图数据关键词查询研究进展

杨书新[1],徐丽萍[1],夏小云[2],徐慧琴[1]
(1.江西理工大学信息工程学院,江西赣州 341000;2.华南理工大学计算机科学与工程学院,广东广州 510006)

摘 要: 图数据关键词查询适用于结构化、半结构化、非结构化数据,使得普通用户在不需要熟悉任何查询语言和底层数据模式情况下能检索数据.目前,图数据关键词查询技术已成为数据库和信息检索领域的研究热点.该文对现有的图数据关键词查询方法进行了综述.首先,介绍了图数据关键词查询的基本概念.然后,对关键技术进行总结和对比分析,包括搜索算法、排序、查询意图和查询评价.最后,对当前工作存在的主要问题及未来研究方向进行讨论.

关键词: 关键词查询; 图数据; 排序; 查询评价
中图分类号: TP311 **文献标识码:** A **文章编号:** 0372-2112 (2014)11-2260-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.11.020

Advances in Keyword Search over Graph Data

YANG Shu-xin[1], XU Li-ping[1], XIA Xiao-yun[2], XU Hui-qin[1]
(1. *School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou, Jiangxi* 341000, *China*;
2. *School of Computer Science and Engineering, South China University of Technology, Guangzhou, Guangdong* 510006, *China*)

Abstract: The technique of keyword search over graph data can be applied to structured data, semi-structured data, and unstructured data. It enables ordinary users to retrieve related information from data without the needs of being familiar with any query languages and underlying data model. Keyword search over graph data is one of research hotspots of the database and information retrieval. The paper proposed a detail review of the current work of keyword search over graph data. Firstly, basic concepts were introduced. Then, some key techniques in the query processing were summarized and analyzed, including search methods, ranking strategies, query intent, and query evaluation. Finally, the existing problems in the current research work and future research issues were also discussed.

Key words: keyword search; graph data; ranking; query evaluation

1 引言

图数据具有通用性,从图中查询信息是适合社会计算时代的搜索方式,近几年受到工业界和学术界的广泛关注[1].图数据关键词查询不仅检索出单个信息实体,而且还找出实体之间的关系.

图数据关键词查询指的是给定图数据 G 和一组关键词 $K=\{k_1,\cdots,k_l\}$,用户在不需要熟悉任何查询语言(如 SPARQL,SQL)和底层数据模式情况下,能方便检索相关信息,适用于结构化、半结构化、非结构化数据[2].其处理过程是根据输入的关键词和词汇倒排索引快速定位包含关键词的节点,运用搜索算法寻找包含输入关键词的结果,并按照排序策略进行排序,返回给用户.

尽管图数据关键词查询技术在结构化数据、半结构化数据、非结构化数据不同领域涌现大量的研究工作,但对该技术的研究尚缺乏全面、系统的总结,因此本文对研究现状进行分析,为未来进一步的研究提供参考.目前国内外该领域与本文相关的研究综述有文献[3,4],文献[3]针对结构化数据-关系数据库,与我们的侧重点不同,文献[4]主要对 2010 年以前的一些搜索算法做了简要的叙述,本文与其不同的是系统地从结果搜索、排序、查询意图等多方面对图数据关键词查询进行分类归纳,并展望未来的研究方向.

2 基本概念

定义 1[5] 将图数据描述为六元组:$G=(V,E,T,L,\omega_v,\omega_e)$

(1) V 为节点集合,$V=\{v_1,v_2,\cdots,v_n\}$.

收稿日期:2014-01-15;修回日期:2014-06-22;责任编辑:孙瑶
基金项目:国家自然科学基金(No.41362015);江西省科技厅青年科学基金(No.20122BAB211035);江西省教育厅科学技术研究项目(No.GJJ14431,No.GJJ13415,No.GJJ13411);江西省教育厅重点项目(No.赣教技字[12770]号)

(2)E 为有向或无向边集合,$E \subseteq V \times V$,$E = \{e_1, e_2, \cdots, e_m\}$.

(3)T 为 G 中出现的词汇集合.

(4)L 为节点标签映射函数,$L(v) \subseteq T$.

(5)ω_v 是节点权重函数,为节点分配一个非负的权重值.

(6)ω_e 是边权重函数,为边分配一个非负的权重值.

对于图 G 中的一个节点 v,如果它包含的文本内容与关键词 k_i 匹配,则称 v 是关于 k_i 的关键词节点.

top-k 查询指的是寻找 k 个相关度高的且包含所有关键词的结果,结果中的节点满足某种结构约束关系.部分研究将结果的结构形式看作是关键词节点之间的最小树,这时,top-k 查询类似于组斯坦纳树问题. Li[2]指出在大图和复杂图中很难找出所有的斯坦纳树,因为这个是 NP 难问题,且结果树呈现的信息不丰富,难以满足用户检索信息的需求.于是一些学者提出使用子图来定义查询结果,要求覆盖 K 的每个关键词,如定义 3~定义 6,其中定义 3~定义 5 为有界子图.对于返回的查询结果定义,我们将其分为最小树和子图两大类.

