构造策略、语义异构解决方案、查询语义、查询处理及优化这 5 个方面论述这些挑战的处理策略.

2.1.2.1 索引构造策略

Hellerstein[30]指出:数据的独立性原理是 P2P 的核心思想,通过索引和查询优化实现.在 P2P 网络中,数据分布在网络中的许多节点上,发起查询的节点将查询再形式化后发送到所有可能具有回答的 peer,在 P2P 网络中求值该查询.此时,避免查询洪泛整个网络至关重要,为此,需要关于 peer 上存储的数据的信息,需要构造 peer 上的索引.

目前主要有两大类索引:模式级索引和实例级索引.前者是基于模式的 PDMS 所特有的,考虑了语义异构问题,它是实现异构环境下查询处理的关键;后者假定索引涉及的所有节点均采用相同的模式,仅提供该模式下数

据分布的信息,它可以有不同的粒度,以便允许用户以不同的准确级别表达查询[8].Edutella[11]使用具有超立方体结构的 super-peer 网络充当网络拓扑,各 super-peer 维护 Super-peer/Peer 索引和 Super-peer/Super-peer 索引,都包括模式索引和实例级索引(property/property set 索引、property value range 索引和 property value 索引).Super-peer/Peer 索引描述 peer 上存储的内容,由 peer 注册时提供给 super-peer;Super-peer/Super-peer 索引由 Super-peer/Peer 索引抽取和概括而成,用于指导 super-peer 骨架内查询路由,避免广播.大多数 DHT 索引以及 P-Tree,P-Ring 索引等都是实例级索引,可分成散列索引和树型索引[31]两种.散列索引是通过散列将原始数据的标识符映射到 ID 的集合,再将 ID 集合映射到负责该 ID 的 peer 产生的,可以具有多种拓扑形态,例如,CAN (content-addressable network)是 d 维空间,Chord 是环.使用这种索引,各节点只有到其他节点的少量传输连接,采用贪婪的路由策略,每一跳查询都向目标接近,最小化额外迂回造成的延迟.树型索引包括 P-Tree[6]和 P-Ring[32,33],它们本质上是数据库中 B+树思想在 P2P 环境下的扩展,能保序地在 peer 之间分布数据项,支持范围查询,并保持大致的存储平衡.其中,P-Tree 定位于每个 peer 只提供一个数据项或者服务的应用[6],而 P-Ring 则高效地支持每个 peer 有大量数据项的情况[33].这些索引的优点是对已知 key 的访问非常高效.但是,为了回答含有任意约束的查询,将需要为每个属性维护单独的索引,在网络和/或数据变化有限时工作良好,但在其他情况下维护代价过大.

实例级索引有多种分布方法:本地分布、集中式分布和分布式分布[22].使用本地索引,各 peer 只维持对自己的数据的引用,而不管其他节点上的数据的信息,这使得 peer 充分自治,peer 能够处理丰富的查询,但是,查询通常通过洪泛获得回答,通信流量大却不保证一定能找到要查找的数据(即使数据存在于网络中).为了提高效率和查询命中率,文献[34]提出了迭代加深方法,文献[35]提出了随机游走方法,文献[36]用 *k*-随机游走改善随机游走的性能,文献[37]提出概然洪泛算法,性能有所改善.使用集中式索引,单个服务器上保存许多 peer 上数据的引用,索引是集中的,但数据是分布的,客户可批量或者增量地更新集中式索引[22].集中式索引容易成为瓶颈和造成单点故障,严重限制了系统的可扩展性和健壮性.使用分布式索引,指向目标的指针存放在多个 peer 上,可有效分散查询负载和索引的存储负载,具有更强的健壮性和可扩展性,大多数 DHT 索引、P-Tree 和 P-Ring 都是分布式索引.

总的来说,当前只有基于模式的索引考虑了模式异构问题、提供语义支持(但当前此类索引的可扩展性有限),其余各类索引均假定系统中所有节点都采用相同的模式,仅提供数据级索引.该领域存在对语义异构支持和高可扩展性脱节的问题,因此不能满足大规模异构信息源 P2P 网络中共享信息的需求.

2.1.2.2 语义异构解决途径

通常,不同数据源间语义异构可能出现在两级[38]:

1) 模式级

数据源可在关系、属性/标签名、数据的规范化程度、详细级别和特定领域的覆盖率上不同.调解模式级异构的问题称为模式匹配(schema matching).

2) 数据级

相同的现实世界的实体使用不同的表示,调解数据级异构的问题称为实体/对象匹配(entity/object matching 或 data deduplication 或 record linkage).

传统信息集成领域采用集中式的策略解决语义异构:采用中介系统,将全局查询分解后发给各数据源;或采用物化的途径,将数据源集合到一个中心位置.这些方法难于扩展,且大规模系统中难于就全局模式达成共识,受模式组成源的频繁变化影响,提供模式和源信息的中心模块会成为瓶颈,或造成单点故障,故无法解决 P2P 网络中节点语义异构的问题,需要信息集成研究扩展到 P2P 领域.

目前,P2P 上解决语义异构主要是通过映射来实现的.针对两级语义异构,有两种映射:

1) 使用模式级映射来表示异构模式之间的关系[10].这种映射可在查询过程中构造:当查询被提出时,进行模式匹配路由来搜索具有匹配模式的节点[15];也可由各节点在查询前预先决定自身模式和邻居的模式间的映射(在这种情况下,邻居共享共同的语义内容)[8,28,39],通过传递关系,获得新的映射.

2) 使用称为映射表的数据级映射,表达两个数据源中数据值间的数据关联,通过检查关联的一致性、从已存在的关联推导出新的关联来自动化映射表的管理[10].

Edutella[11]为语义网提供基础设施.该项目通过采用 RDFS(RDF(resource description framework) schema)实现了基于类、特性或者特性约束来表示模式,定义了描述资源所使用的词汇表[11].Edutella 探索了基于不同种类的中介 peer 的两条研究路线: (1) 建立在显式的中介 peer 之上:采用 PSELO(personalized search engine for learning objects)中介引擎翻译通用 Edutella 查询;(2) 建立在所有的 super-peer 中基于规则的中介之上:采用 super-peer 网络作网络拓扑,经过基于规则的聚簇策略限制各 super-peer 的模式和属性的数量,减少信息集成的工作量,在各 super-peer 上引入全局模式,并映射 peer 模式到该模式解决异构问题.

