网络首发时间：2018-09-21 14:03:32
网络首发地址：http://kns.cnki.net/kcms/detail/11.2223.N.20180920.1619.001.html

ISSN 1000-0054　　清华大学学报（自然科学版）
CN 11-2223/N　　J Tsinghua Univ (Sci & Technol)

知识图谱的 Top-k 摘要模式挖掘方法

罗之皓[1]，　李　劲[1]，　岳　昆[2]，　毛钰源[1]，　刘　琰[1]

（1. 云南大学 软件学院，昆明 650500；2. 云南大学 信息学院，昆明 650500）

摘　要：知识图谱数据具有体量大、内容丰富、类型多样、缺乏统一模式描述特点。提取知识图谱模式信息并形成摘要模式，对于提升知识检索、挖掘质量具有重要研究意义。该文首先给出了摘要模式的判定准则以及摘要模式质量的度量标准，提出了面向知识图谱的 Top-k 摘要模式挖掘问题，并将该问题建模为一个次模函数优化问题；其次，为高效判定摘要模式及度量模式的覆盖质量，提出了基于 Pregel 编程模型的并行化摘要模式判定和质量度量算法；然后，给出了高效求解 Top-k 摘要模式挖掘问题的贪心算法；最后，在真实知识图谱数据上对本文方法进行了验证。实验结果表明：该方法在摘要模式的覆盖度和算法执行效率方面优于已有方法。

关键词：知识图谱；摘要模式挖掘；次模函数；图匹配

中图分类号：TP311.5　　**文献标志码**：A

DOI：10.16511/j.cnki.qhdxxb.2018.26.044

Mining Top-k summarization patterns for knowledge graphs

LUO Zhihao[1], LI Jin[1], YUE Kun[2], MAO Yuyuan[1], LIU Yan[1]

(1. School of Software, Yunnan University, Kunming 650500, China;
2. School of Information, Yunnan University, Kunming 650500, China)

Abstract: Knowledge graph data has large volumes, rich content, diverse types, and lacks a unified model description. Pattern information needs to be extracted from knowledge graphs to improve the quality of knowledge graph retrieval and mining. This paper presents a knowledge graph summarization pattern and quality metrics. This method is used in an algorithm for mining Top-k summarization patterns (Top-k SPM) formulated as a submodular function optimization problem. Then, a Pregel based parallel algorithm is used to validate the algorithm and measure the qualities of summarization patterns. Two efficient greedy algorithms are also presented to solve the Top-k SPM. The efficiency and effectiveness of the method is then verified on real knowledge graph datasets. The tests show that the method outperforms the existing methods in terms of coverage and algorithm execution time.

Key words: knowledge graph; summarization pattern mining; submodular function; graph matching

近年来，在网络信息技术的支撑下，以维基百科、Yago、Freebase 等为代表的包含大量非结构化、异构数据的知识图谱得到了快速发展，并在社交网络，知识检索，生物信息学等领域都有广泛的应用[1-3]。同时，知识图谱数据具有体量庞大、内容丰富、类型多样、动态、无序性强、缺乏统一模式描述等特点[4]。这些特点给用户准确、有效地获取图谱知识带来了巨大的挑战。与传统关系数据相比，知识图谱缺乏统一规范的模式描述。对于用户而言，很难了解、掌握图谱数据包含的模式信息。因此，高效提取知识图谱模式信息，并形成摘要模式（summarization patterns），以此来展示图谱数据信息、并分析不同类型实体之间的相关关系，对于提升知识图谱的知识检索、挖掘质量具有重要研究意义[5-8]。

广义上讲，知识图谱是一种图数据，因此可基于已有的频繁子图模式挖掘算法获得知识图谱的模式信息。然而，直接基于已有的频繁子图模式挖掘算法得到的图谱模式存在以下问题：1）用户很难控制算法的频繁度值，往往产生大量的频繁子图模式；2）模式的复杂程度不易控制；3）不同模式之间往往相互交叠冗余。针对这些问题，Song 等[9]给出了一种新的知识图谱摘要模式挖掘方法。该方法基于已有的图模式挖掘算法得到候选模式集，并将知识图谱模式摘要挖掘建模为一个双目标优化问

收稿日期：2018-07-19

基金项目：国家自然科学基金资助项目(61562091，61472345)；
第二批“云岭学者”培养项目(C6153001)；
云南省应用基础研究计划面上项目(2016FB110)；
云南大学中青年骨干教师培养计划项目；
云南大学青年英才培育计划项目(WX173602)；
云南大学数据驱动的软件工程科技创新团队项目(2017HC012)

作者简介：罗之皓(1993—)，男，硕士研究生。

通信作者：李劲，男，副教授。E-mail：lijin@ynu.edu.cn

题，挖掘知识图谱中典型性强且冗余度低的摘要模式。然而，该方法仍然存在一些不足之处，具体地，首先，该方法目标函数的参数需要反复试验验证进行调优；其次，虽然该方法能够有效确保摘要模式集的非冗余性，模式典型性却有待提升。

针对已有方法的不足之处，本文提出了一种新的知识图谱摘要模式挖掘方法。首先，给出了摘要模式的定义，判定准则以及摘要模式质量的度量标准。提出了基于 Pregel 编程模型的并行化摘要模式判定和质量度量算法。其次，将知识图谱的 Top-k 摘要模式挖掘建模为一个优化问题，证明目标函数满足次模性(submodularity)。给出了 Top-k 摘要模式挖掘贪心算法以及算法加速方法。最后，在真实知识图谱数据上验证了本文提出的模型、算法的优势和有效性。