(1)最小树

定义 2(最小树) 给定图 G 和关键词 $K = \{k_1, \cdots, k_i, \cdots, k_l\}$,假设 $g = \{g_1, g_2, \cdots, g_l\}$,$g_i = \{v_1, v_2, \cdots, v_p\}$,$g_i$ 表示节点文本内容与关键词 k_i 匹配的节点集,$1 \leqslant i \leqslant l$,$l$ 为输入的关键词个数,p 表示 k_i 对应的关键词节点个数.如果图 G 存在 $T = (V', E')$,T 是至少包含 g_i 中一个节点的树,并且 T 在图 G 中所有关于 V' 的树度量值最小,则 T 为最小树.度量值可通过计算 T 的边权重之和、T 的节点权重与边权重的加权和等方式获得.T 的叶子节点只能是关键词节点.

BANKS[6]、DPBF[7]、文献[8]的结果遵循定义 2,与此不同的是,BANKS-Ⅱ[9]、BLINKS[10]的结果虽遵循定义 2,但加了一个约束,就是查询结果集的结果树不允许根节点相同,即在搜索过程中如果遇见根节点相同时,就丢弃评分值不高的结果.

(2)子图

定义 3[1,2](r – 半径斯坦纳图) 给定图 G 和关键词 K,如果 G 中两个节点 u、v 为关键词节点,那么 u 和 v 之间路径上的点(包含 u、v)称为斯坦纳节点.以斯坦纳节点及其相关边构成 G 的子图,且子图中任意一个节点的中心距离最小值不超过 r,称该图为 r 半径斯坦纳图 G_S.

定义 4[11](多中心导出子图) 给定图 G 和关键词 K,如果 G 中节点 u 到关键词节点的距离不超过 $R_{\max}$,则称 u 为中心节点.将中心节点到关键词节点之间路径上的节点称为结构节点.以关键词节点 V_l、中心节点 V_c、结构节点 V_p 及相关边构成的 G_C 称为多中心导出子图.

定义 5[12,13](r – 极大团) 给定图 G 和关键词 K,将由任意两个关键词节点间距离不大于 r 的关键词节点及相关边构成的图 G_r 称为 r – 极大团.

定义 6[14](关联连通簇) 给定图 G 和关键词 K,如果 G 中节点 v 的 r 半径范围内的邻居节点 $N_r(v)$ 至少覆盖 $\min(l, l_0)$ 个关键词,其中 l 为输入的关键词个数,l_0 为给定的阈值,则称 v 为核心节点.通过核心节点而连接的两个节点称为关联连接节点,将从关联连接节点的极大集导出的子图称为关联连通簇(relevancy-connected cluster).

3 关键技术

从现有的工作来看,研究主要围绕搜索算法、排序机制、查询意图、性能评价等问题展开.下面以这些问题为主线,介绍近几年研究进展.

3.1 搜索算法

搜索算法就是根据结果的定义,在图中找到关键词节点之间的关联关系.从查询结果的定义角度去分类,已有的实现方法可分为:基于最小树的方法,基于子图的方法.

(1)基于最小树的方法

基于距离网络启发式等斯坦纳树近似计算思想,学者们在解决结果树搜索问题时提出了不同的方法,较有影响的工作有 BANKS、STAR、DPBF.

BANKS[6]为每个关键词分配一个迭代器,迭代器存放相对应的关键词节点,然后对迭代器中访问的路径长度比较小的节点进行优先扩展,即遍历相邻节点,对访问到的节点使用位与运算来标记关键词访问记录,当发现被遍历节点已经被 K 中每个关键词所访问时,那么就得到一个以当前被访问的节点为根节点的结果树.该方法适用于一般的图模型,但当某个查询关键词匹配很多节点或节点度数比较高,反向扩展算法的执行性能就很差,而且容易丢失相关度高的结果.为提高搜索效率,BANKS-Ⅱ[9]和 BLINKS[10]被提出.

BANKS-Ⅱ采用双向搜索,即在遍历过程中选择入度小的节点优先遍历,在找到结果树之前减少了节点访问数量,从而缩短时间.BLINKS 在建立关键词索引和路径索引基础上提出了基于代价均衡扩展的反向搜索策略,即在遍历过程中尽量让各个迭代器均衡扩展,同时规定了产生的结果树不能有相同根.路径索引采用 Dijkstra 算法构建,把和关键词节点能连通的所有节点

之间最短距离按递增顺序排列,在搜索过程中通过索引查找能够快速得到被访问到的节点与关键词节点距离.BLINKS 的双层索引虽增加了空间开销,但减少了搜索时间.此外,文献[8]采用 Lawler′s procedure[15],相比 BANKS 而言,该方法确保结果的树边权重总和以递增顺序列出,系统的具体实现细节和完整的介绍见文献[16,17].李慧颖[18]借鉴 BLINKS 方法,将其应用到 RDF 数据关键词查询,在实验中将 RDF 数据建模成顶点带标签的实体三元组关联图.为缩小搜索空间,Zhong[19]设计以节点 v 为中心,半径距离为 d 的子图索引,利用此索引提出一种减少探索节点数的方法.

