图数据中 Top-k 属性差异 q-clique 查询

孙焕良[1)] 卢 智[1)] 刘俊岭[1),2)] 于 戈[2)]

[1)](沈阳建筑大学信息与控制工程学院 沈阳 110168)

[2)](东北大学信息科学与控制学院 沈阳 110004)

摘 要 紧密子图发现在许多现实世界网络应用中具有重要的研究意义.提出一种新的紧密子图发现问题——Top-k 属性差异 q-clique 查询,找出图中 k 个节点间属性具有最大差异的 q-clique.属性差异 q-clique 是一种结合图的结构特征和节点属性的紧密子图,在作者合作关系图数据中,该查询可以发现属性(如研究领域或所属单位)上不同的具有紧密合作关系的团队.给出了 q-clique 的属性差异度量,证明了该问题为 NP 难问题.采用分支限界策略,提出一种有效求解问题的算法 AD-Qclique,同时依照 best-first 排序思想优化节点访问次序进一步提高算法性能.ACM 作者信息数据集上的实验表明,算法 AD-Qclique 效率远优于基本算法 BSL,并且结果中作者皆具有较高的 H-index 值及广泛的研究领域.

关键词 图数据;紧密子图;属性差异;分支限界;节点访问次序

中图法分类号 TP311 **DOI 号**: 10.3724/SP.J.1016.2012.02265

Top-k Attribute Difference q-clique Queries in Graph Data

SUN Huan-Liang[1)] LU Zhi[1)] LIU Jun-Ling[1),2)] YU Ge[2)]

[1)](*Department of Information & Control Engineering, Shengyang Jianzhu University, Shenyang* 110168)

[2)](*School of Information Science and Engineering, Northeastern University, Shenyang* 110004)

Abstract The problem of dense subgraph discovery has important research meanings in many applications in real-world networks. This paper discusses a new problem about dense subgraph discovery, Top-k attribute difference q-clique queries, to find k q-cliques in which the dissimilarity between each vertices' attribute is as large as possible. The attribute difference q-clique is a dense subgraph combining both the characters of structure and attributes content, and the queries will find teams in which members' attribute (e.g., research field or affiliation) are different from each other in a co-authorship network. This paper gives the attribute difference measure of q-clique and shows the problem of finding the maximum-difference q-clique is NP-Hard. This paper proposes a query algorithm AD-Qclique for the problem of attribute difference q-clique by using branch and bound strategy and optimizes the node visit order according to best-first order. This paper conducts extensive experiments through ACM author information dataset, which show that AD-Qclique obtains an order of magnitude speed-up comparing to BSL and all the authors in the result set have high H-index value and wide field of study.

Keywords graph data; dense subgraph; attribute difference; branch and bound; node visit order

---

收稿日期:2012-06-30;最终修改稿收到日期:2012-08-27.本课题得到国家自然科学基金重点项目(61033007)、国家自然科学基金项目(61070024,61272179)、中央高校基本科研业务费专项资金(N100704001)资助.**孙焕良**,男,1969 年生,博士,教授,主要研究领域为数据库和数据挖掘等.E-mail: sunhl@sjzu.edu.cn.**卢 智**,男,1988 年生,硕士研究生,主要研究方向为图数据挖掘和空间数据库等.**刘俊岭**,女,1972 年生,博士研究生,主要研究方向为空间数据库与数据挖掘.**于 戈**,男,1962 年生,教授,博士生导师,研究领域为数据库理论与技术、分布式与并行系统等.

1 引 言

近年来，由于社会网络、生物信息等领域的快速发展，使得图数据挖掘成为了一个备受关注的研究领域[1]. 紧密子图是具有特殊结构与性质的子图，紧密子图的发现在许多现实世界网络应用中具有重要的研究意义. 例如，社会网络中的社区发现、金融网络中的统计分析[2]、蛋白质交互网络的功能结构发现[3]、电子商务中的协同过滤[4]、病毒式营销中的影响力团体发现[5]等. 随着网络内容信息的丰富，一些学者开始研究结合图结构和节点属性进行图挖掘的相关问题并取得了一定的成果，例如团队形成[6]、图聚类[7]、近似子图匹配[8]等.

本文提出一种结合图结构和节点属性的紧密子图发现问题—属性差异 q-clique 查询. 作为紧密子图的一种，q-clique 是具有 q 个节点的完全子图. 在社交网络中，q-clique 可以表示 q 个相互认识的用户；在作者合作关系网络中，q-clique 可以表示 q 个彼此合作过的作者. 例如图 1 作者合作关系图中的 3-clique $Q_1=\{r_1,r_2,r_3\}$、$Q_2=\{r_5,r_6,r_7\}$、$Q_3=\{r_5,r_6,r_8\}$、$Q_4=\{r_5,r_7,r_8\}$、$Q_5=\{r_6,r_7,r_8\}$都代表由 3 个彼此合作的作者形成的团队.

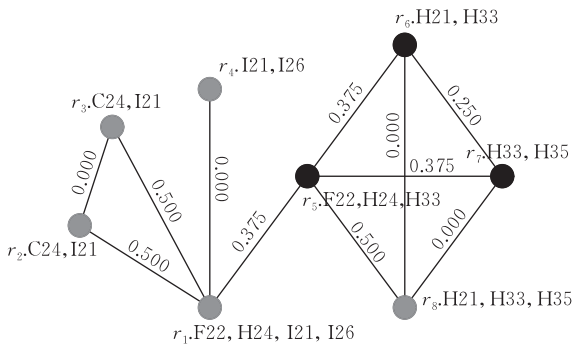


图 1 作者合作关系图

属性差异 q-clique 查询是找出图中节点间属性尽可能地具有较大相异度的 q-clique. 给定 q-clique 差异度量，发现 k 个具有最大差异的 q-clique 称为 Top-k 属性差异q-clique. 该问题旨在使所找出的 q-clique 的属性内容尽可能丰富，同时子图中各节点间的属性内容又尽可能的不同. 图 1 中除 Q_5 只含有 3 种不同的属性值外，Q_1、Q_2、Q_3、Q_4 都拥有 5 种不同的属性值，因而它们具有相同的属性内容丰富程度. 但属性差异 q-clique 问题将 Q_4 视为更好结果，因为 Q_4 不仅属性内容丰富，而且节点间彼此各有不同. 对于作者合作关系图数据中的属性差异 q-clique 查询可以发现属性（如研究领域或所属单位）上不同的合作团队，这样的团队更具有活力或发展潜力. 作为一种特征 q-clique 查询，属性差异 q-clique 查询进一步丰富了紧密子图查询问题.

