大规模演化知识网络中的关系推断

赵泽亚[1,2] 贾岩涛[1] 王元卓[1] 靳小龙[1] 程学旗[1]

[1]中国科学院计算技术研究所中国科学院网络数据科学和与技术重点实验室，北京 100190
[2]解放军信息工程大学，郑州 450000
(zhaozeya@software.ict.ac.cn)

Link Inference in Large Scale Evolutionable Knowledge Network

Zhao Zeya[1,2], Jia Yantao[1], Wang Yuanzhuo[1], Jin Xiaolong[1], and Cheng Xueqi[1]
[1](Key Lab of Network Data Science & Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190)
[2](The PLA Information Engineering University, ZhengZhou, 450000)

Abstract In the era of network big data, the spatiotemporal information of knowledge is richly stored in knowledge networks. Traditional knowledge network representation models are mostly blind to both the spatial and temporal information of vertices and links in the network. And it has been verified that considering the spatial or the temporal information can promote the performance of link inference in knowledge networks. In this paper, we propose the evolutionable knowledge network model which is a heterogeneous knowledge network with vertices and edges anchored in both time and space dimensions. Then based on the model, we further study the link inference problem on evolutionable knowledge networks and propose a knapsack constrained link inference method. Experiments on real data sets suggest the effectiveness and scalability of our proposed method over large-scale networks.
Key words link inference; Evolutionable knowledge network; knapsack problem; link extendable pattern

摘要 网络大数据时代的到来使得知识网络中时空信息越来越丰富. 现有的知识网络描述模型对知识的时空信息的刻画不足. 研究证明，利用网络中知识的时空信息以及相关性，能够提高网络中知识间的关系推断的准确率. 针对以上问题，我们首先提出了一种包含时空信息的演化知识网络表示模型，然后研究在该网络模型上的关系推断问题，提出了一种基于背包问题的知识间关系推断方法. 在多个数据集上的实验证明了所提出的关系推断方法的有效性以及对大规模知识网络的适应性.

关键词 关系推断；演化知识网络；背包问题；链接延展模式
中图法分类号 TP182

网络大数据时代，数据不再仅仅是简单的采集对象，其背后其实蕴含着非常丰富、复杂、关联的知识. 当前网络数据是广泛可用的，所缺乏的只是从中提取知识的能力. 有效利用网络大数据价值的主要任务不仅仅是是获取越来越多的数据，也需要从已有的数据中挖掘更多有用的知识[1]，构建成知识库，便于对知识更充分的利用，因此基于网络的大规模知识库的构建是最近流行的一个研究方向，现有的大规模知识库有 YAGO[2,3]，DBpedia[4]，Probase[5]等.

基于大规模知识库的关系推断是从海量信息中挖掘知识实现知识库增长的有效手段之一[6]，其主要目的是利用已有的大规模知识网络，推断或者预测知识网络中隐含的关系. 目前，关系推断已经在个性化推荐、社区发现、知识问答等方面得到广泛应用[7].

现有的关于知识网络中的关系推断的研究，采用的方法主要有有监督学习、半监督学习以及无监

收稿日期：年-月-日*投稿时不填写此项*；最终修改稿收到日期：年-月-日 *投稿时不填写此项.
基金项目：国家973项目课题 (No. 2014CB340405, 2013CB329602)；国家自然科学基金(No. 61173008, 61232010, 61303244, 61402442)；北京市科技新星计划(No.Z121101002512063)；国家科技支撑计划(No. 2012BAH39B04).

督学习等.目前的研究更多的是基于异构信息网络的关系推断，这里的异构信息网络包含多种不同类型的实体与关系，例如，人物、地点、组织机构、电影、论文等，以及它们之间可能产生的各种类型的关系. 现实中典型的异构信息网络有计算机科学文献网络 DBLP 和互联网电影资料库 IMDB .

研究证明，在含有时间信息的异构网络中进行关系推断时，考虑时间信息得到的推断结果比未考虑时间信息得到的结果更好，例如文献[8,9]，在文献[8]中，考虑时间信息进行推断的正确率比未考虑时间信息的正确率高 10%. 同样的，由相关研究工作[10,11]证实，加入空间信息会对异构信息网络上的关系推断带来更大的提升. 例如，文献[12]已证明，融合了空间信息的关系推断可以获得更好的推断结果，但是在文献[12]中的研究，仅仅考虑了一种类型实体间的关系推断，并非异构信息网络. 目前，基于异构信息网络且对网络中的时空信息加以利用进行关系推断的相关研究还很少.