2002 年,Bernstein 等人提出了局部关系模型(local relational model,简称 LRM)[39],使用 peer、熟人(一组相关联的 peer)、调和公式(peer 和其熟人间的语义依赖)和域关系(peer 和其熟人间的数据翻译规则)表达该模型.受到 LRM 的启发,Hyperion[10]通过 3 种约束数据交换的机制解决语义异构的问题:(1) 映射表:它是提供熟人 peer 的数值间对应关系的二元表,不仅提供数据级映射,而且隐含地提供根本的模式级映射;(2) 映射表达式:使用类 Datalog 语法表达两个不同模式的关联,实现两个不同 peer 间的模式级对应关系.目前,Hyperion 在建立 peer 间熟人关系时半自动创建映射,然后通过人工干预来确保所产生映射的正确性;(3) SQL3 触发器(或 ECA 规则)[40]:用于确保 peer 间的一致性.通过映射表和映射表达式,用单个数据源的模式提出的查询被翻译成一组可在相关数据源上执行的查询,从而获得网络中其他 peer 上的数据.文献[10]给出“正确的翻译”(只检索正确的回答)和“完备的翻译”(检索到全部正确回答而没有不正确的回答)的概念来评价翻译,提出计算完备翻译的算法和测试是否为正确翻译的算法.

Piazza[8]用 XML 建模数据,peer 使用 XML 模式表达自己的模式.它采用 peer 描述和存储描述两级映射解决语义异构问题:“peer 描述”是各 peer 的“世界的视图”之间的映射,用于路由查询;“存储描述”将一个 peer 上存储的数据映射到该 peer 的“世界的视图”.Piazza 首先通过模式匹配产生基础映射,再经过人工干预或者自动技术纠正产生的映射.Piazza 采用 PPL (peer-programming language)来表达映射,这种语言合并了传统数据集成中 LAV(local-as-view)和 GAV (global-as-view)的重要特性:为了使查询回答简单,它采用类似于 GAV 的方式相对于源模式定义目标模式;为了能够在源模式为目标模式的投影或选择的情况下将源模式映射到目标模式,引入 LAV 的形式.并且,Piazza 关心 XML 和 RDF 之间的关系,提出在 XML 和 RDF 节点(node)之间建立映射的算法.

GridVine[12]有两种解决语义异构的机制:(1) 模式继承(schema inheritance):它允许用户从已存在的模式派生出新的模式,重用属于基模式的概念集合,从而隐含地促进互操作性;(2) 语义传播(semantic gossiping):从模式间纯局部映射图开始,以分散的方式促进全局互操作,它是 GridVine 的研究人员在 Chatty Web[28]中提出的.Chatty Web[28]假定不同模式间存在局部映射(可通过人工或者半自动方式建立),映射是选择投影查询,Chatty Web 关注语义映射信息的传播(gossip)协议(即语义传播),目标是随时间推移逐步改进映射的质量,通过局部交互达到全局互操作.Chatty Web 定义了两个评价翻译质量的标准:(1) 内在(intrinsic)标准:即语法相似性,只与被翻译的查询和需要的翻译有关,用于量度翻译造成的信息损失;(2) 外在(extrinsic)标准:即语义相似性,与不同 peer 在特定翻译上达成的语义共识的程度有关.系统中的 peer 在收到查询时,根据这两个标准决定将查询发到哪里;当 peer 收到结果或者反馈(例如,翻译图中存在环,使查询能回到查询发起者或者经过多个翻译环发起查询的 peer 获得返回的查询和数据)时,分析模式级和数据级结果的质量,据此调整标准,以期最终仅将查询转发给最有可能理解该查询的 peer,且通过调整 peer 的每一跳的转发行为不断添加正确的映射,隐含达到全局互操作.目前,语义传播有迭代和递归两种实现策略:使用迭代策略,发起查询的 peer 自己找到并自行处理全部翻译链;使用递归策略,发起查询的 peer 委托其他 peer 翻译查询.实验表明,无论初始延迟还是返回的结果数量的可扩展性,递归实现都优于迭代实现[12].

PeerDB[15]采用全自治的方式描述元数据,各 peer 持有两个元数据集合——本地字典(local dictionary)和导出字典(export dictionary),分别描述本地可访问的数据和能被网络中其他 peer 访问的数据.它使用关系语言描述模式,用户在创建表时,为表和属性指定一些关键字充当同义辞典,利用基于信息检索的关键字匹配途径实现

模式级映射.由于它假定在整个 peer 网络中一致地使用关键字,因此不能处理 peer 的数据值中使用不同字汇表的情况[10].并且,关键字匹配可能会产生不相关的查询再形式化(reformulation)结果,需要用户决定执行哪个查询[12].

Calvanese 等人[41]指出:用一阶语义(FOL(first-order logic)语义)解释 P2P 系统的信息集成从形式角度来看是正确的,但在建模和计算方面存在如下缺点:对模块化、一般性和可判定性支持较弱,将 P2P 系统视为单一平面逻辑理论,因此丢失了系统中以 peer 表达的结构,且远程连接可能传播对一个 peer 的语义产生深远影响的约束;在任意 P2P 互联下,一阶语义的查询回答都是不可判定的,即使单个 peer 具有极端严格的结构.文献[41]提出基于认知逻辑的语义,在模块化、一般性和可判定性方面,基于认知逻辑的途径都优于 FOL 语义途径.在 peer 具有可判定的模式和连接的映射但任意相连的系统中,FOL 途径会导致查询回答不可判定,而基于认识逻辑的方法总保持可判定性.