STAR[20]含两个主要任务:生成包含关键词的树,通过筛查相邻节点和修剪对树进行优化.STAR 的生成树方式类似于 BANKS,为每个关键词分配迭代器,不同的地方在于关键词节点扩展细节不同,STAR 采用轮转方式对迭代器中当前节点进行宽度优先搜索而非距离启发式搜索.

针对 BANKS 产生的结果树不一定是最优问题,Ding[7]提出基于动态规划的方法 DPBF.相比 BANKS 而言,DPBF 在响应时间方面具有一定的优势,但也增大了空间开销.为进一步提高查询效率,Yu[21]在 DPBF 树生长、合并方法基础上,针对星型数据库模式的特殊情况,为星型表元组对应的节点建立索引,离线计算任意两点之间距离并建成索引,在扩展过程中利用索引可以降低时耗.

BANKS、STAR、DPBF 都涉及到路径扩展和优化问题,路径扩展需要遍历相邻节点,路径优化需寻找必要节点之间的最短路径,但它们在路径扩展和优化的处理方式各不相同.为处理无法放入内存的大尺寸图数据,学者们提出了图划分方法[10,22~24],将原图划分成若干分区,从而形成不同层次的多粒度图,在内存里搜索超节点图,当需要访问某个超节点的详细信息,就从缓存中取出.

与 BANKS、STAR、DPBF 采取的在线宽度优先搜索方式不同,Gubichev[25]借鉴 BANKS 和 BLINKS 的思想,提出具有可扩展性的基于路标的概要索引 SketchLS,采用轮转方式对关键词节点的概要索引进行宽度优先搜索访问.CSTree[26]针对结构化数据,使用宽度优先搜索方法建立 Voronoi-path 索引,运用 SQL 语句从索引表中获取关键词 k_i 的 Voronoi-path,在此基础上构造结果树.

对于节点数及边数较大的稠密图来说,考虑到建索引的空间和时间复杂度,一些在查询处理过程中需要将索引存储在内存的方法在单机上难以适用,Zhong[27]提出分布式基于磁盘的索引构建方法.

Park[28]和 Zeng[29]对上述结果树提出了不同的看法.Park[28]提出将结果树的一个关键词对应的关键词节点扩充到 p 个.Zeng[29]结合关系数据库特点,提出一种语义的方法,将图中节点分为三种类型:对象型、混合型、关系型,运用与 BANKS 相似的方法对关系型关键词节点和对象型关键词节点进行扩展,在扩展过程中产生候选结果树.

上述的这些方法都是基于边存在性确定的图.针对不确定图数据的关键词查询,Yuan[30]借鉴 BLINKS 设计的索引和搜索算法,在构建索引时增加概率信息,提出基于概率关键词索引的过滤-验证方法.

总的来说,基于最小树的方法可分为两类:一类是将图装载到内存,在内存里对整个图进行启发式搜索;另一类是对图进行预处理,建立距离约束的邻居节点索引,再利用索引搜索结果.前者对内存空间需求较高,易受内存限制,但对结果的路径长度没有约束.后者利用索引,具有速度优势,缺点是结果树的节点之间距离约束在阈值范围内,且如果阈值设置较高,建立索引所需的耗时和空间成本都较大.

(2)基于子图的方法

为适应复杂数据图的关键词查询,Li[2]提出基于 r-半径斯坦纳图的方法 EASE,搜索结果时根据词汇两两配对索引快速定位到 r-半径图并推算出斯坦纳图.该方法可能会丢失相关度高的结果.由于是将词汇成对组合构建索引,当面临规模大的图、图中词汇比较多的情况,索引构建时间长,占用空间大.

Qin[11]提出基于多中心导出子图的方法,其搜索结果也称为社区.结果产生方法是将所有与某个关键词节点距离不超过 $R_{\max}$ 的节点作为中心节点的候选.为提高查询效率,万洁[31]以多中心子图作为结果,提出利用以前的查询关键词、查询结果、查询热度等历史信息来搜索结果.

由于很多用于揭示关键词节点之间关系的中心节点以及中间节点和查询并不相关,给用户带来理解困难的问题,Kargar[12,13]提出 r-极大团模型 r-cliques,将关键词 k_i 对应的关键词节点集 C_i 组合形成求解空间并寻找最优结果,然后递归分解形成新的多个较小的求解空间,并在新的空间继续寻找最优结果,重复分解求解空间和寻找最优结果,直至找到 top-k 结果为止.

基于最小树和基于子图的结果搜索算法都存在缺陷,如信息的不完整性、冗余性.于是 Gao[14]提出将关联连通簇作为查询答案.当输入关键词 K 时,对关键词节点的 r 半径范围内邻居节点 $N_r(v)$ 执行交集、并集运算,构造图 G',然后对 G'的节点 v 运用聚类算法进行聚类,最终形成一系列关联连通簇.