1 问题定义

给出摘要模式的定义以及摘要模式质量度量标准，进一步定义知识图谱的 Top-k 模式摘要挖掘问题。

$G=\langle V, E, L\rangle$ 为知识图(knowledge graph)，表示实体及其关系。其中：V 表示节点集，$\forall v\in V$ 表示图谱中的一个实体；$E\subseteq V\times V$ 为边集，表示实体之间的关系；L 是标签函数，即 $L(v)$ 是节点 v 的标签(或实体类型)；$L(e)$ 是边 e 的标签，表示 e 连接的 2 个节点之间的关系类型；$P=\langle V_P, E_P\rangle$ 为模式图(pattern graph)，表示标签及标签之间的关系；V_P 是标签集，E_P 是标签关系边集。

与一般图数据不同，知识图谱含有大量实体类型信息，且实体之间的关系多样、复杂。为有效展示知识图谱并提供不同兴趣视角的模式信息，需要对图谱模式的复杂程度进行限制，因此，首先定义受限模式图。

定义 1 受限模式图：给定整数参数 d 和 b。图模式 P 为一个受限模式图，仅当 $d_P\leqslant d$，$|P|\leqslant b$，其中：d_P 为模式 P 中节点之间最长路径长度，$|P|$ 为 P 的节点数和边数的总和。

不是任何受限模式图均为有意义的摘要模式。一个摘要模式 P 是对知识图 G 中节点及节点关系的概括描述，因此，可基于图模式匹配方法来建立受限模式图是否为合法摘要模式的判定准则。一般地，图模式匹配方法有基于图同构的模式匹配[10-11]以及基于图模拟的模式匹配。图同构方法要求模式图和数据图在节点类型和节点拓扑关系严格一致情况下才能判定为匹配，匹配判定是 NP-hard 的，对于大规模知识图谱而言，判定求解困难。图模拟要求模式图和数据图节点类型一致，但不要求节点之间拓扑关系严格保持一致，其判定是多项式可解的，因此，本文基于图模拟匹配来定义合法摘要模式的判定准则，具体地，摘要模式由下面的定义 2 给出。

定义 2 摘要模式：给定知识图 $G=\langle V, E, L\rangle$，模式图 $P=\langle V_P, E_P\rangle$，如果存在二元关系 $R(P, G)\subseteq V_P\times V$，$\forall l\in V_P$ 和每一个知识图节点 $v\in V$ 构成二元关系 $(l, v)\in R$，节点 l' 和 v' 分别为 l 和 v 的子节点。如果对于 P 中所有的边 (l, l')，存在图 G 中对应的边 (v, v')，且 (l, l') 的边标签 (v, v') 的边标签一致，使得二元关系 $(l, v)\in R$，那么图模式 P 为图 G 的一个匹配的摘要。如图 1 所示。

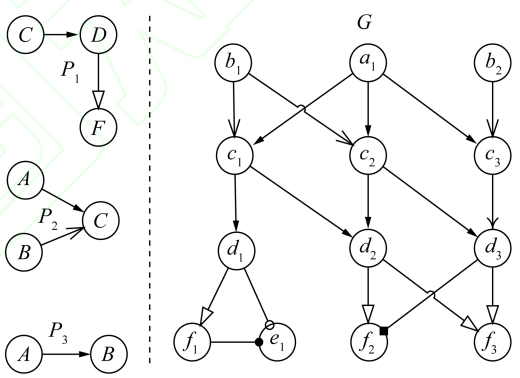


图 1 知识图谱示例

给定一个摘要模式 P，用 $R(P, G)$ 在知识图 G 上决定的子图大小来度量 G_P 的质量，为此，定义摘要模式 P 的覆盖度。

定义 3 摘要模式的覆盖度：给定知识图 G 和模式摘要 P。所有满足二元关系 R 的元组 $(l, v)\in R(P, G)$ 中的实体节点 v 及其连接边构成的 G 的子图记为 G_P，则模式摘要 P 的覆盖度记为 $\mathrm{cov}(G_P, G)$ 并定义如下：

$$\mathrm{cov}(G_P, G)=\frac{|G_P|}{|G|}. \tag{1}$$

其中：$|G_P|$ 表示子图 G_P 边和节点的总数目，$|G|$ 表示图 G 的边和节点数的总和。

例 1 给定参数 $d=2$，$b=5$，图 1 中的 3 个图模式满足受限模式条件。其中只有 P_1、P_2 满足摘要的定义，P_3 中标签为 A 的节点，其子节点的标签为 B，而图 G 中，不存在这样的关系，所以 P_3 不满足摘要的定义。摘要 P_1 根据定义 3，由 P_1 与 G

的二元关系 $R(P_1, G)=\{(C, c_1), (C, c_2), (C, c_3), (D, d_1), (D, d_2), (D, d_3), (F, f_1), (F, f_2), (F, f_3)\}$，得到对应图 G 中节点：c_1、c_2、c_3、d_1、d_2、d_3、f_1、f_2、f_3，并根据这些节点之间的关系构成子图 G_{P_1}，$|G_{P_1}|=18$，$|G|=31$。同理，$|G_{P_2}|=12$，$\mathrm{cov}(G_{P_1}, G)=\frac{18}{31}>\mathrm{cov}(G_{P_2}, G)=\frac{12}{31}$，所以，$P_1$ 的覆盖度高于 P_2。

由定义 3 可知，模式 P 的覆盖度越高，P 对知识图谱数据的概括能力就越强。同时，2 个摘要模式 P_1 和 P_2 的覆盖子图 G_{P_1} 和 G_{P_2} 的节点集、边集也可能存在交集，即摘要模式可能存在覆盖交叠。于是，如何获得概括能力强，相互不冗余的摘要模式集成为需要解决的问题，并将其描述为 Top-k 摘要模式挖掘问题。

定义 4 Top-k 摘要模式挖掘问题（Top-k SPM）：给定常数 k，知识图的 Top-k 摘要模式挖掘问题定义如下：

$$S^* = \arg\max_{|S|=k} F(S) = \arg\max_{|S|=k} \frac{\left|\bigcup_{i=1}^{k} G_{P_i}\right|}{|G|}. \tag{2}$$