属性差异 q-clique 查询问题的挑战在于：(1) q-clique 的属性差异度量定义及 Top-k 属性差异 q-clique 排序；(2) q 值不固定的 q-clique 问题是 NP 难问题，需要设计高效的查询算法.

本文采用二元变量相异度的定义，计算出各邻接节点间的属性相异度并将其作为相应邻接边的边权值，使属性差异的 q-clique 发现问题转换成最大权值 q-clique 发现问题.

一种简单的解决方法便是先找出图中所有的 q-clique，然后计算出它们的差异度，并找出其中具有最大差异度的 q-clique 输出，从图中删除该 q-clique 的节点继续查询. 但是，当 q 值不为常量时，该问题为 NP 难问题，因而随着图中节点数的增加，产生的 q-clique 的数目将会呈指数增长. 由此导致这种方法不仅效率低，而且缺乏良好的可伸缩性.

为解决上述问题，本文提出一种分支限界算法 AD-Qclique 来处理属性差异 q-clique 查询. 在搜索结果的过程中不仅利用图的拓扑结构性质剪去不能形成 q-clique 的树枝，同时利用边上的邻接节点的属性相异度计算相应分支的差异度上界，产生剪枝条件来减小搜索空间，加快搜索过程. 然后，本文依照 best-first 排序思想提出基于优先次序的 AD-Qclique 算法，进一步提高算法性能.

综上所述，本文主要贡献如下：

(1) 提出一种节点间属性相异度尽可能大的紧密子图结构查询问题——Top-k 属性差异q-clique 查询，给出了该问题的形式化定义，并证明该问题具有 NP-Hard 复杂性.

(2) 提出一种分枝限界算法 AD-Qclique，利用边权值形成剪枝条件，剪切掉大部分不符合要求的节点. 提出基于优先次序的 AD-Qclique 算法，进一步提高算法效率.

(3) 使用真实数据集在不同参数下对算法性能及有效性进行实验分析.

本文第 2 节综述相关研究工作；第 3 节定义属性差异 q-clique 问题；第 4 节给出解决问题的有效算法；实验结果及分析在第 5 节中给出；第 6 节为本文的结论.

2　相关工作

随着各种网络数据规模的急剧增长，紧密子图发现相关算法的可伸缩性问题引起了广泛关注[9-12]. Cheng 等人[9]提出了一种在大规模网络上进行最大 clique 枚举的外存算法 ExtMCE，通过扩展原图的核心部分来实现原图的最大 clique 枚举. 同时，Cheng 等人[10]还发明了一种新颖的自顶向下从大 k 值到小 k 值递归计算 k-cores 的外存算法 EMcore. 算法 EMcore 对原图进行图划分，每次处理一部分子图，避免随机访问磁盘数据的庞大开销. Huang 等人[11]提出了一种基于无参数的聚类分析的社区发现算法，能够发现网络中的中心点、边界点以及具有层级的社区结构，具有良好的可伸缩性.

由于现实世界网络构成的图中存在许多不确定性，韩蒙等人[12]提出了在不确定图中发现紧密子图问题，并提出了基于树搜索的 TreeClose 算法及优化策略.

以上工作主要集中于图的拓扑结构特征，而伴随着各种网络的快速发展，网络上海量的内容信息也得到了广泛重视. Lappas 等人[6]研究如何在社会网络中找到一个团队既具备完成特定任务的能力，同时团队中各成员间的沟通成本尽可能低. Zhou 等人[7]提出了基于结构和属性相似性的图聚类算法 SA-Cluster，将属性转化为一种附加的结构，使得属性与结构统一，最终将原图划分为 k 簇并且同一簇中节点属性值尽可能相同. Zhu 等人[8]对同时考虑结构信息和属性信息的近似图匹配问题进行了研究，首先利用算法 SA-Cluster 对图建立索引，然后对各划分之间的连接关系建立索引，最后使用贪心算法找出最佳路径匹配. Kargar 等人[13]提出 r-cliques 的概念作为一种关键字搜索问题的新方法，要求包含关键字内容的各节点间的最短路径长度不能超过 r，使它们之间更加紧密；同时，建立索引避免搜索图中所有的节点，缩减搜索空间.

与上述工作相比，本文所提出的 Top-k 属性差异 q-clique 查询，在结合拓扑结构和属性内容发现紧密子图的同时，希望子图中各节点属性内容尽可能地具有差异，使得子图中属性内容尽可能丰富且节点的属性内容各有特点.

3　问题定义

本节给出属性差异 q-clique 问题的背景和相关定义.

3.1　q-clique

给定一个带权简单无向图 $G=(V,E,w)$，其中 V 是无向图 G 的顶点集，$E\subseteq V\times V$ 是图 G 的边集，$w(u,v)$ 表示以顶点 u 和 v 为端点的边的权值，用 A^G 表示图中每个顶点的相关属性. 对于图 G 中每个顶点 $v\in V$，$A^G(v)$ 表示该节点的属性值，它为一个列表 $[a_1,a_2,\cdots,a_n]$，其中 a_i 为节点 v 的一个属性值. 以图 1 为例，图中每个节点代表一位学者，具有“研究领域”属性，图中边表示彼此具有合作关系. 根据 ACM Computing Classification System [1998 Version]的命名规则[14]，节点 r_1 的研究领域属性 $A^G(r_1)=[\mathrm{F22},\mathrm{H24},\mathrm{I21},\mathrm{I26}]$.

定义 $n=|V|$，$m=|E|$，图 G 的大小用 $|G|$ 表示，其值为图中节点的数目，即 $|G|=n$. 给定一个顶点集 $S\subseteq V$，定义由 S 形成的图 G 的导出子图为 $G_S=(V_S=S,\ E_S=\{(u,v):u,v\in S,\ (u,v)\in E\})$. 图 G 中任一顶点 $v\in V$ 的邻接点的集合为 $nb(v)=\{u:(u,v)\in E\}$，同时顶点 v 的度为 $d(v)=|nb(v)|$.

图 G 的一个 clique 是使得导出子图为所有节点相互连接的完全图的顶点集的子集，q-clique 即为具有 q 个节点的 clique. 图 1 中 $\{r_5,r_6,r_7\}$ 即为一个 3-clique，表示 3 个彼此具有合作关系的学者组成的团队.