当前,语义异构解决方案方面的研究主要集中在如何在 peer 间建立映射,以便提供 peer 间的语义互操作上.目前使用的映射有很大的局限性:多数现有系统使用的模式元素间的映射就是简单的相等或者包含语句,而使用更复杂映射(特别是合取查询)的途径不能扩展到大量数据源[42].此外,无论是查询前预先确定语义映射,还是在查询提出后根据关键字匹配确定映射,都不能依查询来解释语义冲突.这里通过例子来说明什么是依查询解释语义冲突.例如,有两个数据库 DB1 和 DB2,其模式如图 2 所示.在提供关键字或者预定义映射的情况下,DB1 中的 University 相对于 Faculty 的作用和 DB2 中 Employer 相对于 Employee 的作用很相似,因此会断言 University 和 Employer 是同义词.这个断言对于“List Names of Employers of Engineering Related Professionals”这样的查询是正确的,然而对于查询“List Names of people who work in Academic Institutions”,这个同义词关系不再成立,使用这个同义词断言将产生错误的结果.此外,随着网络进化会产生大量的语义映射,需要研究管理这些映射的技术:找出应该建立而尚未建立映射的节点集,寻求处理不同质量的映射的途径,逐步改善映射,从而改进节点从其他节点获得数据的能力.

DB1:
Faculty (SS#, Name, Dept, Sal, Sal_Type, Affiliation, University,…)
Faculty: each tuple stores the information of a faculty member;
SS#: the identifier of a faculty member (primary key);
Name: the name of a faculty member;
Dept: the department to which a faculty member is affiliated;
Sal: the amount of annual salary paid to a faculty member;
Sal_Type: the type of salary, such as base salary, bonus;
Affiliation: the affiliation of a faculty member, such as teaching, research;
University: the university where a faculty member is employed; …

DB2:
Emloyee (ID, Name, Type, Employer, Dept, CompType, Comp, Affiliation,…)
Employee: each tuple stores the information of an employee;
ID: the identifier of an employee (primary key);
Name: the name of an employee;
Type: the job category of an employee, such as executive, manager;
Employer: the name of the employer, such as Microsoft, Motorola;
Dept: the department where an employee works;
CompType: the type of compensation given to an employee, such as base salary;
Comp: the amount of annual compensation for an employee;
Affiliation: the name of university where an employee works as a visiting professor; …

Fig.2 Schemas of DB1 and DB2

图 2 DB1 和 DB2 的模式

2.1.2.3 查询语义

分布式查询传递不完全同步产生了分布的语义.文献[39]用关系空间(relational space)定义 LRM 的模型理

论语义,一个关系空间是关系数据库的一个有限集合.它使用一阶逻辑语言描述每个数据库的模式,用调和公式(coordination formula)表达关系空间中的数据库间的协调以及查询.此时,查询的全局回答是这样计算的:将调和公式中各原子公式分别在该公式指定的数据库上本地求值,然后按照组成调和公式的连接词和量词,递归合成和映射这些结果.文献[10]定义映射表逻辑语义,给出处理本地查询和全局查询语义的形式化表示:本地查询类似于集中式系统中的查询,执行时只使用本地 peer 的数据;当执行全局查询时,用 peer 网络中驻留在其他 peer 上的数据来补充本地检索到的数据,允许用不同模式返回结果.因此,本地查询可视为平凡的全局查询.在 PIER[7]中,Ryan 等人认为在 P2P 环境中 ACID(atomicity, consistency, isolation, and durability)特性不可能同时实现,因此放松一致性来达到其他特性.基于此设计原理定义了 PIER 的查询语义——加宽可达快照(dilated-reachable snapshot),可达快照(reachable snapshot)是在查询从客户节点发出时可达的节点发布的数据的集合.考虑到时钟和查询处理全局同步的难度,Ryan 采用了可达节点发布数据的本地快照的(略微时间加宽的)并集,其中每个本地快照是查询消息到达该节点时刻的快照.Piazza[8]通过扩展必然的回答(certain answers)的概念定义查询回答语义:必然的回答是在每个可能的一致的数据实例中都成立的回答.给定一个 PDMS N、存储的关系 D 的一个实例以及一个查询 Q,Piazza 的查询语义是找到 Q 的全部必然的回答.

2.1.2.4 查询处理

当前,PDMS 的查询处理策略有数据传递、查询传递、代理传递、基于 DHT 的查询处理、突变查询计划[31].数据传递是将计算结果所需的数据移动到查询的发起者,在查询发起处进行所有操作,其缺点是数据传递的开销导致更长的响应时间,查询完成的时间随着数据量的增加呈指数增长.因此,大多数 PDMS 采用查询传递,将查询向数据移动,只有满足查询的部分才传递给查询的发起者作进一步处理,削减了网络中移动的数据量.代理传递是查询传递的发展,属于代码传递[15],代理中携带查询和处理代码到远程站点执行,只返回代理产生的回答;在数据提供者处执行对数据的处理使 peer 共享计算能力,同时携带代码和数据可以有效地进行任何功能.PeerDB[15]率先使用代理传递,用关系匹配代理找到有希望的 peer,然后由数据检索代理翻译和提交 SQL 查询到这些 peer,将结果发送回产生查询的主代理.Edutella[43]提出可扩展基于模式的查询处理方法——推出携带代码的查询求值计划,它类似于代理传递,查询求值计划和用户定义的代码被从客户推到执行它们的 super-peer,由 supper-peer 提供查询优化器产生好的查询计划,在数据源附近进行标准查询处理和用户定义的操作.为了利用 DHT 提供的高可扩展性以及查询效率保证,一些查询引擎建立在 DHT 上:PIER[7]将对称哈希连接、Fetch-Matches 连接算法扩展到 DHT.传统的分布式查询要求子计划分布、主动和同步地通信,要求分布的状态,Papadimos 等人[44,45]提出了突变查询计划(mutant query plans,简称 MQP),查询计划使用 XML 表示,也可包含逐字的 XML 数据、引用资源定位(uniform resource locator,简称 URL)或抽象的资源名(uniform resource name,简称 URN).各服务器用局部的、可能不完备的知识尽可能多地部分求值查询计划,将部分结果合并成新的突变的查询,传递给能够继续处理的其他服务器.MQP 放开传统分布式查询处理模型的两个约束:集中的优化和同步计划.因此,MQP 允许分布式的查询求值同时保持本地的状态,在任何一点(除了 MQP 被传递的短暂时期),在面对 MQP 重载时,服务器(peer)可自由选择延迟求值、按相似性分组 MQP 或将工作负载转到另一服务器,而不占用网络其他地方的资源.另外,MQP 可以处理不完全的元数据、进行分布式优化、尊重执行站点的自治和本地策略,即使在求值的过程中也能随服务器和网络条件的改变而加以调整.然而,MQP 不能支持用户定义的操作符,它不使用流水(pipeline),限制了在分布式查询处理中的普遍应用[43].