其中：$S=\{P_1, P_2, \cdots, P_k\}$ 为含 k 个受限摘要模式组成的集合，G_{P_i} 为摘要模式 P_i 在知识图 G 上的覆盖子图，$|G_{P_i}|$ 为图 G_{P_i} 的节点数和边数的总和。不难看出，Top-k SPM 等价于一个最大 k 覆盖问题，因此是 NP-hard 的。然而，*Top*-k *SPM* 问题的目标函数是满足次模性，这为高效近似求解该问题提供了理论基础。

定义 5 次模函数[12]：给定集合 U，函数 $F: 2^U \to \mathbb{R}^+$ 是 U 上的次模函数，仅当对于任意两个集合 $S \subseteq T \subseteq U$，任意元素 $j \in U \setminus T$，有 $F(S \cup \{j\}) - F(S) \geqslant F(T \cup \{j\}) - F(T)$。

定理 1 Top-k SPM 的目标函数满足次模性。

证明：设 U 为全集，设任意两个集合 S 和 T，有 $S \subseteq T \subseteq U$，元素 $P_j \in E/T$。$|S|=n$，$|T|=m$，$m \geqslant n$。

记

$$\sigma(S \mid P_j) = F(S \cup \{j\}) - F(S),$$
$$\sigma(T \mid P_j) = F(T \cup \{j\}) - F(T),$$

于是有

$$\sigma(S \mid P_j) = \frac{\left|\left(\bigcup_{i=1}^{n} G_{P_i}\right) \cup G_{P_j}\right|}{|G|} - \frac{\left|\bigcup_{i=1}^{n} G_{P_i}\right|}{|G|} = \frac{\left|G_{P_j} - \left(\bigcup_{i=1}^{n} G_{P_i}\right)\right|}{|G|} \geqslant 0。$$

同理，$\sigma(T \mid P_j) = \frac{\left|G_{P_j} - \left(\bigcup_{i=1}^{m} G_{P_i}\right)\right|}{|G|} \geqslant 0$。由此可得：$\sigma(S \mid P_j) \geqslant \sigma(T \mid P_j) \geqslant 0$。综上，$F(S \cup \{P_j\}) - F(S) \geqslant F(T \cup \{P_j\}) - F(T)$，因此，$F(S)$ 满足次模性。

需要说明的是：与文[9]中将摘要模式挖掘问题定义为一个双目标函数优化问题不同，本文将 Top-k SPM 定义为一个次模函数最大化问题。由此，挖掘过程中无需人为指定折中因子。次模函数优化过程中寻求边际效用最大化（marginal utility maximization）的数学性质也决定了结果集由高覆盖度且相互之间非冗余的模式组成，保证了求解质量。虽然 Top-k SPM 是 NP-hard 的，但由文[13]结论可知，贪心法求解 Top-k SPM 可得到保证近似下界 $1-\frac{1}{e}$ 的近似解。

2 Top-k 摘要模式挖掘算法

首先介绍给定知识图 G 的一个受限模式图 P，基于 Pregel 编程模型的判定 P 为合法摘要模式的判定方法，以及求 P 在 G 上的覆盖子图 G_P 的方法。基于此，进一步介绍基于贪心法挖掘 Top-k 摘要模式集，并讨论基于 lazy update 的挖掘算法加速方法。

2.1 摘要模式判定及其覆盖子图求解

给定受限模式 P 和知识图 G，判定 P 为 G 上的合法摘要模式的关键是：确定二元关系 $R(P, G)$ 是否存在，进一步地，如果 P 为 G 上的合法摘要模式，那么，P 在 G 上的覆盖子图 G_P 依赖于求解 $R(P, G)$ 中的所有二元元组。由定义 2 可知，求解 $R(P, G)$ 的过程是一个基于图模拟判定 P 是否匹配 G 的过程。由文[14]可知，判定算法是多项式计算复杂度的，即 $O((|V_P|+|V|)(|E_P|+|E|))$。

首先介绍判定知识图 G 中任意节点 $u \in V$ 匹配的方法。由定义 2 可知，判定节点 $u \in V$ 是否匹配取决于：如果 u 的标签是 $l \in V_p$，且 l 是 p 中的叶子节点，则 u 是匹配的；如果 v 的标签为 l，且在 p 中 l 有子结点 $l'_1, l'_2, \cdots, l'_c$，那么，v 在 G 中具有 $l'_1, l'_2, \cdots, l'_c$ 标签的子节点，且 v 与其 l'_i（$i=1, 2, \cdots, c$）标签子节点之间的边与边 (l, l'_i) 的标签一致，则 v 匹配。

基于节点匹配判定，进一步可判定 P 是否为合法摘要模式，具体地，将 P 中节点进行拓扑排序，并生成一个由 P 决定的标签拓扑序列，记为 l_1，l_2，…，l_t。对知识图 G 中的结点，按照 P 中标签拓扑序列的逆序判定相应标签的节点是否匹配，最终，如果 G 中存在 l_1，l_2，…，l_t 标签的匹配节点，那么，P 被判定为一个合法的摘要模式。在对 P 进行判定过程中，被判定为匹配的节点及与这些节点相邻的具有相应标签的边即构成覆盖子图 G_P。

由于实际应用中的知识图谱含有大量节点和边，判定 P 和求解 G_P 是耗时的。为高效地判定、求解，本文提出了一种基于 Pregel 编程模型[15]的并行化的摘要模式判定及覆盖子图的求解方法。

Pregel 编程模型是一种以节点为中心的并行迭代式图计算模型。基本思想是：整个计算分为多个超步(super step)迭代执行。在每一个超步，图中节点并行地接收、合并其邻居结点发来的消息，根据消息更新节点自身的状态，然后向其邻居节点发送新消息。当图中没有消息传递时，整个 Pregel 过程结束。基于 Pregel 编程模型的摘要模式判定及覆盖子图的求解算法分为 2 个步骤：