3.2　属性差异 q-clique

节点中的属性大多为二元变量，二元变量是只有两种状态 0 或 1 的变量，其中 0 表示该变量不出现，1 表示该变量出现. 本文采用二元变量相异度的定义计算出各邻接节点间的属性相异度，如定义 1.

定义 1.　节点间的属性相异度. 给定一个图 G，具有二元变量属性值的两个节点 u 和 v，其中 $u,v\in V(G)$ 且 $(u,v)\in E(G)$，节点间的属性相异度定义为

$$diff(u,v)=\begin{cases}\dfrac{r+s}{q+r+s+t}, & r\neq 0 \text{ 且 } s\neq 0\\ 0, & \text{其它}\end{cases}\tag{1}$$

其中，r 是节点 u 值为 1 而节点 v 值为 0 的属性值的数目；s 是节点 u 值为 0 而节点 v 值为 1 的属性值的数目；q 是节点 u 和节点 v 值都为 1 的属性值的数目；t 是节点 u 和节点 v 值都为 0 的属性值的数目.

图 1 中 $q+r+s+t=8$，该值为图中所有属性值的个数. 以 r_1，r_2 节点为例，$A^G(r_1)=[F22,H24,I21,I26]$，$A^G(r_2)=[C24,I21]$，则 $r=3$（来自于 F22，H24，I26），$s=1$（来自于 C24），$diff(r_1,r_2)=0.500$.

本文将图 G 中任意两邻接点 u，v 之间的属性相异度作为 $w(u,v)$ 的取值，从而得到 q-clique 的属性差异度的定义.

定义 2. q-clique 的属性差异度. 设 $Q\subseteq G$ 为图 G 中的一个 q-clique，其节点集表示为 $\{v_1,v_2,\cdots,v_q\}$，该 q-clique 的属性差异度定义为

$$dvalue_q(Q)=\sum_{i=1}^{q}\sum_{j=i+1}^{q}w(v_i,v_j) \tag{2}$$

其中 $w(v_i,v_j)=diff(v_i,v_j)$ 为以 v_i 和 v_j 为端点的边的权值，即两个节点之间的属性相异度.

图 1 中 r_5，r_6，r_7 节点组成的 3-clique 的属性差异度为 $dvalue_3(\{r_5,r_6,r_7\})=w(r_5,r_6)+w(r_5,r_7)+w(r_6,r_7)=0.375+0.375+0.250=1.000$.

当两个 q-clique 具有相同属性差异度时，并不代表着二者就具有相同的重要性. 下面给出 clique 的跨度定义来对具有相同差异度的 q-clique 进行区别.

定义 3. clique 的跨度. 假设 $C\subseteq G$ 为图 G 中的一个 clique，其边集表示为 E_C，该 clique 的跨度定义为

$$span(C)=\arg\max_{e\in E_C}w(e)-\arg\min_{e\in E_C}w(e) \tag{3}$$

其中，$w(e)$ 为边 e 的权值，$\arg\max_{e\in E_C}w(e)$ 为边集中的最大边权值，$\arg\min_{e\in E_C}w(e)$ 为边集中的最小边权值.

为满足子图中节点彼此间各有特点的目标，当两个 q-clique 的差异度相同时，较小跨度的 q-clique 将更有意义.

本文中 q-clique 具有较大的属性差异度将被视为较好的结果，而当两个 q-clique 的差异度相同时，将跨度较小的视为较好的结果. 由于输出所有的 q-clique 缺乏实际意义，本文试图找出 k 个具有最大属性差异度且具有最小跨度的 q-clique. 同时，由于一些节点可能出现在多个具有较大属性差异的 q-clique 中而导致查询结果出现大量相同节点，所以输出的 k 个结果应该避免出现公共节点. 本文研究问题如问题 1.

问题 1. Top-k 最大属性差异 q-clique 查询.

给定一个图 G，一个结果数目参数 k，一个节点数目参数 q，Top-k 属性差异 q-clique 查询将找出一个集合 S，S 为图 G 中具有最大属性差异度且具有最小跨度的 k 个 q-clique 组成的集合且 S 中 q-clique 彼此之间没有公共节点.

定理 1. 问题 1 Top-k 最大属性差异 q-clique 是一个 NP 难问题.

证明. 首先，将问题 1 转换为其判定版本，给定一个图 G，一个参数 q，一个常量 c，是否存在一个具有属性差异度 c 的 q-clique. 假定图 G 中每条边都具有权值 $\frac{2c}{q(q-1)}$，故而问题 1 被转换为证明一个图中是否存在一个 q-clique，因为 q-clique 的差异度一定为 c. 由于后者已证明为 NP 难问题[15]，因此问题 1 也为 NP 难问题. 证毕.

4 查询处理算法

本文首先给出了属性差异 q-clique 问题的基本算法，然后提出一种分支限界算法. 利用图节点信息，设计了基于优先访问次序的优化查询算法.

4.1 基本算法 BSL

一种直接的 Top-k 属性差异 q-clique 查询问题的解决方法是首先利用图的拓扑结构性质找出图中所有的 q-clique，然后计算出它们的属性差异度，并找出其中具有最大差异度的结果输出. 同时，从图中删除该 q-clique 中的节点避免后续结果出现公共节点，重复上述查询过程直到 k 个结果输出或者图中没有 q-clique 为止.

算法 1 描述了基本算法 BSL 的细节. 步 5 对过程 1 进行调用找出图中所有的 q-clique. 然后步 6～9 对 q-clique 的差异度进行计算并将具有最大差异度的 q-clique 存于 Q 中，最后步 10～11 将 Q 中结果存于 L_q 中输出. 过程 Recursive_Search_BSL 中步 5～11 借鉴文献[16]中提出的 clique 遍历问题的剪切技术，其思想是利用 clique 中节点度的性质对无法形成目标 clique 的分支进行预先剪切，提高了 clique 遍历效率.

算法 1. Baseline Algorithm(BSL).