要使 P2P 网络得到广泛的应用需支持复杂查询,包括范围查询、连接查询、聚集查询、递归查询.

- 范围查询

DHT 随机的散列函数适合负载平衡,但不能支持范围查询[46].为了支持近似范围查询,文献[47]用位置保持哈希散列范围,文献[48]提出了改进文献[47]来支持确切范围查询的方法,但它对首次提出的范围查询不提供任何性能保证.它们都要事先分区:分区太大,就容易过载;分区太小,就需要太多跳步[22];它们通常不支持或不能高效支持简单的 key 搜索操作,对相关范围会产生很差的碎片,造成存储负载平衡差或访问不高效.RDFPeers[9]用位置保持哈希函数为 RDF 的 subject,predicate 和 object 产生 key,用 MAAN(multi-attribute addressable network)

支持范围查询.针对简单位置保持哈希函数对分布不均匀的属性不一定能产生均匀哈希值的问题,文献[9]提出了在输入属性的分布函数连续且分布可预知情况下的均匀位置保持哈希函数.然而,RDFPeers 仍需要依赖于虚拟节点改善负载平衡,而不能有效处理查询负载倾斜问题.

Aspnes 等人提出 Skip Graphs[49],它是建立在 skip 列表上的随机结构,能支持范围查找.但即使当索引完全一致时,它也只提供概率保证.Skip Graphs[50]可选择支持范围查询或负载平衡,但是不能同时支持两者[22]. Ganesan 等人[50]提出用两个 Skip Graphs 来提供负载平衡,但是文献[50]没有给出查询倾斜的实验结果.

PePeR(peer-to-peer range)[51]系统的性能依赖于插入操作的某种启发式算法,不提供任何性能保证,即使在索引完全一致后,搜索性能仍可能是线性的[32].

传统数据库研究表明,tries 是支持范围查询的最实际数据结构.PHT(prefix Hash tree)[52]在任意结构的 overlay 上添加树,将范围查询翻译成若干 DHT 搜索,其缺点是不够高效.GridVine[12]提出了使用 trie 结构的 P2P 重叠网络支持范围查询的方法,用保序哈希产生数据项的 key,提出顺序的 min-max traversal 和并行的 shower 范围查询算法.即使数据倾斜,GridVine 也能保证对数级搜索性能.GridVine 将此结构用于支持 RDF 三元组的范围、前缀、后缀和子串查询,但是 GridVine 也没有考虑查询倾斜的负载平衡.总的来说,基于前缀匹配/tries 的途径(文献[53–55])可支持范围查询,但不能用于任意数值属性(如浮点数)的范围查找[32].

Crainiceanu 等人将 B+树思想扩展到 P2P 环境,提出 P-Tree[6]和 P-Ring[32,33],能有效地支持范围查询和保持大致的存储平衡,但也没有考虑查询倾斜的情况.另外,当 peer 后继指针不一致或在查询过程中 peer 负责的范围改变时,P-Tree 和 P-Ring 会遗漏范围查询的结果.因此,Linga 等人[56]提出了在范围索引中可证明地确保查询的正确性及系统、项的可用性的技术.

除了 GridVine 以外,上述各种方法都假设系统中所有的节点使用共同的模式,其查询处理策略没有考虑 P2P 网络的语义异构问题.

- 连接查询

PIER[7]将对称哈希连接算法扩展到 DHT 上,交替地在每个输入关系上建造和探测哈希表,处理相等连接.在参加连接的一个表已用连接属性散列存储时,只需扫描另一个表,此时称为 Fetch Matches 算法.然而,PIER 不能处理语义异构 peer 上的连接查询.

文献[57]提出分配固定数量的 peer 充当范围哨(range guard),对连接属性的域进行分区,一个分区有一个范围哨,查询只发送到范围哨上,但没有指出选择范围哨的高效算法[22].

- 聚集查询

有时,用户需要了解系统特性的聚集或作为一个整体了解数据集合[16].例如,P2P 网络上的应用程序管理员为获得使用趋势的信息,需要用网络中主机上的数据计算聚集函数(如主机的平均生命周期),以便据此调整应用程序的行为[5].这样就需要聚集查询.

Stanford 大学 Peers 组研究无结构 P2P 网络上的聚集查询,取得了一些初步的成果.文献[58]解决了谓词上的 COUNT 查询(如计算属于 stanford.edu 域的节点的数量),它定义和研究“节点聚集”问题以及 P2P 网络的可计算性,给出能用来计算任何基本的聚集函数的一般策略,可用于为特定任务平衡准确性和效率[5].文献[59]注意到聚集函数的许多 best-effort 算法的查询语义不够清楚,很难对返回的结果相关联的保证(guarantee)进行推理,提出了称为“单一站点有效性”的正确性条件,以及保持动态网络中有效性的一类算法,在现实和人造的网络拓扑上进行了实际验证.

PIER[27]提出了层次聚集的方法.PIER 不显式分组节点,而是按照查询广播中的方法将节点组织成树,各节点计算它的本地聚集,并将结果发送到查询中指定的根,消除了接收所有数据进行聚集的节点的瓶颈问题.设计者也指出这种方法的局限性:为了使各节点恰好只发送一个部分聚集,各节点必须知道何时已从每个孩子接收了数据,对于可分配的聚集和算术聚集,算法工作良好,但并不能改善整体聚集.

- 递归查询

递归查询是允许使用自身来定义其结果的查询,对查询本身具有递归结构的关系非常有用,可用于充当查

询分布式网络图的接口,用于理解和控制 P2P 和 overlay 的结构特性[60].PIER[60]用距离向量协议(类似路由表计算时用的)计算递归查询,发起递归查询的节点发送一个事实的集合到它的邻居,邻居依次基于自身的本地信息更新这个集合,然后将结果传得更远.文献[60]指出实现递归查询有两种设置:(1) 嵌入的网络查询:网络中的每个节点都嵌入查询功能,可用于计算从一个节点开始可达的节点的集合、两个节点间的最短路径、两个节点的路径数;(2) 外部的网络查询:查询由类似于 PIER 的分离的基于 DHT 的查询基础设施执行,这种设置可以用来设计 P2P 网络上的分布式爬虫[60].PIER 已用 Datalog 描述递归查询产生爬虫爬行 Gnutella 网络.