步骤 1 初始化：首先，为记录 G 中节点的匹配状态，以及记录覆盖子图信息，对于 G 中每个节点 v 设置 2 个变量，分别为：匹配状态 m_v，覆盖节点集 C_v。m_v 是个 Bool 变量(T 为 true，F 为 false)，m_v=true，判定节点 v 匹配。C_v 是集合，用来存储与节点 v 相关的已匹配的节点集。由定义 2 可知，G 中标签为 P 中叶子标签的所有节点均为已匹配节点，例如图 2 中，对于给定的 P 来说，G 中的 f_1、f_2、f_3 均为匹配节点，于是有 $m_{f_1}=T$，$m_{f_2}=T$，$m_{f_3}=T$，以及 $C_{f_1}=\{f_1\}$，$C_{f_2}=\{f_2\}$等等。其他 v 是否匹配取决于其是否具有已匹配的相应标签的子节点，例如图 2 中，c_1 是否匹配取决于 d_1 或者 d_2 是否匹配。于是，给定 P，可初始化 m_v，例如，$m_{c_1}=m_{d_1}\vee m_{d_2}$(即结点 c_1 匹配，仅当其任意一个 D 标签子节点匹配)，$m_{d_1}=m_{f_1}$等. 由于 c_1、d_1 等非叶子标签节点在 Pregel 消息传递前无法判定是否匹配，因此 $C_{c_1}=\varnothing$，$C_{d_1}=\varnothing$ 等。

步骤 2 Pregel 消息传递：Pregel 过程的消息从已判定为匹配的叶子标签节点开始。在图 2 中，从 f_1、f_2、f_3 开始。每个已匹配的节点根据 P 中规定的边标签向其父节点发消息，消息内容为：节点 ID，节点标签，例如，f_1 向 d_1 发送消息 msg(f_1，F)。需要注意的 f_2 不会向 d_3 发送消息，因为边

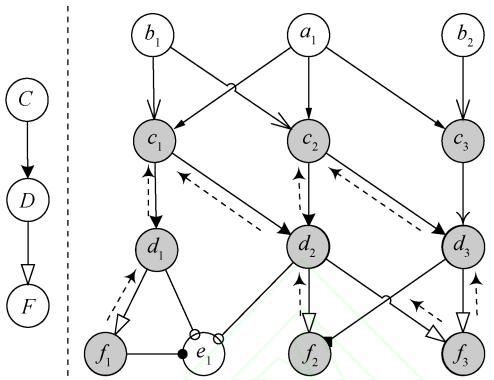


图 2 基于 Pregel 的摘要模型判定及其覆盖子图求解

(f_2，d_3)与 P 中 F 到 D 的边标签不同。接收到消息的节点将消息进行合并，例如 d_2 同时接收到 msg(f_2，F)，msg(f_3，F)，将它们按标签合并为 msg($\{f_2$，$f_3\}$，F)，进而调用节点消息处理函数 vprog 处理合并后的消息，具体地，根据消息更新匹配判定逻辑表达式，例如，对于 d_2 节点，其逻辑表达式为：$m_{d_2}=m_{f_2}\vee m_{f_3}$，由接收到的消息可知 f_2、f_3 已匹配，因此，更新逻辑表达式为：$m_{d_2}=T\vee T$，于是，d_2 判定为匹配。图 2 中的 Pregel 过程经历 2 个超步(在图中用不同有向线段表示)完成消息传递。最终，c_1、c_2 都是被判定为匹配 C 的标签点，因此模式 P 为一个合法的摘要模式，同时，其覆盖子图由所有灰色节点及相应边组成。

2.2 挖掘算法

Top-k 摘要模式挖掘算法的基本思想是：首先，基于频繁子图模式挖掘算法，例如 GRAMI[16]，得到模式子图，并选择满足 $d_P\leqslant d$，$|P|\leqslant b_P$ 条件的子图作为受限模式图；其次，对每个受限模式图 P，采用节 2.1 描述的 Pregel 过程判定 P，且如果 P 为合法模式，则同时求解其覆盖子图 G_P；再次，根据式(2)中目标函数 $F(S)$，每次选取 $F(S)$最大边际效用的 P 加入 S，直到 $|S|=k$；最后，算法输出 S 作为 Top-k 摘要模式挖掘结果。图 3 中算法 1 给出了 Top-k 摘要模式挖掘算法的完整描述。

算法 1 的时间复杂度分析如下：基于 Pregel 模型判定 P，求解其覆盖子图 G_P，时间复杂度为 $O((|V|+|V_P|)(|E|+|E_P|))$[14]。贪心法求解摘要模式集 S，时间复杂度为 $O(|P|)$。所以算法 1 的时间复杂度为

$$O(|P|(|V|+|V_P|)(|E|+|E_P|)+|P|).$$

算法 1 subTopk（Top-k 摘要模式挖掘算法）
输入：图 $G=(V, E)$，k，最大路径长度 d，图模式大小参数 b_p，图 G 的频繁模式子图集 P，结果集 $S\leftarrow\varnothing$；
输出：摘要集 S；
(1) 基于 GRAMI 做频繁子图挖掘，得到频繁子图集合 FG；
(2) 按照 $d_p\leqslant d$，$|P|\leqslant b_p$ 条件，筛选 FG 得到受限模式图集合 RP；
(3) 调用 Pregel，判定 $\forall P\in$ RP 是否为合法的摘要模式，如果 P 合法，则添加到摘要模式集 PSet 中，即：
PSet$\leftarrow$PSet$\cup\{P\}$，并生成其对应的覆盖子图 G_P；
(4) 初始化解集 $S\leftarrow\varnothing$；
(5) 选择具有最大边际效用模式 P^*，并加入解集：
$P^*\leftarrow\arg\max_{p\in \mathrm{PSet}}\sigma(S|P)$；$S\leftarrow S\cup\{P^*\}$；
其中：$\sigma(S|P)=F(S\cup\{P\})-F(S)$，
$F(S\cup\{P\})=(|V_{G_S\cup G_P}|+|E_{G_S\cup G_P}|)/|G|$，
$F(S)=(|V_{G_S}|+|E_{G_S}|)/|G|$。
(6) 输出摘要集 S，$|S|=k$。