针对知识网络中时空信息的不断丰富，而现有的一些知识网络模型无法很好的刻画这些信息的问题，我们首先提出一个融合时间与空间信息的演化知识网络表示模型. 基于演化知识网络提出了一种新的关系推断方法. 由于知识网络中的关系推断是知识挖掘的重要手段，而在知识挖掘中我们最关注的无疑是推断结果的正确性，因此我们提出的新的关系推断方法旨在提高关系推断的正确率及对大规模数据的适应性. 总结起来，本文贡献可归纳为以下两点：

（1）本文提出了一个演化知识网络表示模型，将知识的时空信息融入到整个知识网络中，为知识的演化和计算提供更多的信息.

（2）本文研究了基于演化知识网络的关系推断方法. 具体的讲，提出一种基于混合背包问题的关系推断方法 KP-LIM，提高关系推断的正确率和推断效率.

实验证明，与当前流行的关系推断方法相比，我们提出的关系推断方法得到了更好的推断效果，在正确率上有 8%-37%的提高，且在千万规模的数据集上的实验证明，我们的方法依然有效.

下面我们详细介绍一下关系推断的相关研究工作。目前主流的关系推断方法是运用机器学习的算法进行关系推断，他们基本上可被分为两类，有监督学习方法[13-17]和无监督学习方法[9,18,19,20]. 其中工作[12]是有监督学习方法经典代表，它将关系推断问题当成一个分类问题，利用经典的逻辑回归方法训练模型实现关系推断. 尽管有监督学习的方法比较流行，但是他们也存在许多弊端，例如训练复杂度高、平衡性较差、难以选择合适的特征等. 相反，无监督的方法不需要关于数据分布的先验知识，避免了有监督学习的训练复杂度高等问题，对于大规模数据具有更强的适应性. 无监督的方法主要是通过定义一些指标来刻画网络中实体间的相似度来实现关系推断，例如文献[9,20]，其中文献[8]是近期无监督关系推断的典型代表，它以经典的共同邻居(CN)方法为基础，加入节点连通性、边连通性以及部分时间信息等信息进行关系推断，但该方法只利用了网络中的局部信息. 我们的提出的推断方法 KP-LIM 也是一种无监督学习方法，该方法定义了一个拓扑特征，链接延展模式(LE 模式)，将全图的结构特征以及网络中的时空间信息融入到背包问题的参数中，利用背包问题的求解对 LE 模式进行选择，再利用选出的模式实现关系推断.

另一方面，目前流行的关系推断方法大部分是应用于异构信息网络上的，即网络中的实体与边的类型是多种多样的，例如文献[8,14,20]等. 近期又有许多工作将时间信息融入了异构信息网络中，并利用这些信息来提高关系推断的准确性. 我们的提出的演化知识网络模型既包含知识的时间信息也包含知识的空间信息，并利用这些信息进一步提高了关系推断的准确率. 需要特别指出的是，YAGO2[4]已经提出了一个基于时空信息知识网络的模型 SPOTL，但是这个模型主要解决了 YAGO2 知识库上的知识的检索与查询的问题，并未将时空信息应用到关系推断问题上.

综上所述，由于现有的知识网络对于知识的时空信息的描述能力有限，导致在进行关系推断时无法对时空信息进行充分的利用，限制关系推断的正确率的提高，因此我们提出一种融合了时空信息的知识演化网络模型，并提出一种基于该网络的推断方法，提高关系推断的正确率.

本文的主要结构如下：第一部分提出一个演化知识网络模型；第二部分中我们提出一种关系推断方法 KP-LIM，并对该方法进行的详细介绍；第三部分主要介绍了实验采用的数据集，实验结果，以及结果分析. 第四部分对全文进行总结并提出几点下一步研究的方向.

1 演化知识网络模型

在这部分中，我们主要提出一个演化知识网络模型和定义在该网络上的一种特殊的子网络，链接延展模式.

1.1 演化知识网络

演化知识网络是一个异构的演化的多重图，且图中的节点和边都包含时间与空间信息. 具体定义如下:

定义 1. 演化知识网络. 给定一个时间集合T，空间集合S，则演化知识网络$G_{T,S}$可定义为一个八元组：

$$G_{T,S}=(V,E,\phi,\varphi,\theta,\tau,\lambda,\eta)$$

其中，V是演化知识网络中节点的集合，E有向边的集合，它的具体表示形式是一个三元组(u,v,r)，这里$u,v\in V$，$r\in R$，其中R是边的所有类型构成的集合；$\phi:V\to A$是节点类型的计算函数，使得每个节点通过该计算函数，可得到唯一的类型$\phi(v)\in A$，这里A为顶点的所有类型构成的集合；$\varphi:E\to R$表示在边集合中，某一条边的计算函数，且每一个实体对间最多有$|R|$条边. θ表示图中边的时间属性信息，用来描述一条边的发生以及存在的时间信息. τ是边的空间属性信息；λ是节点的时间属性信息，η是节点的空间属性信息.