输入：带权简单无向图 G，不小于 1 的正整数 k，不小于 3 的正整数 q

输出：G 中具有最大属性差异度的无公共节点的 k 个 q-clique 的集合 L_q

1. 初始化长度为 k 的数组 L_q；
2. 初始化栈 S；
3. for i from 1 to k do
4. $Q\leftarrow\varnothing$，$dvalue(Q)\leftarrow 0$；
5. Recursive_Search_BSL($G,q,0,\varnothing,V(G),S$)；
6. while $S\neq\varnothing$ do

```
7.      Q′←pop S;
8.      if Q′具有最大的属性差异度 then
9.        Q←Q′;
10.   if Q≠∅ then
11.     Q 添加入 L_q, G=G−Q;
12. Return L_q.
Recursive_Search_BSL(G,q,l,R,P,S)
1. C←∅;
2. if l=q then
3.   S←push R;
4. for 每个 P 中的节点 v do
5.   if d(v)<q−1 then
6.     continue;
7.   for 每个 nb(v)中的节点 u do
8.     If u.id>v.id then
9.       C←C∪{u};
10.  if |P∩C|<q−l then
11.    continue;
12.  Recursive_Search_BSL(G,q,l+1,R∪v,P∩C,S).
```

本文所研究的问题中参数 q 不是固定常量，问题为 NP 难问题，因而随着图中节点数的增加，产生的 q-clique 的数目将会急剧增长. 基本算法 BSL 遍历出图中所有的 q-clique 将导致算法效率极低且缺乏良好的可伸缩性.

4.2　分支限界算法 AD-Qclique

为解决 BSL 算法的不足，本文提出一种分枝限界算法 AD-Qclique. 在搜索结果的过程中，AD-Qclique 算法不仅利用图的拓扑结构性质，而且利用图中边上的邻接节点属性相异度计算相应分支的差异度上界，产生剪枝条件来减小搜索空间.

为了方便设计图中边权值形成的剪枝条件，本文给出如下定义.

定义 4.　节点的 t-边权和. 给定一个图 G 和一个正整数 t，$u\in V(G)$ 为图中一个节点且 $nb(u)\geq t$，节点 u 的 t-边权和定义为

$$sum_t(u)=\sum_{i=1}^{t}w(e_i) \tag{4}$$

其中 $w(e_i)$ 为节点 u 邻接边的边权值中第 i 大边权值.

例如，图 1 中节点 r_5 的 2-边权和 $sum_2(r_5)=w(r_5,r_8)+w(r_5,r_6)=0.875$.

定义 5.　节点集合的 t-边权和. 给定一个图 G 和一个正整数 t 和一个节点集合 $P\subseteq G$ 且 $\forall u\in P$ 满足 $nb(u)\geq t$，集合 P 的 t-边权和定义为

$$sum_t(P)=\arg\max_{u\in P} sum_t(u) \tag{5}$$

图 1 中 $\{r_5,r_6,r_7\}$ 的 2-边权和 $sum_2(\{r_5,r_6,r_7\})=sum_2(r_5)=0.875$.

定义 6.　clique 的差异度上界函数. 给定一个图 G，一个 clique $R\subseteq G$，一个正整数 q 且 $|R|<q$，P 为一个候选节点集合且 P 中任一节点都与 R 中所有节点相连接，该 clique 的差异度上界函数定义为

$$d_q^+(R)=dvalue_{|R|}(R)+\sum_{i=|R|}^{q-1}sum_i(P) \tag{6}$$

图 1 中 clique $\{r_5,r_6\}$ 的差异上界函数值为 $d_3^+(\{r_5,r_6\})=w(r_5,r_6)+sum_2(r_7)=0.375+0.625=1.000$.

定理 2.　给定一个图 G，δ 为 Top-k 属性差异 q-clique 中的最小属性差异度，一个 clique $R\subseteq G$，$|R|<q$，$P\subseteq V(G)$ 为一个候选节点集合且 P 中任一节点都与 R 中所有节点相连接，如果 $d_q^+(R)<\delta$，则 R 不会成为 Top-k 属性差异 q-clique.

证明.　令 Q 为由 R 扩展而得到的所有 q-clique 的集合中具有最大属性差异度的 q-clique. 因此要想证明定理 2，只需证明 $\sum_{i=|R|}^{q-1}sum_i(P)\geq dvalue_q(Q)-dvalue_{|R|}(R)$. 因为 $Q\subseteq G$，对任一正整数 t，$\forall u\in Q-R$ 满足 $sum_{t,Q}(u)\leq sum_{t,G}(u)$. 同时 $V(Q)-V(R)\subseteq P$，则不等式成立. 因此 $dvalue_q(Q)\leq d_q^+(R)\leq\delta$，定理 2 成立.　证毕.

算法 AD-Qclique 的具体执行过程如算法 2 所示.

算法 2.　AD-Qclique.

输入：带权简单无向图 G，k，q

输出：G 中具有最大属性差异度的无公共节点的 k 个 q-clique 的集合 L_q

```
1. 初始化长度为 k 的数组 L_q;
2. Q←∅, dvalue(Q)←0;
3. for i from 1 to k do
4.   Recursive_Search_AD(G,q,0,∅,V(G),Q);
5.   if Q≠∅ then
6.     将 Q 添加入 L_q, G=G−Q;
7.   else
8.     break;
9. Return L_q.
Recursive_Search_AD(G,q,l,R,P,Q)
1. C←∅;
2. if l=q then
3.   if Q=∅ then
4.     Q←R;
5.   if R 具有最大的属性差异度 then
6.     Q←R;
7. for 每个 P 中的节点 v do
8.   if v 不能形成 q-clique then
9.     continue;
10.  if d+(R∪v)<dvalue(Q) then
11.    continue;
12.  Recursive_Search_AD(G,q,l+1,R∪v,P∩C,Q).
```

与算法 BSL 相同，算法 AD-Qclique 同样是每次查询出最大属性差异 q-clique 将其输出，然后从图中删除该 q-clique 的节点避免后续结果出现公共节点.但不同的是，在查询 Top-1 属性差异 q-clique 的过程中，算法 AD-Qclique 并没有遍历所有的 q-clique，而是利用定理 2 的剪枝条件（见过程 2 步 10）剪切掉差异度上界值小于当前最大属性差异度的分支，减小搜索空间.为清晰地描述算法 AD-Qclique 的执行过程，本文在图 2 中给出了算法进行 Top-1 属性差异 3-clique 查询时的空间状态树.