图 3 算法 1

贪心算法执行过程中，每次迭代都要进行其余模式对于当前解集的边际效用的计算。候选模式集大时，极为耗时。得益于次模函数的次模性，可使用 lazy update[12] 策略对贪心算法进行优化。本文节 3 实验结果表明：该策略可有效提升算法执行效率。

基于 lazy update 策略的 Top-k 摘要模式挖掘算法的基本思想是：给定摘要提取的数目 k、图 G 和基于图 G 的频繁模式子图挖掘结果集 P，首先对 P 中的所有模式子图做筛选，满足约束性图模式条件的图模式进行模式匹配，不满足约束性图模式条件的则舍去。并分别计算剩余模式子图对应的次模函数边际效益值，按照函数值做非递增排序，函数值最大对应的摘要模式图 P 添加到摘要集 S 中；在选第二元素及其以后的过程中，首先更新 F 值最大的模式子图的函数值，若更新之后该子图对应的函数 F 仍然最大，那么该元素即为当前最优值，直接添加到集合 S 中，如果 F 不是最大值，那么就按照贪心算法重复此部分的计算并排序；按照该步骤进行计算，直至 S 的元素个数达到 k 个。算法如图 4 所示。

算法 2 是算法 1 的加速算法，最坏的情况和算法 2 的时间复杂度一样，为 $O(|P|(|V|+|V_P|)(|E|+|E_P|)+|P|)$；而最好的情况是，每次选取元素的时候只用计算一次，所以最优情况下的时间复杂度为 $O(|P|(|V|+|V_P|)(|E|+|E_P|)+k)$。

算法 2 luTopk（基于 lazy update 的 Top-k 摘要模式挖掘算法）
输入：图 $G=(V, E)$，k，最大路径长度 d，图模式大小参数 b_p，图 G 的频繁模式子图集 P，结果集 $S\leftarrow\varnothing$；
输出：摘要集 S；
(1) 基于 GRAMI 做频繁子图挖掘，得到频繁子图集合 FG；
(2) 按照 $d_p\leqslant d$，$|P|\leqslant b_p$ 条件，筛选 FG 得到受限模式图集合 RP；
(3) 调用 Pregel，判定 $\forall P\in$ RP 是否为合法的摘要模式，如果 P 合法，则添加到摘要模式集 PSet 中，即：
PSet$\leftarrow$PSet$\cup\{P\}$，并生成其对应的覆盖子图 G_P；
(4) 初始化解集 $S\leftarrow\varnothing$；
(5) 根据每个摘要模式的边际效益大小做非递增排序，记更新前的摘要集为 S'，更新后的摘要集为 S，若 $\sigma(S|P_1)>\sigma(S|P_2)$，更新 S，直接将 P_1 添加到摘要集 S 中。否则按照当前摘要集 S'，选择具有最大边际效用模式 P^*，并加入解集，具体的：
if $(\sigma(S|P_1)>\sigma(S|P_2))$
$\{S\leftarrow S\cup\{P_1\}$；RP$\leftarrow$RP$-\{P_1\}$；$\}$
else $\{$RP$\leftarrow$sort(RP)；$P^*\leftarrow\arg\max_{P\in \mathrm{PSet}}\sigma(S|P)$；$S\leftarrow S\cup\{P^*\}$；$\}$
(6) 输出摘要集 S，$|S|=k$。

图 4 算法 2

3 实验方法与结果

3.1 实验环境设置

1）实验数据集。

利用 3 个真实的图数据集上测试了本文的算法，其中 Caida 是自治系统商业关系网络，Yago 是开源的知识图库，Stanford 是斯坦福大学提供的网页关系网络。Caida：节点数为 26 475，边数为 106 762，节点标签有 40 类，边标签有 4 类。Yago：采用 Yago 源数据集的一个子集作为实验数据，节点集文本为 1.32 MB(100 000 个节点)，边表文本为 5.34 MB(348 089 条边)，节点标签为 6 753 类，边标签为 37 类。Stanford：节点集文本为 3.08 MB(281 903 个节点)，边表文本为 35.7 MB(2 312 498 条边)，节点标签为 11 478 类，边标签为 120 类。

2）度量标准。

本文提出的方法旨在利用较少的数据表达更多原数据的内容，故使用覆盖度作为覆盖性的度量标准，由于图数据量比较庞大时，图模式挖掘需要消耗大量计算机内存和挖掘时间，为实验便携性，采用挖掘部分图模式进行摘要选取，定义如下：

$$\mathrm{cov}(G_S, G_{P_{\mathrm{sum}}}) = \frac{|G_S|}{|G_{P_{\mathrm{sum}}}|}. \quad (3)$$

其中：S 表示算法得到的解集，是模式图的一个集合；$G_S = \bigcup_{s_i \in S}^{k} G_{s_i}$ 表示 S 中包含的所有摘要模式的覆盖范围，是所有模式图匹配到的子图总和。挖掘得到的所有图模式，对图 G 具有一个最大覆盖数 $G_{P_{\mathrm{sum}}} = \bigcup_{p_i \in P}^{m} G_{p_i}$，$G_{P_{\mathrm{sum}}}$ 表示整个挖掘的图模式集 P 中所有图模式匹配到的子图的并集。由此可以看出，cov 越大，解集的覆盖性就越好，cov 最大值为 1。