2.1.2.5 查询优化

由于大 P2P 网络的特性相当易变,需要运行时再优化,提交的查询不太可能在整个查询处理过程中都保持原来的性质,因此,传统的静态查询优化和执行技术在这种环境中无效.文献[61]提出了自适应的查询优化方法 Eddies,对流入和流出算子的数据流率进行观察,在此基础上决定如何在算子间路由元组.然而,在 P2P 环境中,每个节点的 eddy 只能看到路由到或经过该节点的数据,本地 eddy 实例很难作出全局决策,eddy 如何在 P2P 环境中合作仍是悬而未决的问题[27].

PIER[27]由查询书写者进行优化,这是非常初级的优化方法,这种方法要求使用 PIER 的应用程序设计者是 Internet 系统专家,比起 SQL 应更熟悉 Click 路由器等数据流程.然而实际用户更喜欢用 SQL[27],因此,需要更精密的优化器.文献[43]讨论了 Edutella 上的 3 种优化策略:(1) 当一个逻辑资源(logical resource,简称 LR)对应于多个物理资源(physical resource,简称 PR)或资源方向(resource direction,简称 RD)时,先对所有的物理资源求“并”,再对产生的结果应用逻辑资源上的操作符;(2) 将连接操作推入“并”中来增加分布程度;(3) 尽量在一个主机上收集关于一个逻辑资源的数据.文献[44]为 MQP 提出合并、吸收、延期 3 种查询优化方法,尽量在本地计算更多的算子,减少网络中传输的中间查询结果的大小.目前,他们正在研究如何在查询优化器的一次运行中找出所有需要延期的候选算子[44].Piazza[8]认为冗余是造成查询的执行时间和响应时间较长的原因,指出查询展开和重写的重复应用会导致包含冗余子表达式的大的再行式化(reformulation)结果,提出了查询优化策略:修剪 reformulation goals、最小化 reformulation、嵌套 XML 查询的包含和最小化,预计算语义路径,将 reformulation 过程看作在再形式化空间中搜索,要求最终会被剪除的 reformulation 在展开其任何子树前剪除.

2.2 连续查询

连续查询(continuous queries,简称 CQ)是能够执行较长的一段时间、监视底层的数据流语义来触发用户定义的行为的查询,它将被动的网络结构转换成一个主动的网络结构,在大量数据被频繁地远程更新的分布式的网络环境中非常有用[14].连续查询主要是数据流(data stream)领域研究的问题.最初的 CQ 系统使用客户/服务器模型.多个连续查询系统不共享计算,各系统独立运作,只关心有效地处理自己内部的连续查询,因此造成了大量的重复工作,负载不均衡,资源得不到充分的利用[14].这样的系统难于扩展和维护,并且容易形成单点故障[62].由于 P2P 模型具有高可扩展性和健壮性等优点,研究人员开始采用 P2P 模型来解决连续查询领域中的问题,连续查询领域已经成为倍受关注的 P2P 应用环境.

PeerCQ[62]认为拥有标识越多的 peer,则映射到该 peer 更多的 CQ 的概率越大.它通过 CQ 到标识符、peer 到标识符和 CQ 到 peer 三次映射基于 DHT 进行服务划分,为能力强的 peer 映射更多的标识,从而实现整个系统的负载平衡.文献[62]认为,将相似的查询分布到连续的标识区域,则映射到同一 peer 的概率更大.通过两阶段选择 CQ 执行节点,在相似的 CQ 间共享计算,减少重复操作,提高系统的利用率.但是,PeerCQ 只允许一个 peer 的 $2r$ 个邻居中同时只有一个加入或者离开操作,限制了节点的自治性.由于节点的故障和节点的离开在行为上非常类似,要求邻居中一次只能有一个故障节点在 P2P 自治环境中是很难保证的.

P2P-DIET[13]是基于 super-peer 网络的选择性信息分发(publish/subscribe)系统,主要用于数字图书馆等领域,处理的数据以文件为单位,提供广告(advertisement)、通知(notification)和预约(rendezvous)的功能.当文件出现后,由文件资源的提供者自动地向 super-peer 提供信息变化的情况.CQ 的执行完全由服务访问点(super-peer)进行,客户节点只提供资源而不参加 CQ 的处理,查询提出后,使用最小生成树将 CQ 传播到所有的 super-peer.这样,消息数多,网络带宽消耗也大.另外,该系统不提供负载平衡功能.

CQ-Buddy[14]是基于无结构 P2P 网络的连续查询系统,提出“普遍深入的连续查询(pervasive continuous queries)”的概念,解决 peer 的频繁断开连接打扰连续查询执行的问题.它提出了谓词相似性、投影相似性和数据源相似性 3 种相似性标准.对于新提出的连续查询,接收该查询的节点首先确定该 CQ 与自身缓冲池中正在运行的 CQ 的相似性:如果与现存正在运行的 CQ 相似,则将新到查询添加到现存查询之上;如果与所有正在运行的 CQ 都不相似,则 CQ-Buddy 有两种可选择的处理策略:(1) SELF-HELP:如同单个连续查询系统一样,peer 发起一个新的处理任务来处理它接收到的新查询;(2) BUDDY-HELP:peer 向它的伙伴 peer(buddy peer)寻求帮助来处理查询,然后伙伴 peer 代表这个 peer 执行查询并向该 peer 提供连续查询的结果.但是,它建立在无结构 P2P 重叠网络之上,因此不可避免地具有无结构 P2P 网络的缺点:在寻求 BUDDY-HELP 的过程中,不能保证在有限步内找到实际存在的相似查询.因此,会造成重复计算,并且影响系统扩展.

MIT 的 Medusa[63]使用 Chord 将存储分布式数据流及其互连的目录集合分布到各节点,为流处理引擎 Aurora 提供网络基础设施.