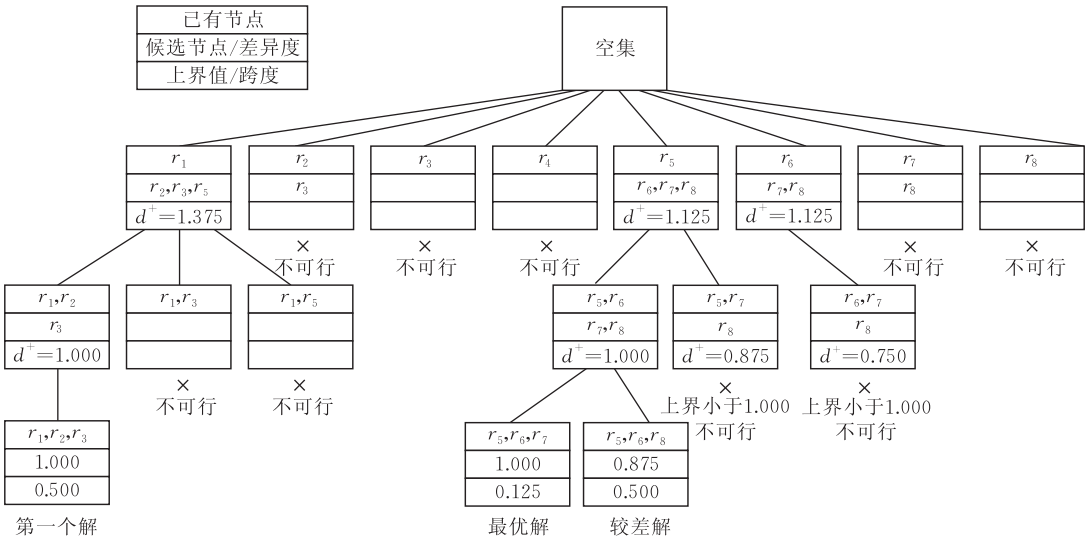


图 2 算法 AD-Qclique 的状态空间树

算法首先通过 r_1 搜索到第一个 3-clique $\{r_1,r_2,r_3\}$，而 r_2,r_3,r_4 都无法形成 3-clique，所以剪切相应分支.然后，算法通过 r_5 搜索到 $\{r_5,r_6,r_7\}$，虽然 $\{r_1,r_2,r_3\}$ 与 $\{r_5,r_6,r_7\}$ 的属性差异度相等，但是其跨度为 0.500 大于后者的 0.125，因此 $\{r_5,r_6,r_7\}$ 为当前最优解.由于 $\{r_5,r_6,r_8\}$ 的属性差异度 0.875 小于当前最优解的属性差异度 1.000，$\{r_5,r_6,r_8\}$ 被丢弃.当搜索到 $\{r_5,r_7\}$ 分支时，该分支的差异度上界值 0.875 小于当前最优解的 1.000，此时算法进行剪枝处理.同理，$\{r_6,r_7\}$ 分支被剪切掉.最后 r_7 和 r_8 都无法形成 3-clique，算法结束，输出最优解 $\{r_5,r_6,r_7\}$.如图 2 所示，算法在利用结构性质缩减搜索空间的同时，利用定理 2 的剪枝条件避免了 $\{r_5,r_7\}$ 和 $\{r_6,r_7\}$ 分支的进一步节点访问.

4.3 基于优先次序的 AD-Qclique 算法

算法 AD-Qclique 搜索过程中必须逐个节点遍历来寻找 q-clique，因此具有大差异度的结果越早出现，越能利用剪枝条件缩小搜索空间.基于 best-first 排序思想，遍历时要优先选择可能形成具有较大属性差异度 q-clique 的节点，而最直接的方法是在进行搜索过程之前，重新排列节点序列.

算法 3. 基于优先次序的 AD-Qclique（算法 3 替换算法 2 的步 1～2）.

输入：带权简单无向图 G，不小于 1 的正整数 k，不小于 3 的正整数 q

输出：G 中具有最大属性差异度的无公共节点的 k 个 q-clique 的集合 L_q

1. 依照主键对图 G 中的节点进行排序；

2. 初始化长度为 k 的数组 L_q，$Q\leftarrow\varnothing$，$dvalue(Q)\leftarrow 0$.

基于优先次序的 AD-Qclique 算法如算法 3 所示.算法 3 在算法 2 执行之前对节点访问次序进行排序优化处理.

下面将讨论依照不同主键的优先次序方法.

（1）基于节点度次序.现实世界网络通常具有幂律分布和富人俱乐部性质[17]，即网络中少量的节点具有大量的边并且这些节点倾向于彼此之间互相连接.因此，节点度越大的节点越可能拥有大量的内容属性且能够形成 q-clique 的可能性越大.当基于节点度的优先次序时，若最先加入已有节点集合的节点的度小于 $q-1$ 时，则算法可以直接结束，后续节点不可能形成 q-clique.

（2）基于邻接边权值次序.具有较大属性差异度的 q-clique 一定具有较大的边权值，若节点的邻接边具有较大的边权值，则该节点构成大属性差异度的 q-clique 可能性会更高.由此衍生出 3 种策略：基于节点 1-边权和的优先次序、基于节点 $(q-1)$-边权和的优先次序和基于节点所有邻接边的边权之和的优先次序.

表 1 中列出了本文根据不同主键排序产生的基于优先次序算法的名称.

表 1　基于优先次序算法表

算法名称	排序参照主键
Deg+AD-Qclique	节点的度
Maxw+AD-Qclique	节点的 1-边权和
Sumw+AD-Qclique	节点所有邻接边的边权和
Qmax+AD-Qclique	节点的($q-1$)-边权和

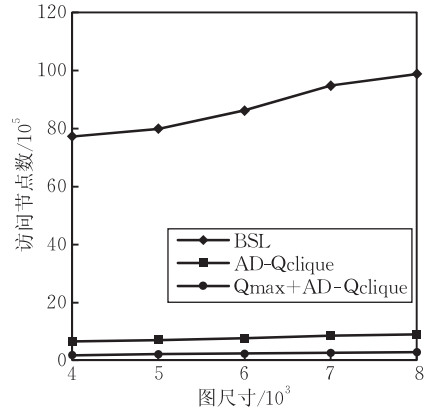
5　实验分析

本文使用多组实验考察算法在不同数据大小和不同参数影响下的性能和结果质量. 本文算法使用 C++语言实现，实验使用一台拥有 2 个 3.00GHz Xeon(R) CPU 和 3.25 GB 的内存，运行 Windows Server 2003 操作系统的服务器.