无论是基于最小树还是子图的搜索算法,在建结构索引过程中都对路径长度有限制,实现思路都是利

用图遍历.

3.2 排序

由于需"猜想"关键词之间的拓扑结构,因此,在把检索结果呈现给用户之前,需对搜索到的结果进行度量排序[1].现有方法的结果度量方式多样,通过分析,我们将其分为基于结构的排序、基于内容和结构混合的排序.

(1)基于结构的排序

基于结构的排序策略主要是衡量结果中节点的重要程度以及关键词节点之间关系的紧密程度,影响排序的因素主要有节点权重、边权重、节点数、边数等.

对于最小树的结果排序,BANKS[6]和 BANKS-Ⅱ[9]、STAR[20]对树的根节点和叶子节点权重的平均值、边权重进行综合计算,将计算值作为排序依据,DPBF[7]以结果树的边权重之和作为排序依据.BANKS 设计了两种评分函数,式(1)为其中的一种,BANKS-Ⅱ的评分函数 $S(T)$如式(2)所示,STAR 的评分函数没有明确给出.式中 E_{score}为树中所有边权重计算结果,N_{score}为树中叶子节点和根节点权重的平均值,λ 为微调参数,E 表示根节点到关键词节点的路径中所有边权重之和,N 表示树叶子节点和根节点权重之和.BANKS、BANKS-Ⅱ等采用启发式搜索,按照相关性的近似顺序生成答案,不能保证与真实的排序顺序完全一致.

$$S(T) = (1-\lambda)E_{\text{score}} + \lambda N_{\text{score}} \tag{1}$$

$$S(T) = EN^{\lambda} \tag{2}$$

STAR[20]以结果树的所有边权值总和作升序排列.在实验中,STAR 对边随机赋予 0 到 1 范围内的值.为保障结果的多样性,Golenberg[8]除计算结果树中节点和边权重的综合值之外,还对树结构进行同构判断,如果某结果树与先前产生的结果同构则给予一定的惩罚值.

Zhong[5]以结果树的根节点到关键词节点的路径中所有边权重之和作为排序依据.边的权重计算如式(3),其中 $\omega_e(e)$表示边 e 的权重初始化函数,$\omega_v(u)$表示节点 u 的权重函数.节点的权重计算采用 PageRank 方式迭代计算,直至权重稳定为止.Zhong 与 Golenberg 在给查询结果评分时都对信息重复的结果给予一定的惩罚,但两者的信息重复判断依据不一样,前者是结果树的根节点是否相同,后者是结果树的结构是否同构.

$$\omega_e'(e) = \left(1 - \sqrt{\frac{\omega_v(v)+\omega_v(u)}{2\cdot\max\{\omega_v(y)\,y\in V\}}}\right)\cdot\omega_e(e) \tag{3}$$

对于基于子图的结果度量,r-cliques[12,13]是以任意两个关键词节点之间距离总和作为排序依据.Qin[11]计算社区不同中心节点到各关键词节点的路径中所有边权重的总和,取最小值作为社区排序依据.万洁[31]基于 PageRank 算法,计算社区中节点与查询关键词的相关度,将相关度的平均值作为排序依据.Gao[14]综合考虑查询关键词覆盖数量和簇类的边权重总和,如果在边权重相同情况下覆盖的关键词越多,查询结果的排序就靠前.

(2)基于内容和结构混合的排序

在基于内容和结构混合的排序方法中,内容评分又分为 TF-IDF[2,10,26]、语言模型[32,33].

EASE[2]对关键词在 r-半径图的内容相关度采用基于 TF-IDF 的评分函数计算,将其和关键词节点之间的结构紧密度进行综合计算,将计算值作为排序依据.BLINKS[10]基于匹配语义和图距离语义,将根节点权重、根节点到关键词节点的路径中所有边权重之和、关键词节点与关键词匹配情况(引用 IR 领域中的文本排序方法)综合计算.CSTree[26]将基于 TF-IDF 方式计算的节点权重和路径的边权重进行加权求和.

针对 RDF 图,Elbassuoni[32]采用基于统计语言模型的评分函数,将检索到的子图中三元组当作虚拟文档,统计三元组谓词在虚拟文档中与关键词共同出现次数,推断出可能的查询模式结构,并将与该模式匹配的结果赋以高分,排在队列前面.Mass[33]也采用基于统计语言模型的评分函数,与文献[32]不同的是,前者除了计算查询关键词在结果树中出现的频率之外,还计算图中查询关键词出现的频率,再结合节点权重、边权重形成评分函数.对于节点的文本包含词汇比较多的情况,该方法的计算比较耗时.

在上述的各种度量计算公式中,节点、边权重计算公式各不一样,如节点权重计算有的是默认为 1,有的是简单按节点的出入度数来处理,有的是按度数结合 PageRank 算法分配权重.基于结构的排序与基于内容和结构混合的排序策略相比,前者排序方式比较单一,实现简单,当结构相同时难以进行有效的排序,后者虽然计算复杂,但在排序中考虑到查询结果的内容与关键词的相关度,在查准率方面有一定优势.