3) 实验环境

本文提出的算法均采用 scala 语言实现，所有实验在一台 16 GB 内存、8 核(8 个 Intel 3.4 GHz CPU)的实验室 PC 机上完成。Apache Spark 采用 2.0 版本。

3.2 实验内容

如表 1 所示，频繁度为图模式的挖掘标准，使用 GRAMI 挖掘工具根据该图模式挖掘得到相应的数目的图模式。模式图最大值和最小值(是单个图模式边数和节点数总和的最大值或最小值)用于衡量一个图模式的大小，挖掘得到的所有图模式对图 G 做图模式匹配得到的所有子图，存在一个对原图数据的最大覆盖，即式(3)中的 $G_{P_{\mathrm{sum}}}$，最大覆盖数为 $|G_{P_{\mathrm{sum}}}|$。

表 1 图模式挖掘的参数

数据集	频繁度	模式图数目	模式图最大值	模式图最小值	最大覆盖
Caida	100	692	15	3	4 698
Yago	38	872	11	3	10 468
Stanford	400	542	7	3	17 952

使用 GRAMI 进行频繁模式子图挖掘，整个过程耗时，挖掘结果也需要消耗计算机的大量内存，通过多次调整频繁度，保证了避免内存溢出的情况下使得挖掘数目一定。数据集 Caida 频繁度取 100，受限模式参数 $d=2$，$b=15$，满足首先模式条件的图模式共 692 个。数据集 Yago 中的频繁度取 38，受限模式参数 $d=2$，$b=11$，满足受限模式条件的图模式共 872 个。Stanford 频繁度取 400，受限模式参数 $d=2$，$b=7$，满足受限模式条件的图模式共 542 个。

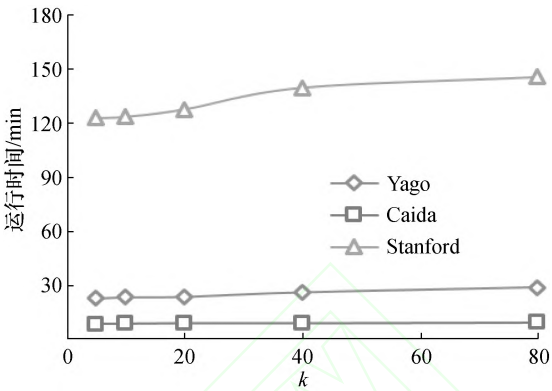


图 5 subTopk 算法在 3 个数据集上的运行时间

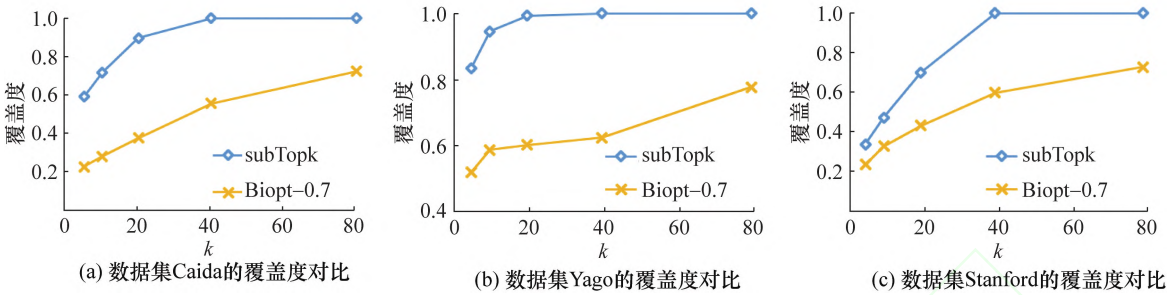
图 5 给出了 3 个数据集摘要选取的运行时间。运行时间包含了所有图模式的匹配以及算法进行摘要选取 2 部分。随着摘要模式数目的增多，时间消耗越大。在整个算法运行中，图模式匹配消耗了大量的时间，例如数据集 Stanford 中，进行图模式匹配的时间开销约为 120 min。随着摘要选取数目的增加，图模式匹配消耗的时间不变，只增加了图模式选取所消耗的时间，故而时间增幅比较小。由此可以看出，传统的频繁模式子图挖掘得到的结果存在大量冗余图模式，通过本文算法筛选，能够挖掘到覆盖能力更强的摘要模式。

3.3 算法实验对比

1) 双目标函数与 subTopk 的实验对比。

采用文[9]中的双目标优化函数(BiOpt)做实验对比，从覆盖度和时间消耗两方面进行对比。BiOpt 中，α 的值需要用户自定义，α 的值取值依次为 0.3、0.5、0.7，经过实验对比，α 取值为 0.7 的时候效果最好。

图 6 中给出了 3 个数据集使用 subTopk 算法进行摘要选取的覆盖结果。图 6a、6b、6c 分别是数据集 Caida、Yago 和 Stanford 的实验结果。由表 1 可知，挖掘得到的摘要模式集对应一个最大覆盖数目，采用这个最大数目来进行覆盖度的衡量。例如图 6b 是在 Yago 数据上运行的结果，最大覆盖数目为 10 468，也就是说挖掘出的 872 个图模式经过摘要判定后得到一系列子图，将这些子图做并，得到一个最大子图，对应节点数和边数的总和为10 468，根据这个最大覆盖数，来进行摘要覆盖度的计算。当 $k=5$ 时，top-5 的摘要模式集的覆盖数目为8 714，覆盖度为 0.832 4，当 $k=40$ 时，