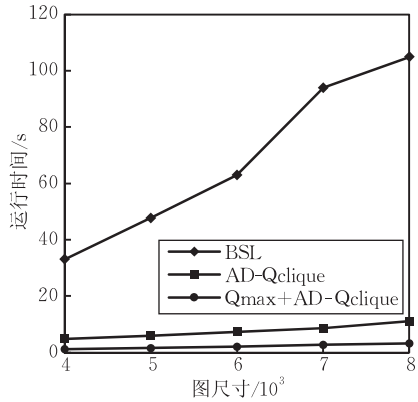
5.1　算法效率分析

本文针对问题相关参数进行了实验分析，对各算法的效率进行了比较.

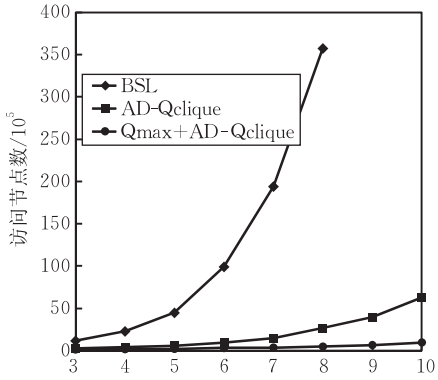
由于 BSL 算法缺乏可伸缩性致使其无法在较大的图上正常运行，因此实验 1 在较小尺寸图上比较算法 AD-Qclique 与算法 BSL 的运行效率. 如图 3(a)所示，算法 AD-Qclique 比 BSL 的访问节点数减少了约 28%，而算法 Qmax+AD-Qclique 则减少了 86%. 同时，图 3(b)中显示了算法的运行时间，与图 3(a)中的节点访问数相一致. 与算法 BSL 相比，算法 Qmax+AD-Qclique 获得了约 37 倍的平均加速比.



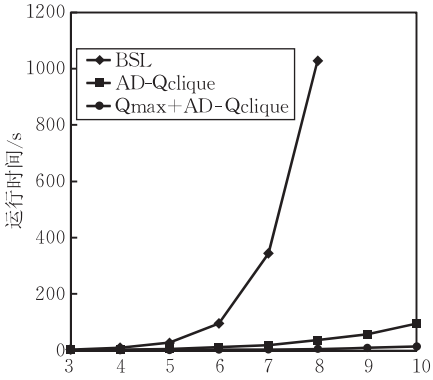
(a) 访问节点数对比(k=10, q=6)



(b) 运行时间对比(k=10, q=6)



(c) 参数q对算法访问节点的影响($|G|$=8000, k=10)

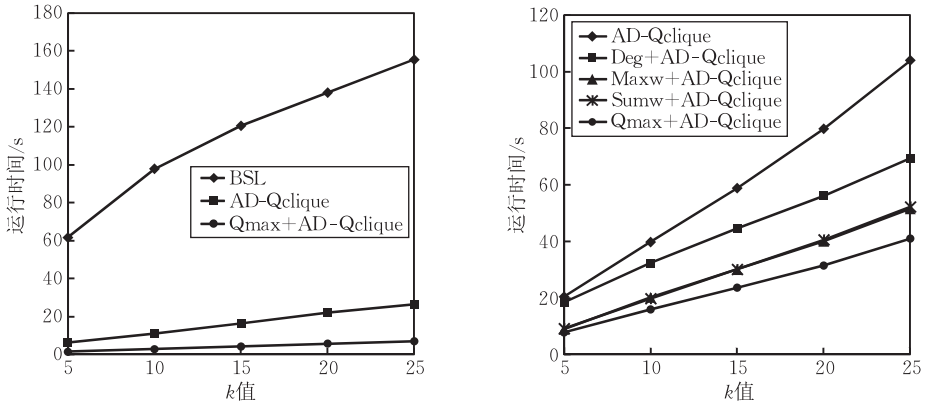


(d) 参数q对算法运行时间的影响($|G|$=8000, k=10)

图 3　算法 AD-Qclique 与算法 BSL 效率对比

实验表明，在利用图的拓扑结构性质的同时利用边权形成的剪切条件能够较大地提高算法的运行效率. 对节点访问次序进行优化处理后，算法的效率进一步提高.

图 3(c)、(d)分别分析了参数 q 对算法访问节点数和运行时间的影响. 与图 3(a)、(b)对比，可以证实，参数 q 固定的 q-clique 问题为多项式时间复杂度. 而当 q 值变化时，q-clique 问题为一个具有 NP 难复杂性的问题. 同时由图 3 可知，算法的访问节点数与算法的运行时间的趋势基本一致. 实验 2 评价了参数 k 对算法运行时间的影响，因为本文中的所有算法皆为迭代地查询 Top-1 最大属性差异 q-clique，因而算法的运行时间会随着 k 值的增大而呈多项式增长. 图 4(a)、(b)则验证了这一结论.