3.3 查询意图

由于只使用有限数量的词汇来表达查询需求,关键词的上下文意义不够明确,很多查询结果并不是用户想要的.为了尽可能将所需结果呈现在用户面前,人们除了在排序模型上下功夫之外,还在查询意图方面做了一些研究,如查询重写[34~37]、查询表达式[38].

Zeng[34]提出三阶段的关键词查询范式,采取猜想查询意图及排序、收集用户反馈等手段来逐步明确查询语义.Zhang[35]以斯坦纳图作为结果,运用贝叶斯定理等知识,提出概率查询重写方法.林子雨[36]结合用户兴趣节点集预测情况,运用蚁群优化算法,提出基于概念漂移的查询结果动态优化方法.王新军[37]吸取 Web 检索扩展中的共现思想,当用户对查询结果不满意时,

就在数据图上计算并选择与查询 K 密切程度最高的 L 个词作为查询扩展候选词.

为支持高级查询,允许用户输入必须包含的、可选的、需排除的关键词,Pan[38]引入查询分块图概念,用 Bron-Kerbosch 算法在查询分块图找出满足这些约束条件的结果.

根据查询重写发生在开始查找的前后关系,我们将文献[35]归为事先方法,将文献[34,36,37]归为事后方法.这几种方法都与用户有交互,事后方法主要是利用查询结果、历史日志等数据来获取用户感兴趣的信息,从而进行有针对性的查询重写.

3.4 评价指标

研究者们主要使用时间、查准率(Precision)、查全率(Recall)来对查询效率及查询结果质量进行分析,在实验过程中用于测试的查询关键词是从数据中随机选择的,查询包含的关键词个数一般为 2~6 个,查询次数一般都达到 50 以上.部分文献实验的查询次数较少,没有达到 50 次.对于一个搜索系统来说,用于测试的查询次数不宜低于 50[39].

结合前几节所述内容,表 1 列出了主要方法的对比情况.部分方法由于结合应用领域的数据图特点,不具通用性,因此表 1 没有列出,如文献[32,33]结合了 RDF 图三元组特点.从表 1 可看出,部分方法只关注时间,未关注结果的有效性.表 1 中,在"比较对象"这一列中出现的一些上文未提到的方法,是非图数据关键词查询方法或作者优化之前的方法.

EASE[2]和文献[5]没有给出具体的算法,CSTree[26]运用 SQL 语句来获取结果,因此,在这里就没列出它们的时间复杂度."时间复杂度"一列中符号说明:l 表示查询关键词的个数,m 表示图的边数,n 表示图的节点数,n_i 表示第 i 个 QSUBTREE 的节点数量,k 表示 Top-k 查询的 k,N_p 表示图分割的块数,ε 表示路径权值优化参数,$w_{\max}$表示输入图中的最大边权值,$w_{\min}$表示输入图中的最小边权值,$D_{\max}$表示输入图中节点的最大度,V_i 表示关键词 k_i 对应的关键词节点集.

4 未来研究方向

尽管国内外学者对图数据关键词查询进行研究,并取得了一定的成果,但还没有成熟的产品,仍存在一些实际问题值得进一步研究.

(1)算法效率可扩展性不好,难以适应规模比较大的数据

为满足扩展性,学者提出图划分和集群系统下的处理方法,但在集群环境下的负载均衡、I/O 优化需要更多的研究.

除了从方法本身考虑效率、扩展性之外,可以利用一些其他技术,如云计算、图形处理器(Graphic Process Unit,GPU).作为并行计算的一个分支,GPU 已经被应用在图论领域,如 GPU 加速的 PageRank 算法[40]、Grafight[41](使用 GPU 技术的图遍历引擎).一些已有的图数据关键词查询方法的节点权重计算公式就是基于 PageRank 算法.GPU 硬件和编程技术的发展,给图数据关键词查询带来了新的机遇.然而,如何有效地把相关计算移植到 GPU 上运行是一个挑战.

(2)排序模型是通过经验性的方法构建,理论依据不充分

现有方法对内容相关度、结构紧凑度的计算凭着直观感觉,调参比较困难.以统计学习为基础的排序学习是将多种特征通过融合的方式构建成一个排序函数,在搜索引擎优化、问答系统等领域得到广泛应用,拥有理论及应用上多方面的优点[42].因此,将排序学习应用在图数据关键词查询的排序方面将是一个值得研究的内容.

(3)性能评价缺少统一的标准

现有文献采用的数据集、评价指标并不完全一致,测试的规模也不大,缺乏测试基准,难以进行一个比较全面的评价.因此,建立一个与 INEX、TREC 类似的评价平台是必要的.除从效率、有效性方面去评价方法之外,还要考虑方法的适用范围,如图规模、图的稠密度等,这样有利于后续的研究评估.此外,在现有的使用 Recall、Precision 评价指标中加入评价指标 nDCG(Normalized Discounted Cumulative Gain)也值得考虑.