(a) 数据集Caida的覆盖度对比 (b) 数据集Yago的覆盖度对比 (c) 数据集Stanford的覆盖度对比

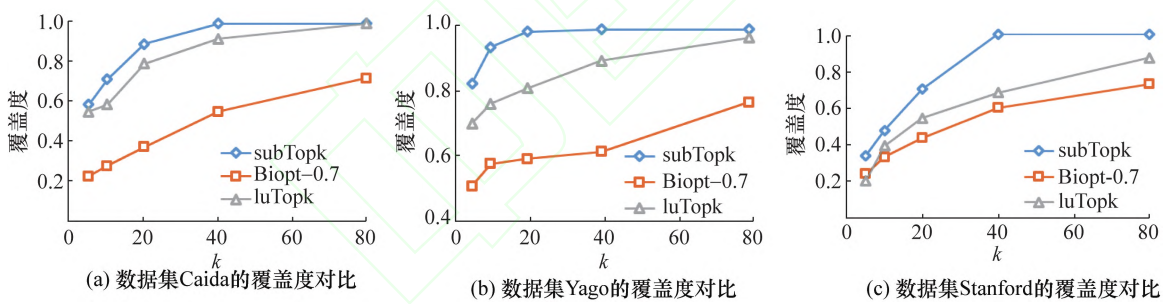
图 6 subTopk 与 BiOpt 覆盖度对比

top-40 的摘要模式集覆盖数目达到最大覆盖，所以覆盖度为 1。覆盖度计算如式(3)所示。BiOpt 折中系数 α 取 0.7 进行实验对比。从实验结果中可以得到，使用 subTopk 选取的摘要模式图在覆盖范围方面均优于 BiOpt；而在在选取摘要数目较少的时候，覆盖范围的优势比较明显。

2）luTopk 算法与 subTopk 和 BiOpt 算法的对比。

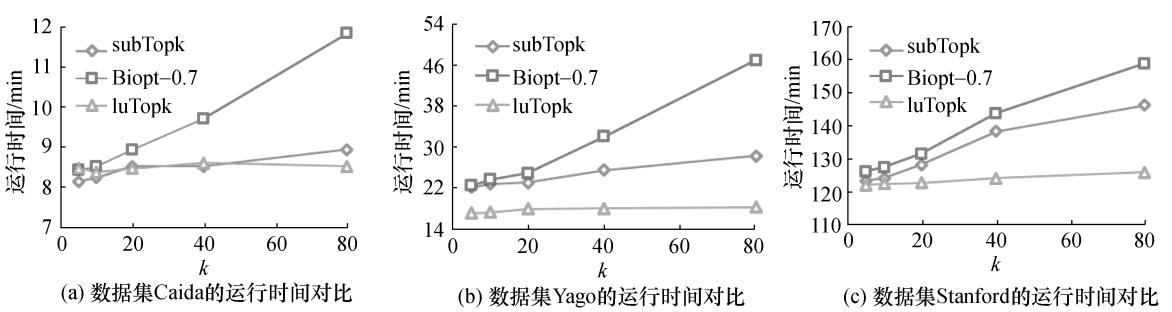
采用 luTopk 的方式优化传统的贪心算法，图 7 给出了 luTopk 与 subTopk、BiOpt 选取的覆盖性对比。图 7a、7b、7c 分别是数据集 Caida、Yago 和 Stanford 的实验结果。同样的，采用最大覆盖数目来进行覆盖度的计算。取 $\alpha=0.7$ 的 BiOpt 作为实验对比。luTopk 的覆盖性比 subTopk 的覆盖性弱一些，但优于 BiOpt。同样的，随着摘要数目的增多，覆盖性逐渐接近于最大覆盖。

时间方面如图 8 所示。luTopk 的时间开销最少，并且随着摘要选取数目的增多，luTopk 的运算时间小幅增加，可以看到，luTopk 在摘要选取数目



(a) 数据集Caida的覆盖度对比 (b) 数据集Yago的覆盖度对比 (c) 数据集Stanford的覆盖度对比

图 7 luTopk 与 subTopk 和 BiOpt 的覆盖度对比覆盖性对比



(a) 数据集Caida的运行时间对比 (b) 数据集Yago的运行时间对比 (c) 数据集Stanford的运行时间对比

图 8 luTopk 与 subTopk 和 BioOpt 的运行时间对比

较多的时候，加速效果明显。

总的来说，subTop-k SPM 覆盖性和时间开销 2 个方面都明显优于 BiOpt；在牺牲一定覆盖性的情况下，luTopk 能够获得更少的时间开销。

3.4 Top-k SPM 实际案例

以 Yago 数据集为例，分别采用 subTopk、BiOpt 算法挖掘 top-2 的摘要模式集，结果如图 9 所示。使用 subTopk 挖掘的摘要模式集如 $\{P_1, P_2\}$所示，其的覆盖度为 0.696 3；使用 BiOpt 挖掘的摘要模式集如 $\{P_3, P_4\}$所示，其覆盖度为 0.379 7。圈节点中的数字表示节点 ID，旁边的数字是该节点的标签 ID。