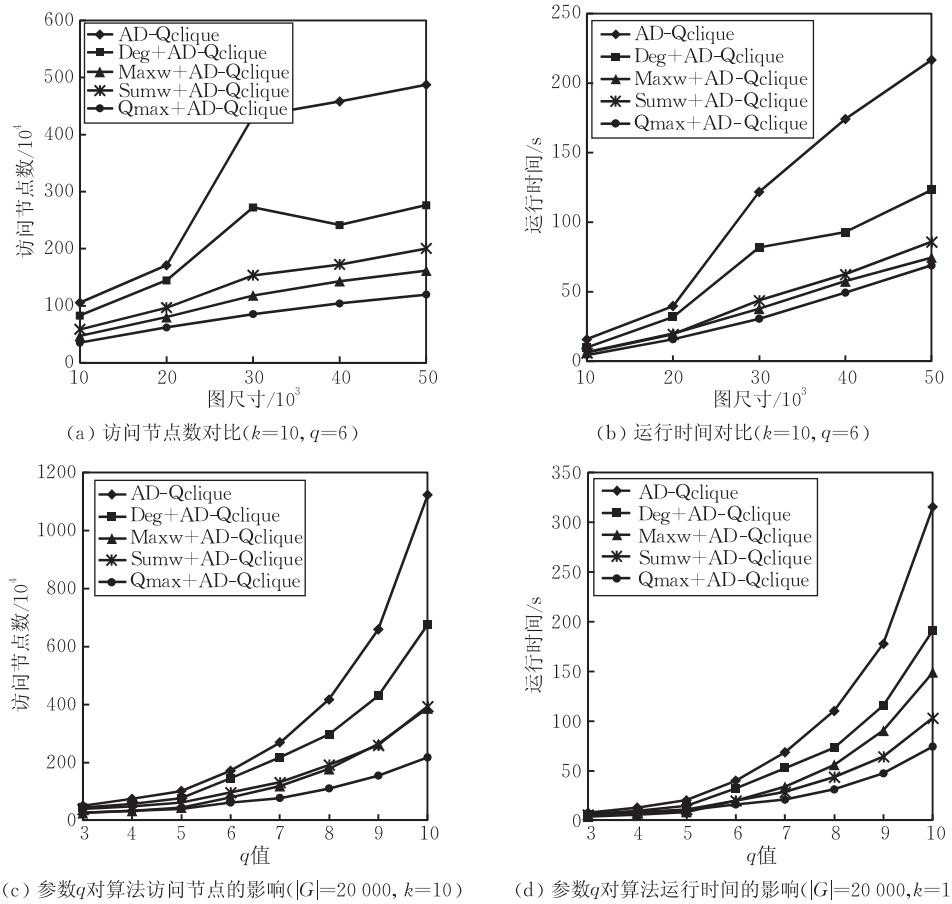


(a) 参数k对算法运行时间的影响($|G|=8000$, $q=6$)　(b) 参数k对算法运行时间的影响($|G|=20\,000$, $q=6$)

图 4　参数 k 对算法 AD-Qclique 和 BSL 的影响

为分析节点访问次序对算法带来的进一步效率提升,实验 3 对大图尺寸下基于不同优先次序的 AD-Qclique 算法的节点访问数和运行时间进行了分析.除算法 Qmax+AD-Qclique 外,其它算法不用每次查询都进行节点排序,因此其它算法的排序开销作为预处理手段而没有算入运行时间中.

如图 5(a)、(b)所示,随着图尺寸的不断增长,算法 AD-Qclique 和算法 Deg+AD-Qclique 的访问节点数和运行时间都表现不稳定.与之相反,算法 Maxw+AD-Qclique、Sumw+AD-Qclique 和 Qmax+AD-Qclique 始终都保持稳定且高效.实验表明,基于邻接边权值次序的优化策略比基于节点度次序的策略效果好.图 5(c)、(d)分别显示了 $|G|=20000$, $k=10$ 时算法的节点访问数和运行时间随 q 值改变时的变化情况.与图 3(c)、(d)的情况相同,随着 q 值的增长,算法节点访问数和运行时间同样呈指数级增长.



(a) 访问节点数对比($k=10$, $q=6$)　(b) 运行时间对比($k=10$, $q=6$)

(c) 参数q对算法访问节点的影响($|G|=20\,000$, $k=10$)　(d) 参数q对算法运行时间的影响($|G|=20\,000$, $k=10$)

图 5　基于优先次序的 AD-Qclique 算法效率对比

最后，图 3、图 4 和图 5 表明，由于算法 Qmax+AD-Qclique 的排序主键与剪切条件最为相近，该算法拥有最高的运行效率.

5.2　算法结果分析

表 2 显示了在 20000 节点的合作关系图中进行 Top-3 属性差异 6-clique 查询的结果并给出了作者的个人研究领域数和科研产出度量值 H-index. 表 2 括号中的值依次为个人研究领域数和 H-index 值.

表 2　Top-3 属性差异 6-clique 查询结果

Rank 1	Rank 2	Rank 3
A. Choudhary(52,37)	J. D. Ullman(43,86)	S. Sahni(46,40)
G. C. Fox(56,30)	H. G. Molina(50,80)	V. Prasanna(49,39)
Ian T. Foster(50,90)	G. Weikum(37,44)	S. K. Das(46,14)
D. Reed(41,35)	J. Hellerstein(39,51)	A. Sussman(28,23)
K. Kennedy(36,63)	M. E. Lesk(22,19)	F. Dehne(33,21)
C. H. Koelbel(11,20)	J. F. Naughton(36,48)	R. J. Leblanc(12,15)

图 6 显示了 20000 节点数据集的作者个人研究领域数和对应作者数的平滑散点图. 图 6 中的分布情况表明真实数据集中只有极少部分人拥有广泛的研究领域. 结合表 2 中查询结果和图 6 中的个人研究领域数分布情况，说明合作关系网络中较大属性差异 q-clique 中节点不仅具有极强的研究水平，同时有着相比于大部分学者更为广泛的研究领域.

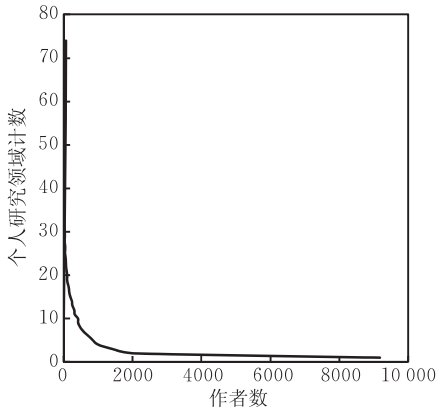


图 6　作者个人领域数分布情况($|G|=20000$)

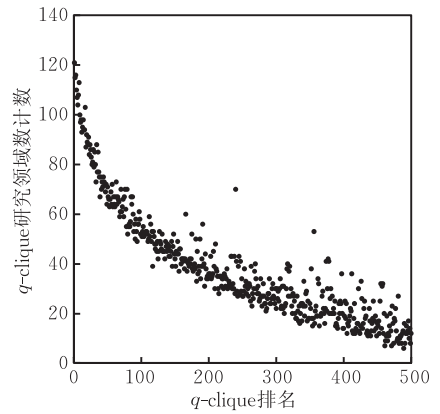


图 7　结果 q-clique 研究领域数分布情况($|G|=20000$, $q=6$)

图 7 显示了在 20 000 节点的合作关系图中 Top-500 属性差异 6-clique 的不同研究领域属性值计数的分布情况. 如图 7 所示，6-clique 的差异度越大则其中包含的不同研究领域属性值越多. 实验验证本文中属性差异度的定义能够满足属性差异 q-clique 问题中对属性内容丰富程度的要求.