(4)不能很好地支持高级查询

目前的方法对查询需求的表达能力不够强大,难以表达用户的查询需求.ROU[38]虽对必须、排除、可选的关键词约束进行了研究,但可扩展性、排序等问题没有得到很好的解决.因此,支持高级查询的技术有待进一步的研究.

此外,查询重写、复杂环境下的 top-k 关键词查询也是值得关注的研究方向.在这里,复杂环境主要是指图不确定、图数据频繁更新.在实际应用中,除了碰到图不确定情况外,还会面临图数据经常会更新问题,如增加或删除图的一些节点,影响结果的实时性.Xu[43]针对更新问题,提出了连续关键词查询方法.由于该方法是基于模式图框架,且需要维护不同查询的 top-k 结果,难以适用于通用的图和大量的关键词查询.

表 1　查询方法

方 1 法	结果定义	排序策略	测试数据集	评价指标	比较对象	适用图	时间复杂度
BANKS[6]	最小树	○	DBLP,&1	Time,真实排序与实际排序的差异绝对值	*	有向	$O(l(m+n)\log n+2l\cdot n)$
BANKS-Ⅱ[9]	最小树	○	DBLP,IMDB,&2	Time,Recall,Precision	BANKS	有向,无向	$O(l(m+n)\log n+2l\cdot n)$
Zhong[5]	最小树	○	DBLP	*	*	无向,有向	–
DPBF[7]	最小树	○	DBLP,IMDB	Time,Memory,Consumption,total edge-weight	BANKS,BANKS-Ⅱ,RIU-E,RIU-T	无向,有向	$O(3^l n+2^l((l+\log n)n+m))$
Golenberg[8]	最小树	○	Mondial	Time	*	有向	$O\left(\sum_{i=1}^{k}(n_i\cdot l(m+n\log n))\right)$
BLINKS[10]	最小树	●	DBLP,IMDB,&2	Time,Recall,Precision	BANKS-Ⅱ	有向	$O(l\log N_p\cdot n+l((l-1)+\log N_p(m+n)))$
CSTree[26]	最小树	●	DBLP,IMDB	Time,Precision,Recall	EASE,BLINKS	无向,有向	–
STAR[20]	最小树	○	DBLP,IMDB,YAGO	Time,average scores	BANKS,BANKS-Ⅱ,DPBF,BLINKS	无向	$O\left(\frac{1}{\varepsilon}\frac{w_{\max}}{w_{\min}}m\cdot N(n\log n+m)\right)$
SketchLS[25]	最小树	○	IMDB,&3	the relative difference,Time	DPBF,STAR,MST,BANKS-Ⅱ,Sketch	有向,无向	$O(l^2\log^2 n\cdot(D_{\max}+l\log n))$
EASE[2]	子图	●	DBLife,DBLP,IMDB	Time,Precision	DPBF,DBLife,InfoUnit	无向	–
Qin[11]	子图	○	DBLP,IMDB	Time,Memory	PDall,PDk,BUall,BUk,TDall,TDk	有向,无向	$O(l(n\log n+m))$
r-cliques[12,13]	子图	○	DBLP,IMDB	Time,Precision 紧密度	Qin[20]	无向,有向	$O\left(\left(\sum_{i=1}^{l}\mid V_i\mid\right)^2\cdot k\right)$
Gao[14]	子图	○	DBLP	*	*	有向,无向	$O(n\log n)$

标记:

(1)○为基于结构的排序,●为基于内容和结构混合的排序.

(2)&1 代表 IIT Bombay 的硕博论文信息库,&2 代表 US Patent Database,&3 代表 Social networks 数据 Slashdot、Youtube、Flickr、Orkut.

(3)* 表示文献没有明确的相应信息.

5　结束语

本文从搜索算法等方面对现有的方法进行了分析和总结.从研究现状来看,图数据关键词查询问题的研究从理论的完善到方法的实际应用需要结合新理论和新技术,还有一些创新性研究和大量的完善工作要做.

参考文献

[1] 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 – 32.
Ma Shuai,Li Jia,Liu Xudong,et al.Graph query:New search in a social computing Era[J].Journal of Communications of the China Computer Federation,2012,8(11):26 – 32.(in Chinese)

[2] Li G,Ooi BC,Feng J,et al.Ease:An effective 3-in-1 keyword search method for unstructured,semi-structured and structured data[A].Proceedings of the ACM SIGMOD International Conference on Management of Data[C].New York:Association for Computing Machiner,2008.3 – 914.

[3] 林子雨,杨冬青,王腾蛟,等.基于关系数据库的关键词查询[J].软件学报,2010,21(10):2454 – 2476.
Lin Ziyu,Yang Dongqing,Wang Tengjiao,et al.Keyword search over relational databases[J].Journal of Software,2010,21(10):2454 – 2476.(in Chinese)

[4] Aggarwal CC,Wang H.Managing and Mining Graph Data[M].New York:Springer-Verlag,2010.249 – 274.

[5] Zhong Ming,Liu Mengchi.Ranking the answer trees of graph search by both structure and content[A].Proceedings of 1st Joint International Workshop on Entity-Oriented and Semantic Search,JIWES'12- Co-located with the 35th ACM SIGIR Conference[C].New York:Association for Computing Machiner,2012.344 – 350.