3 总结及展望

现有研究项目分别从 P2P 网络的语义重叠网络和重叠网络两个方面改善现有 P2P 网络的性能和服务质量,两者之间缺乏衔接.它们或者解决语义互操作性而只提供有限的可扩展性(Hyperion,Edutella,PeerDB,Piazza 等),或者解决了可扩展性问题但不处理语义互操作性(PIER,RDFPeers 等).然而,为了在大规模语义异构网络中高效、健壮地共享信息,充分发掘分布在网络边缘的信息源的价值,满足 P2P 研究领域中对语义支持的需求以及数据库领域中提高可扩展性的需求,未来的 P2P 系统应该既提供语义异构支持又高可扩展,并且应至少提供关系完备的查询处理能力.当前,P2P 数据管理的研究成果不能直接结合产生这样的系统,需要研究将两者融合的技术.为此,今后需要在以下几个方面继续加以研究:

(1) 模式映射管理方法

当前主要研究如何在 peer 间建立映射提供 peer 间的语义互操作的方法.将来需要全新的方法依查询解释语义冲突,需要能够表达更复杂的语义关系同时能够扩展到大量数据源的映射方法;并且,随着网络进化,会产生大量的语义映射,需要研究管理这些映射的技术,包括:找出应该建立而尚未建立映射的节点集、发现 P2P 局部交互产生的矛盾的映射、解决矛盾映射的方法;寻求处理不同质量的映射的途径,逐步改善映射,从而改进节点从其他节点获得数据的能力.

(2) 高可扩展性语义索引构造和维护方法

当前,只有基于模式的索引考虑了模式异构问题,提供语义支持(但当前此类索引的可扩展性有限),其余各类索引均假定系统中所有节点都采用相同的模式,仅提供数据级索引.要将语义异构支持和高可扩展性相融合,构造高可扩展的语义索引是关键.为了给查询优化算法提供信息,需要研究在索引中动态维护统计信息、网络拓扑参数等额外信息的方法.

(3) 查询处理能力和查询优化

在即兴查询方面,信息检索领域中全面的关键字搜索仍是悬而未决的问题;分级关键字搜索需要对文档整个集合的全局统计(如文档频率),目前还没有收集和维护统计信息的健壮、高效、分布式方法.另外,在数据库查询领域,现有的范围查询(除 GridVine 外)、连接查询、聚集查询算法均假定网络中使用相同的模式,利用高可扩展的实例级索引(绝大多数利用重叠网络)进行高效的分布式查询处理.然而,P2P 网络语义异构是不可避免的.因此,今后需要研究利用高可扩展语义索引提供上述查询能力的途径;研究在上述运算中 Null 值的处理策略,以及包含否定的、跨数据源、多操作符的查询处理方法.由于大的 P2P 网络的特性相当易变,需要运行时再优化,提交的查询不太可能在整个查询处理过程中都保持原来的性质,每个节点的查询执行引擎和优化器只能看到路由到或经过该节点的数据,很难作出全局决策,因此,P2P 动态环境下的自适应查询优化极具挑战性,仍是悬而未决的问题.

在连续查询方面,现有的 P2P 连续查询系统通过相似查询聚簇共享计算.然而在聚簇过程中,有的提供了负

载平衡功能但限制了网络的动态性(PeerCQ);有的不提供负载平衡功能且广播消耗大量带宽(P2P-DIET);有的基于无结构重叠网络扩展性差(CQ-Buddy).因此,将来需要研究新的 P2P 连续查询方法,有效支持 P2P 网络动态性,保证在有限跳步内找到相似查询,减少带宽消耗,提供有效的负载平衡.

References:

[1] Gribble SD, Halevy AY, Ives ZG, Rodrig M, Suciu D. What can databases do for peer-to-peer? In: Mecca G, Simeon J, eds. Proc. (informal) of the 4th Int'l Workshop on the Web and Databases (WebDB). Santa Barbara, 2001. 31−36.

[2] Ling B, Lu ZG, Ng WS, Qian WN, Zhou AY. PeerIS: A peer-to-peer based information retrieval system. Journal of Software, 2004,15(9):1375−1384 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/15/1375.htm

[3] Aberer K, Hauswirth M. An overview on peer-to-peer information systems. In: Litwin W, Lévy G, eds. Proc. in Informatics (14). Waterloo: Carleton Scientific, 2002. 171−188.

[4] Batini C. A survey of data quality issues in cooperative information systems. In: Tutorial of the 23rd Int'l Conf. on Conceptual Modeling (ER 2004). Shanghai: Fudan University, 2004.

[5] Bawa M, Cooper BF, Crespo A, Daswani N, Ganesan P, Garcia-Molina H, Kamvar S, Marti S, Schlosser M, Sun Q, Vinograd P, Yang B. Peer-to-Peer research at Stanford. ACM SIGMOD Record, 2003,32(3):23−28.

[6] Crainiceanu A, Linga P, Gehrke J, Shanmugasundaram J. Querying peer-to-peer networks using p-trees. In: Amer-Yahia S, Gravano L, eds. Proc. of the 7th Int'l Workshop on Web and Databases. New York: ACM Press, 2004. 25−30.

[7] Heubsch R, Hellerstein JM, Lanham N, Loo BT, Shenker S, Stocia I. Querying the Internet with PIER. In: Freytag JC, Lockemann PC, Abiteboul S, Carey MJ, Selinger PG, Heuer A, eds. Proc. of the 29th Int'l Conf. on Very Large Data Bases. San Fransisco: Morgan Kaufmann Publishers, 2003. 321−332.

[8] Tatarinov I, Halevy A. Efficient query reformulation in peer data management systems. In: Weikum G, Konig AC, DeBloch S, eds. Proc. of the ACM SIGMOD Int'l Conf. on the Management of Data. Paris: ACM, 2004. 539−550.

[9] Cai M, Frank M. RDFPeers: A scalable distributed RDF repository based on a structured peer-to-peer network. In: Tolles-Efinger L, ed. Proc. of the 13th Int'l World Wide Web Conf. New York: Sheridan Printing, 2004. 650−657.

[10] Kementsietsidis A, Arenas M. Data sharing through query translation in autonomous sources. In: Nascimento MA, Ozsu MT, Kossmann D, Miller RJ, Blakeley JA, Schiefer KB, eds. Proc. of the 30th Int'l Conf. on Very Large Data Bases. San Fransisco: Morgan Kaufmann Publishers, 2004. 468−479.

[11] Nejdl W, Siberski W, Sintek M. Design issues and challenges for RDF and schema-based peer-to-peer systems. ACM SIGMOD Record, 2003,32(3):41−46.

[12] Aberer K, Cudré-Mauroux P, Hauswirth M, van Pelt T. GridVine: Building Internet-scale semantic overlay networks. Lecture Notes in Computer Science, 2004, 3298:107−121.