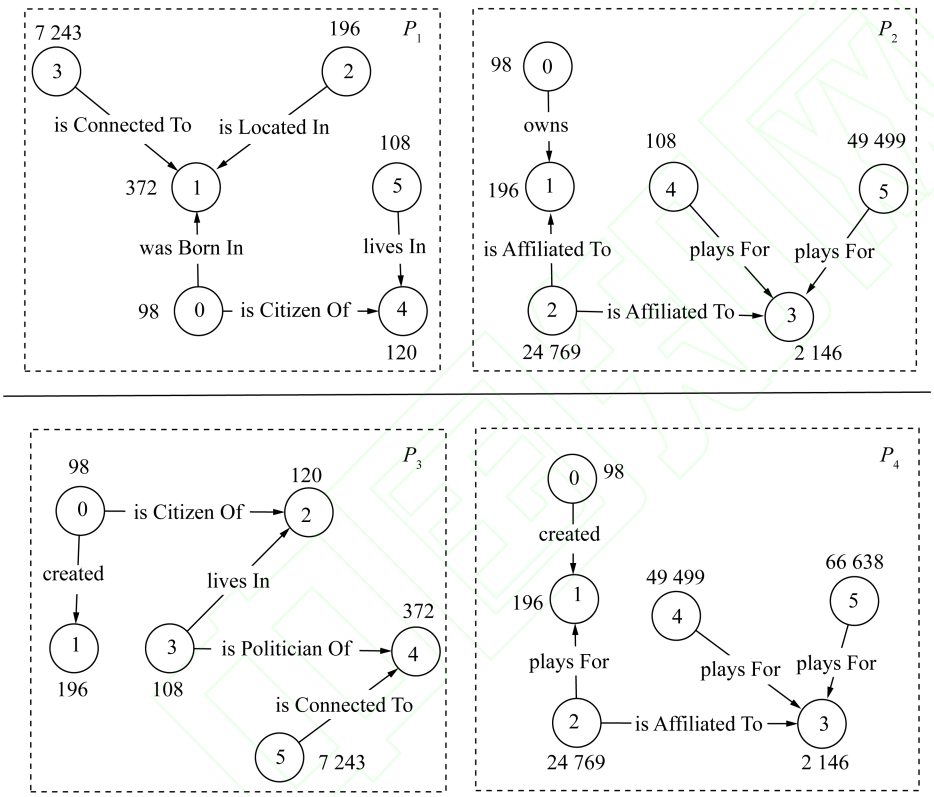


图 9 在 Yago 数据集上的 Top-k SPM 实际案例

4 结 论

高效提取知识图谱模式信息，并形成摘要模式，对于提升知识图谱的知识检索、挖掘质量具有重要研究意义。本文提出了一种新的知识图谱摘要模式挖掘方法。该方法将知识图谱的摘要模式挖掘问题建模为一个次模函数优化问题，并基于 Pregel 编程模型进行摘要模式判定和质量度量，进而基于贪心法进行 Top-k 摘要模式挖掘。在真实的知识图谱数据上将本文的方法与已有方法进行了实验对比，从模式摘要集质量、算法运行时间两方面验证了本文方法的优势和有效性。

基于本文工作，将来拟在以下 2 个方面继续开展研究工作：带标签层次信息的摘要模式挖掘和针对动态知识图谱的摘要模式挖掘均是需要研究的课题。

参考文献 (References)

[1] QIAN J W, LI X Y, ZHANG C H, et al. Social network de-anonymization and privacy inference with knowledge graph model [J]. IEEE Transactions on Dependable and Secure Computing, 2017. DOI: 10.1109/TDSC.2017.2697854.

[2] SHI B X, WENINGER T. Discriminative predicate path mining for fact checking in knowledge graphs [J]. Knowledge-Based Systems, 2016, 104:123-133.

[3] SHI L X, LI S J, YANG X R, et al. Semantic health knowledge graph: Semantic integration of heterogeneous medical knowledge and services [J]. BioMed Research International, 2017, 2858423.

[4] 王萍. 网络环境下的领域知识挖掘 [D]. 上海：华东师范大学，2010.
WANG P. Domain knowledge mining in network environments [D]. Shanghai: East China Normal University, 2010. (in Chinese)

[5] 陈池，王宇鹏，李超，等. 面向在线教育领域的大数据研究及应用 [J]. 计算机研究与发展，2014，51(S1)：67－74.
CHEN C, WANG Y P, LI C, et al. The research and application of big data in the field of online education [J]. Journal of Computer Research and Development, 2014, 51(S1): 67－74. (in Chinese)

[6] SANG S T, YANG Z Z, WANG L, et al. SemaTyP: A knowledge graph based literature mining method for drug discovery [J]. BMC Bioinformatics, 2018, 19(1): 193－193.

[7] KEMMAR A, LEBBAH Y, LOUDNI S. Interval graph mining [J]. International Journal of Data Mining, Modelling and Management, 2018, 10(1): 1－22.

[8] 高俊平，张晖，赵旭剑，等. 面向维基百科的领域知识演化关系抽取 [J]. 计算机学报，2016，39(10)：2088－2101.
GAO J P, ZHANG H, ZHAO X J, et al. Evolutionary relation extraction for domain knowledge in Wikipedia [J]. Chinese Journal of Computers, 2016, 39(10): 2088－2101. (in Chinese)

[9] SONG Q, WU Y H, DONG X L. Mining summaries for knowledge graph search [C]//Proceedings of 2016 IEEE International Conference on Data Mining. Barcelona, Spain: IEEE, 2016: 1215－1220.

[10] BABAI L. Graph isomorphism in quasipolynomial time [extended abstract] [C]//Proceedings of the 48th Annual ACM Symposium on Theory of Computing. Cambridge, USA: ACM, 2016: 684－697.

[11] SAMSI S, GADEPALLY V, HURLEY M, et al. Static graph challenge: Subgraph isomorphism [C]//Proceedings of 2017 IEEE High Performance Extreme Computing Conference. Waltham, USA: IEEE, 2017: 1－6.

[12] KRAUSE A, GOLOVIN D. Submodular function maximization [M]//BORDEAUX L, HAMADI Y, KOHLI P. Tractability. Cambridge: Cambridge University Press, 2014: 71－104.

[13] DVOŘÁK W, HENZINGER M, WILLIAMSON D P. Maximizing a Submodular function with viability constraints [J]. Algorithmica, 2017, 77(1): 152－172.

[14] MA S, CAO Y, FAN W F, et al. Capturing topology in graph pattern matching [J]. Proceedings of the VLDB Endowment, 2011, 5(4): 310－321.

[15] MALEWICZ G, AUSTERN M H, BIK A J C, et al. Pregel: A system for large-scale graph processing [C]//Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data. Indianapolis, USA: ACM, 2010: 135－146.

[16] ELSEIDY M E, ABDELHAMID P, SKIADOPOULOS S, et al. GraMi: Frequent subgraph and pattern mining in a single large graph [J]. Proceedings of the VLDB Endowment, 2014, 7(7): 517－528.

（责任编辑 金延秋）