[6] Bhalotia G,Nakhe C,Hulgeri A,et al.Keyword searching and browsing in databases using BANKS[A].Proceedings of the International Conference on Data Engineering[C].Washington:IEEE Computer Society,2002.431 – 440.

[7] Ding Bolin,Yu Jeffrey Xu,Wang Shan,et al.Finding top-k

min-cost connected trees in databases[A]. Proceedings of the International Conference on Data Engineering[C]. New Jersey: IEEE Computer Society, 2009. 836 – 845.

[8] Golenberg K, Kimelfeld B, Sagiv Y. Keyword proximity search in complex data graphs[A]. Proceedings of the ACM SIGMOD International Conference on Management of Data[C]. New York: Association for Computing Machiner, 2008. 927 – 940.

[9] Kacholia V, Pandit S, Chakrabarti S, et al. Bidrectional expansion for keyword search on graph databases[A]. Proceedings of the 31st International Conference on Very Large Data Bases[C]. New York: Association for Computing Machiner, 2005. 505 – 516.

[10] He H, Wang H, Yang J, et al. BLINKS: ranked keyword searches on graphs[A]. Proceedings of the ACM SIGMOD International Conference on Management of Data[C]. New York: Association for Computing Machiner, 2007. 305 – 316.

[11] Qin L, Yu JX, Chang L, et al. Querying communities in relational databases[A]. Proceedings of the International Conference on Data Engineering[C]. New Jersey: IEEE, 2009. 724 – 735.

[12] Kargar M, An A. Keyword search in graphs: finding r-cliques[J]. VLDB Journal, 2011, 4(10): 681 – 692.

[13] Kargar M, An A. Efficient top-k keyword search in graphs with polynomial delay[A]. Proceedings of the International Conference on Data Engineering[C]. Los Alamitos: IEEE Computer Society, 2012. 1269 – 1272.

[14] Gao BJ, Chen Z, Kang Q. Information-complete and redundancy-free keyword search over large data graphs[A]. Proceedings of the 21st ACM International Conference on Information and Knowledge Management[C]. New York: Association for Computing Machiner, 2012. 2639 – 2642.

[15] Lawler E. A procedure for computing the k best solutions to discrete optimization problems and its application to the shortest path problem[J]. Management Science, 1972, 18(7): 401 – 405.

[16] Achiezra H, Golenberg K, Kimelfeld B, et al. Exploratory keyword search on data graphs[A]. Proceedings of the ACM SIGMOD International Conference on Management of Data[C]. New York: Association for Computing Machiner, 2010. 1163 – 1166.

[17] Sagiv Y. A personal perspective on keyword search over data graphs[A]. Proceedings of the 16th International Conference on Database Theory[C]. New York: Association for Computing Machiner, 2013. 21 – 32.

[18] 李慧颖, 瞿裕忠. KREFAG: 基于实体三元组关联图的 RDF 数据关键词查询方法[J]. 计算机学报, 2011, 34(5): 825 – 834.
Li Huiying, Qu Yuzhong. KREFAG: Keyword query approach over rdf data based on entity-triple association graph[J]. Chinese Journal of Computers, 2011, 34(5): 825 – 834. (in Chinese)

[19] Zhong M, Liu M. Efficient keyword proximity search using a frontier-reduce strategy based on d-distance graph index[A]. International Database Engineering and Applications Symposium[C]. New York: Association for Computing Machiner, 2009. 206 – 216.

[20] Kasneci G, Ramanath M, Sozio M, et al. STAR: stenier-tree approximation in relationship graphs[A]. Proceedings of the IEEE 25th International Conference on Data Engineering[C]. New Jersey: IEEE, 2009. 868 – 879.

[21] Yu Xiaohui, Shi Huxia. CI-Rank: Ranking keyword search results based on collective importance[A]. Proceedings of the IEEE 28th International Conference on Data Engineering[C]. New Jersey: IEEE Computer Society, 2012. 78 – 89.

[22] Dalvi BB, Kshirsagar M, Sudarshan S. Keyword search on external memory data graphs[J]. VLDB Journal, 2008, 1(1): 1189 – 1204.

[23] Rose Catherine K, Sudarshan S. Graph clustering for keyword search[A]. Proceedings of the 15th International Conference on Management of Data[C]. Pitamputa: Computer Society of India, 2009. 971 – 982.

[24] Yang S, Yan X, Zong B, et al. Towards effective partition management for large graphs[A]. Proceedings of the ACM SIGMOD International Conference on Management of Data[C]. New York: Association for Computing Machiner, 2012. 517 – 528.

[25] Gubichev A, Neumann T. Fast approximation of steiner trees in large graphs[A]. Proceedings of the 21st ACM International Conference on Information and Knowledge Management[C]. New York: Association for Computing Machiner, 2012. 1497 – 1501.