[13] Chirita PA, Idreos S, Koubarakis M, Nejdl W. Publish/Subscribe for RDF-based P2P networks. Lecture Notes in Computer Science, 2004,3053:182−197.

[14] Ng WS, Ooi BC, Shu YF, Tan KL, Tok WH. Efficient distributed continuous query processing using peers. Technical Report, NUS-CS01-03, Kent Ridge: National University of Singapore, 2003. 1−13.

[15] Ng WS, Ooi BC, Tan KL, Zhou AY. PeerDB: A P2P-based system for distributed data sharing. In: Dayal U, ed. Proc. of the 19th Int'l Conf. on Data Engineering (ICDE). Bangalore: IEEE Computer Society Press, 2003. 633−644.

[16] Daswani N, Garcia-Molina H, Yang B. Open problems in data-sharing peer-to-peer systems. Lecture Notes in Computer Science, 2003,2572:1−15.

[17] Harren M, Hellerstein JM, Huebsch R, Loo BT, Shenker S, Stoica I. Complex queries in DHT-based peer-to-peer networks. Lecture Notes in Computer Science, 2002,2429:242−259.

[18] Joseph S, Hoshiai T. Decentralized meta-data strategies: Effective peer-to-peer search. IEICE Trans. on Communications, 2003, E85-B(6):1740−1753.

[19] Weikum G. Towards a statistically semantic Web. Lecture Notes in Computer Science, 2004,3288:3−17.

[20] Galanis L, Wang Y, Jeffery SR, DeWitt DJ. Processing queries in a large peer-to-peer system. Lecture Notes in Computer Science, 2003,2681:273−288.

[21] Bawa M, Manku GS, Raghavan P. SETS: Search enhanced by topic segmentation. In: Callan J, Cormack G, Clarke C, Hawking D, Smeaton A, eds. Proc. of the 26th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2003. 306−313.

[22] Risson J, Moors T. Survey of research towards robust peer-to-peer networks: Search methods. Technical Report, UNSWEE-P2P-1-1, Sydney: University of New South Wales, 2004. 1−36.

[23] Shen HT, Shu YF, Yu B. Efficient semantic-based content search in P2P network. IEEE Trans. on Knowledge and Data Engineering, 2004,17(7):813−826.

[24] Balke WT. Supporting information retrieval in peer-to-peer systems. Lecture Notes in Computer Science, 2005,3485:337−352.

[25] Ganesan P, Sun Q, Garcia-Molina H. Adlib: A self-tuning index for dynamic peer-to-peer systems. In: Kawada S, ed. Proc. of the 21st Int'l Conf. on Data Engineering (ICDE). Tokyo: IEEE Computer Society, 2005. 256−257.

[26] Aberer K. Guest editor's introduction. ACM SIGMOD Record, 2003,32(3):21−22.

[27] Huebsch R, Chun B, Hellerstein J, Loo BT, Maniatis P, Roscoe T, Shenker S, Stoica I, Yumerefendi AR. The architecture of PIER: An Internet-scale query processor. In: Stonebraker M, Weikum G, DeWitt D, eds. Proc. of the 2005 Conf. on Innovative Data Systems Research. Asilomar: VLDB, 2005. 28−43.

[28] Aberer K, Cudre-Maroux P, Hauswirth M, Van Pelt T. Start making sense: The chatty Web approach for global semantic agreements. Journal of Web Semantics, 2004,1(1):72−86.

[29] Ooi BC, Shu YF, Tan KL. Relational data sharing in peer-based data management systems. ACM SIGMOD Record, 2003,32(3): 59−64.

[30] Hellerstein JM. Toward network data independence. ACM SIGMOD Record, 2003,32(3):34−40.

[31] Zhou AY. The understanding and consideration on several frontiers of database. In: Proc. of the 2004 Symp. on Database Development Strategy (in Chinese). Shenyang, 2004 (in Chinese with English abstract).

[32] Crainiceanu A, Linga P, Machanavjjhala A, Gehrke J, Shanmugasundaram J. P-Ring: An index structure for peer-to-peer systems. Technical Report, TR2004-1946, New York: Cornell University, 2004. 1−19.

[33] Crainiceanu A, Linga P, Machanavajjhala A, Gehrke J, Shanmugasundaram J. An indexing framework for peer-to-peer systems. In: Weikum G, Konig AC, DeBloch S, eds. Proc. of the ACM SIGMOD Int'l Conf. on the Management of Data (Demo). Paris: ACM, 2004. 939−940.

[34] Yang B, Garcia-Molina H. Improving search in peer-to-peer networks. In: Rodrigues LET, Raynal M, Chen WSE, eds. Proc. of the 22nd Int'l Conf. on Distributed Computing Systems. Washington: IEEE Computer Society, 2002. 5−14.

[35] Lü Q, Ratnasamy S, Shenker S. Can heterogeneity make Gnutella scalable? Lecture Notes in Computer Science, 2002,2429:94−103.

[36] Lü Q, Cao P, Cohen E, Li K, Shenker S. Search and replication in unstructured peer-to-peer networks. In: Ebcioglu K, Pingali K, Nicolau A, eds. Proc. of the 16th Int'l Conf. on Supercomputing. New York: ACM Press, 2002. 84−95.

[37] Banaei-Kashani F, Shahabi C. Criticality-Based analysis and design of unstructured peer-to-peer networks as "complex systems". In: Yokokawa M, ed. Proc. of the 3rd IEEE/ACM Int'l Symp. on Cluster Computing and the Grid. Tokyo: IEEE Computer Society, 2003. 351−358.

[38] Doan A, Noy D, Halevy A. Introduction to the special issue on semantic integration. ACM SIGMOD Record, 2004,33(4):11−13.

[39] Bernstein P, Giunchigalia F, Kementsietsidis A, Mylopoulos J, Serafini L, Zaihrayeu I. Data management for peer-to-peer computing: A vision. In: Fernandez MF, Papakonstantinous Y, eds. Proc. of the 5th Int'l Workshop on the Web and Databases (WebDB 2002). Madison: Wisconsin, 2002. 89−94.

[40] Kantere V, Kiringa I, Mylopoulos J, Kementisietsidis A, Arenas M. Coordinating peer databases using ECA rules. Lecture Notes in Computer Science, 2004, 2944:108−122.