6　结　论

本文提出了一种结合拓扑结构和属性内容的紧密子图发现问题——属性差异 q-clique 发现. 为解决图中 q-clique 数目过多的问题，本文提出了一种分支限界算法 AD-Qclique，该算法在搜索的过程中利用边上的节点属性相异度形成剪枝条件，从而较大程度地减少了节点访问数目. 本文还对节点访问次序进行了优化，进一步提高了算法性能. 最后，本文利用 ACM 作者信息数据集，对查询的有效性与算法的效率进行了测试. 实验表明，AD-Qclique 的效率远远高于基本算法 BSL，且属性差异 q-clique 查询返回的作者都具有较高的 H-index 值及广泛的研究领域，表明了属性差异 q-clique 问题具有重要的实际应用价值.

参　考　文　献

[1] Dou Bing-Lin, Li Shu-Song, Zhang Shi-Yong. Social network analysis based on structure. Chinese Journal of Computers, 2012, 35(4): 741-753(in Chinese)
(窦炳琳，李澍淞，张世永. 基于结构的社会网络分析. 计算机学报，2012，35(4)：741-753)

[2] Boginski V, Butenko S, Pardalos P M. Statistical analysis of financial networks. Computational Statistics & Data Analysis, 2005, 48(2): 431-443

[3] Mascia Franco, Cilia Elisa, Brunato Mauro, Passerini Andrea. Predicting structural and functional sites in proteins by searching for maximum-weight cliques//Proceedings of the 24th AAAI Conference on Artificial Intelligence. Atlanta, USA, 2010: 1274-1279

[4] Reddy P K, Kitsuregawa M, Sreekanth P, Rao S Srinivasa. A graph based approach to extract a neighborhood customer community for collaborative filtering//Proceedings of the 2nd International Workshop on Databases in Networked Information Systems(DNIS'02). Aizu, Japan, 2002: 188-200

[5] Leskovec J, Adamic L A, Huberman B A. The dynamics of viral marketing//Proceedings of the 7th ACM Conference on Electronic Commerc. Ann Arbor, USA, 2006: 228-237

[6] Lappas Theodoros, Liu Kun, Terzi Evimaria. Finding a team of experts in social networks//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 467-476

[7] Zhou Yang, Cheng Hong, Yu Jeffrey Xu. Graph clustering

based on structural/attribute similarities. Proceedings of the VLDB Endowment, 2009, 2(1): 718-729

[8] Zhu Linhong, Ng Wee Keong, Cheng James. Structure and attribute index for approximate graph matching in large graphs. Information Systems, 2011, 36(6): 958-972

[9] Cheng James, Ke Y, Fu Ada Wai-Chee. Finding maximal cliques in massive networks by H*-graph//Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data (SIGMOD'11). Indianapolis, USA, 2010: 447-458

[10] Cheng James, Ke James, Chu Shumo, Özsu M Tamer. Efficient core decomposition in massive networks//Proceedings of the 2011 IEEE 27th International Conference on Data Engineering (ICDE'11). Washington, DC, USA, 2011: 51-62

[11] Huang Jianbin, Sun Heli, Han Jiawei, Deng Hongbo, Sun Yizhou, Liu Yaguang. Shrink: A structural clustering algorithm for detecting hierarchical communities in networks// Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'10). Toronto, Canada, 2010: 219-228

[12] Han Meng, Li Jian-Zhong, Zou Zhao-Nian. Finding K close subgraphs in an uncertain graph. Journal of Frontiers of Computer Science and Technology, 2011, 5(9): 791-803(in Chinese)
(韩蒙, 李建中, 邹兆年. 从不确定图中发现 K 紧密子图. 计算机科学与探索, 2011, 5(9): 791-803)

[13] Kargar Mehdi, An Aijun. Keyword search in graphs: Finding r-cliques. Proceedings of the VLDB Endowment, 2011, 4(10): 681-692

[14] Coulter N, French J, Horton T, Mead N, Rada R, Ralston A, Rodkin C, Rous B, Tucker A, Wegner P, Weiss E, Wierzbicki C. Computing classification system 1998: Current status and future maintenance (report of the CCS update committee). ACM Computing Reviews, 1998, 39(1): 1-24

[15] Downey Rod G, Fellows Michael R. Fixed-parameter tractability and completeness II: On completeness for W[1]. Theoretical Computer Science, 1995, 114(1-2): 109-131

[16] Liu Guimei, Wong Limsoon. Effective pruning techniques for mining quasi-cliques//Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases. Antwerp, Belguim, 2008, 2: 33-49

[17] Zhou S, Mondragon R J. The rich-club phenomenon in the Internet topology. IEEE Communication Letters, 2004, 8(3): 180-182

SUN Huan-Liang, born in 1969, professor. His research interests include spatial database and data mining, etc.

LU Zhi, born in 1988, master candidate. His research interests include graph data mining and spatial database, etc.

LIU Jun-Ling, born in 1972, Ph. D. candidate. Her research interests include spatial database and data mining, etc.

YU Ge, born in 1962, professor, Ph. D. supervisor. His main research interests include database theory and technology, distributed and parallel system, etc.

Background

This paper focuses on the research for dense subgraph discovery in large graphs. Dense subgraph discovery is an important aspect of network analysis since density is an indication of importance in almost any network. Recently, many existing dense subgraph discovery methods, such as Ext-MCE, EMcore, SHRINK and so on, have been devised. However, most of them only focus on the topological structure and largely ignore the vertex properties which are often informative. In this paper, the problem of attribute difference q-clique discovery is proposed to find q-clique by combining topological structure and attribute content. An attribute difference q-clique of an attributed graph is a subgraph with q vertices that are completely connected and the dissimilarity between each two adjacent vertices' attribute is as large as possible. This problem aims to find q-cliques with abundant attribute content and nodes contained by a q-clique are difference from each other. This paper converts the problem into finding largest weighted q-clique by calculating the dissimilarity between endpoints' attribute of each edge as the weight of the edge and proposes a branch and bound algorithm AD-Qclique and a novel and efficient node visit order to improve the performance of the algorithm. Finally, the paper conducts extensive experiments to evaluate the performance of AD-Qclique and the results of queries through real authors cited network dataset. Experiment results show that AD-Qclique obtains an order of magnitude speed-up comparing to Baseline Algorithm and the found results show teams with wide field of study and high authority.

This work is supported in part by the Key Program of National Natural Science Foundation of China (61033007), National Natural Science Foundation of China (61070024, 61272179) and Fundamental Research Funds for the Central Universities (N100704001). The foundations focus on the research of various areas of spatial data and graph data. Our group has been working on the research of data mining for many years, and some good papers have been published in worldwide conferences. This paper proposes a novel problem of attribute difference q-clique queries.