[26] Li Guoliang, Feng Jianhua, Zhou Xiaofang. Providing built-in keyword search capabilities in RDBMS[J]. VLDB Journal, 2011, 1(20): 1 – 19.

[27] Zhong Ming, Liu Mengchi. A distributed index for efficient parallel top-k keyword search on massive graphs[A]. Proceedings of the 12th ACM International Workshop on Web Information and Data Management, Co-located with CIKM[C]. New York: Association for Computing Machiner, 2012. 27 – 32.

[28] Park, C S. An effective keyword search method for graph-structured data using extended answer structure[A]. Proceedings of the 13th International Conference on Computational Science and Its Applications[C]. Berlin: Springer Verlag, 2013. 620 – 635.

[29] Zeng Z, Bao ZF, L ML, et al. A semantic approach to keyword search over relational databases[A]. Proceedings of the 32th International Conference on Conceptual Modeling[C].

Berlin:Springer Verlag,2013.241 – 254.

[30] Yuan Y, Wang GR, C L, et al. Efficient keyword search on uncertain graph data[J]. IEEE Transactions on Knowledge and Data Engineering,2013,25(12):2767 – 2779.

[31] 万洁,张文胜,朱青,等.基于历史信息提升关键字查询效率[J].小型微型计算机系统,2011,32(11):2192 – 2197.
Wan Jie, Zhang Wen-Sheng, Zhu Qing, et al. Improving keyword search efficiency on history information[J]. Journal of Chinese Computer Systems, 2011, 32(11):2192 – 2197. (in Chinese)

[32] Elbassuoni S, Blanco R. Keyword search over rdf graphs[A]. Proceedings of the ACM International Conference on Information and Knowledge Management[C]. New York: Association for Computing Machiner,2011.237 – 242.

[33] Mass Y, Sagiv Y. Language models for keyword search over data graphs[A]. Proceedings of the 5th ACM International Conference on Web Search and Data Mining[C]. New York: Association for Computing Machiner,2012.363 – 372.

[34] Zeng Z, Bao Z, Ling TW, et al. iSearch: An interpretion based framework for keyword search in relational databases[A]. Proceedings of the 3rd International Workshop on Keyword Search on Structured Data[C]. New York: Association for Computing Machiner,2012.3 – 10.

[35] Zhang L, Tran T, Rettinger A. Probabilistic query rewriting for efficient and effective keyword search on graph data[J]. Proceedings of the VLDB Endowment, 2013, 6(14): 1642 – 1653.

[36] 林子雨,邹权,赖永炫,等.关系数据库中的关键词查询结果动态优化[J].软件学报,2014,25(3):528 – 546.
Lin Ziyu, Zou Quan, Lai Yongxuan, et al. Dynamic result optimization for keyword search over relational databases[J]. Journal of Software,2014,25(3):528 – 546. (in Chinese)

[37] 王新军,闫实,彭朝晖,等. Extractor:支持查询重构的高效数据库关键词检索系统[J].电子学报,2014,42(2):209 – 216.
Wang Xinjun, Yan Shi, Peng Chaohui, et al. Extractor: A query-reformulation embeeded efficient keyword search system over relational databases [J]. Acta Electronica Sinica, 2014, 42(2):209 – 216. (in Chinese)

[38] Pan YF, Wu YQ. ROU: Advanced keyword search on graph [A]. Proceedings of the 22nd ACM International Conference on Information and Knowledge Management[C]. New York: Association for Computing Machiner,2013.1625 – 1630.

[39] Manning CD, Raghavan P, Schutze H. Introduction to Information Retrieval [M]. New York: Cambridge University Press,2008.139 – 161.

[40] VDuong N T, Nguyen Q A P, Nguyen A T, et al. Parallel pagerank computation using GPUs[A]. Proceedings of the Third Symposium on Information and Communication Technology[C]. New York: Association for Computing Machiner, 2012.223 – 230.

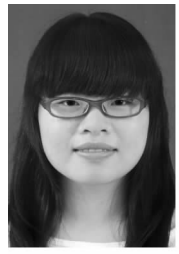
[41] GitHub Inc. Grafight. https://github.com/ashwinraghav/Grafight[OL].2014 – 4 – 11.

[42] 花贵春,张敏,刘奕群,等.基于查询聚类的排序学习算法[J].模式识别与人工智能,2012,25(1):118 – 123.
Hua Guichun, Zhang Min, Liu Yiqun, et al. Learning to rank based on query clustering [J]. Journal of Pattern Recognition and Artificial Intelligence, 2012, 25(1): 118 – 123. (in Chinese)

[43] Xu YW, Guan JH, Li FR, et al. Scalable continual top-k keyword search in relational databases[J]. Data and Knowledge Engineering,2013,86:206 – 223.

作者简介

杨书新　男,1978 年 5 月出生,江西九江人,博士、硕士生导师、副教授,主要研究方向为数据管理、信息检索.
E-mail:yimuyunlang@sina.com



徐丽萍　女,1990 年 5 月出生,江西贵溪人,硕士研究生,主要研究方向为信息检索.
E-mail:xuliping0519@163.com