[41] Calvanese D, De Giacomo G, Lenzerini M, Rosati R. Logical foundations of peer-to-peer data integration. In: Deutsch A, ed. Proc. of the 23rd ACM SIGMOD-SIGACT-SIGART Symp. on Principles of Database Systems (PODS 2004). Paris: ACM, 2004. 241−251.

[42] Nejdl W, Siberski W. Schema-Based peer-to-peer systems. Lecture Notes in Computer Science, 2005,3485:323−336.

[43] Brunkhorst I, Dhraief H, Kemper A, Nejdl W, Wiesner C. Distributed queries and query optimization in schema-based P2P-systmes. Lecture Notes in Computer Science, 2004,2944:184−199.

[44] Papadimos V, Maier D. Distributed queries without distributed state. In: Fernandex MF, Papkonstantinou Y, eds. Proc. of the 5th Int'l Workshop on the Web and Databases (WebDB 2002). Madison: Wisconsin, 2002. 95−100.

[45] Papadimos V, Maier D, Tufte K. Distributed query processing and catalogs for peer-to-peer systems. In: Proc. (online) of the 1st Biennial Conf. on Innovative Data Systems Research (CIDR 2003). Asilornar: Wisconsin, 2003.

[46] Agrawal BAM, Seshan S. Mercury: Supporting scalable multi-attribute range queries. Computer Communication Review, 2004, 34(4):353−366.

[47] Gupta A, Agrawal D, Abbadi AE. Approximate range selection queries in peer-to-peer systems. In: Proc. (online) of the 1st Biennial Conf. on Innovative Data Systems Research (CIDR 2003). Asilornar: Wisconsin, 2003.

[48] Sahin OD, Gupta A, Agrawal D, Abbadi AE. A peer-to-peer framework for caching range queries. In: Rundensteiner E, ed. Proc. of the 20th Int'l Conf. on Data Engineering. Boston: IEEE Computer Society, 2004. 165−176.

[49] Aspnes J, Shah G. Skip graphs. In: Proc. of the 14th Annual ACM-SIAM Symp. on Discrete Algorithms. Philadelphia: Society for Industrial and Applied Mathematics, 2003. 384−393.

[50] Ganesan P, Bawa M, Carcia-Molina H. Online balancing of range-partitioned data with applications to peer-to-peer systems. In: Nascimento MA, Ozsu MT, Kossmann D, Miller RJ, Blakeley JA, Schiefer KB, eds. Proc. of the 30th Int'l Conf. on Very Large Data Bases. San Fransisco: Morgan Kaufmann Publishers, 2004. 444−455.

[51] Daskos A, Ghandeharizadeh S, An X. PePeR: A distributed range addressing space for peer-to-peer systems. Lecture Notes in Computer Science, 2004,2944:165−176.

[52] Ramabhadran S, Ratnasamy S, Hellerstein JM, Shenker S. Brief announcement: Prefix Hash tree. In: Chaudhuri S, Kutten S, eds. Proc. of the 23rd Annual ACM Symp. on Principles of Distributed Computing. St.John's: ACM, 2004. 368.

[53] Aberer K. P-Grid: A self-organizing access structure for P2P information systems. Lecture Notes in Computer Science, 2001,2172:179−194.

[54] Freedman MJ, Vingralek R. Efficient peer-to-peer lookup based on a distributed trie. Lecture Notes in Computer Science, 2002,2429:66−75.

[55] Rowstron A, Druschel P. Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems. Lecture Notes in Computer Science, 2001,2218:329−350.

[56] Linga P, Crainiceanu A, Gehrke J, Shanmugasundaram J. Guaranteeing correctness and availability in P2P range indices. In: Ozcan F, ed. Proc. of the ACM SIGMOD Int'l Conf. on Management of data. Baltimore: ACM, 2005. 323−334.

[57] Triantafillow P, Pitoura T. Towards a unifying framework for complex query processing over structured peer-to-peer data networks. Lecture Notes in Computer Science, 2004,2944:169−183.

[58] Bawa M, Garcia-Molina H, Gionis A, Motwani R. Estimating aggregates on a peer-to-peer networks. Technical Report, 2003-24, Stanford: Stanford University, 2003. 1−13.

[59] Bawa M, Gionis A, Garcia-Molina H, Motwani R. The price of validity in dynamic networks. In: Weikum G, Konig AC, DeBloch S, eds. Proc. of the ACM SIGMOD Int'l Conf. on the Management of Data. Paris: ACM, 2004. 515−526.

[60] Loo BT, Huebsch R, Hellerstein JM, Roscoe T, Stoica I. Analyzing P2P overlays with recursive queries. Technical Report, UCB//CSD-04-1301, Berkeley: Computer Science Division, UC Berkeley, 2004. 1−5.

[61] Avnur R, Hellerstein JM. Eddies: Continuously adaptive query processing. In: Chen WD, Naughton JF, Bernstein PA, eds. Proc. of the 2000 ACM SIGMOD Int'l Conf. on Management of Data. Dallas: ACM, 2000. 261−272.

[62] Gedik B, Liu L. PeerCQ: A decentralized and self-configuring peer-to-peer information monitoring system. In: Titsworth FM, ed. Proc. of the 23rd IEEE Int'l Conf. on Distributed Computer Systems. Providence: IEEE Computer Society, 2003. 490−499.

[63] Zdonik S, Stonebraker M, Chemiack M, Centintemel U, Balazinska M, Balakrishna H. The aurora and medusa project. IEEE Data Engineering Bulletin, 2003,26(1):3−10.

附中文参考文献:

[2] 凌波,陆志国,黄维维,钱卫宁,周敖英.PeerIS:基于 Peer-to-Peer 的信息检索系统.软件学报,2004,15(9):1375−1384. http://www.jos.org.cn/1000-9825/15/1375.htm

[31] 周敖英.若干数据库前沿技术的理解与思考.见:2004 年数据库发展战略研讨会.沈阳,2004.

余敏(1980－),女,江西波阳人,博士生,主要研究领域为数据库理论与技术,P2P 数据管理.

李战怀(1961－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据库理论与技术.

张龙波(1968－),男,博士生,副教授,主要研究领域为数据流管理,查询处理.