在这个演化知识网络模型中，我们记录了图中节点与边的时间和空间信息. 这里的时间信息是一系列离散的时间戳，空间信息则是一系列离散的地理位置信息. 演化知识网络的可演化性主要体现在可通过感知网络中产生的新变化，与自身进行比较，发现新知识，并实现自我更新. 网络中的节点和边都有时间戳信息，他们都会随着时间的变化而演变，例如对于当前国家元首这个节点，会随着节点任职期满，而自动更新为前任领导人，这便体现了网络的时空可演化性.

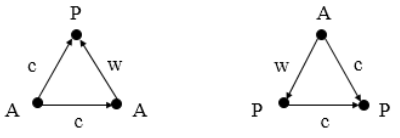
1.2 链接延展模式

基于我们提出的演化知识网络，本文着重研究在该网络上的关系推断问题. 关系推断的主要目的是，利用知识库中已有的知识作为基础，推断出两实体间可能存在的新关系，这里我们做的关系推断不仅仅要推断出新的边，还要给出边的类型. 推断的主要思路是，首先构造出所有可能存在的链接延展模式，然后建立一个混合背包问题模型，将每一个模式看作背包问题中待选择的物品，通过背包问题的求解，选择出对于关系推断有意义的模式，利用这些模式在图中进行匹配，推断出新的关系.

首先引入演化知识网络中的链接延展模式的定义，简称为 LE(Link Extendable) 模式.

定义 2. 链接延展模式. 已知一个关系集R，知识网络$G_{T,S}$，我们定义$G_{T,S}$上的一个子网络$H=(V',E')$，$V'\subseteq A$，$E'\subseteq R$，在这个子网络中任意两节点都可通过一条边进行关联，如果这个子网络中有n个节点，则称其为n元模式，我们将这个子网络叫做链接延展模式.

由于子网络中节点的个数越多，计算复杂度越大且对关系推断的结果提升较小，因此本文重点研究 3 元模式. 为了便于理解，下面我们简单列举一些典型 LE 模式. 假设$A'=\{author(A),paper(P)\}$，$R'=\{write(w),cite(c)\}$，若子网络中的点集为$V'=\{V1(A),V2(P),V3(A)\}$，边集$E'=\{w(V1,V2)$，$c(V3,V2)$，$c(V1,V3)\}$，则该 LE 模式由图 1(a)表示；同理，若子网络为$V'=\{V1(P),V2(A),V3(P)$，$E'=\{w(V2,V1)$，$c(V2,V3)$，$c(V1,V3)\}$，则该 LE 模式可刻画为 1(b).

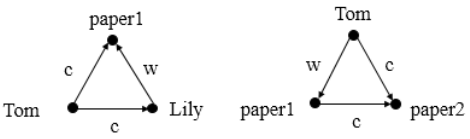


(a) (b)

Fig. 1 Examples for LE-Patterns

图 1 LE 模式样例示意图

在真实的演化知识网络中，凡是满足 LE 模式要求的子网络均称为 LE 模式的实例h，例如图 2 中(c)图表示图 1(a)的一个实例，(d)代表图 1(b)的一个实例.



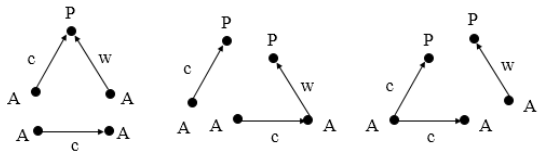
(c) (d)

Fig. 2 LE-Patterns instances

图 2 不同 LE 模式的实例示意图

由图可知，对于不同的 LE 模式的定义我们可以找到它相应的实例，且对用同一个 LE 模式可以有多个不同的实例. 在进行关系推断时，我们需要将

LE 模式进行分解，使其成为可用来实现关系推断的新的 LE 模式. 例如图 2(a)表示的一个 LE 模式，我们可将其拆解为 3 个可用于关系推断的新的模式，如图 3 所示：



(e) (f) (h)

Fig. 3 LE-Patterns for link prediction

图 3 用于关系推断的 LE 模式示意图

在图 3(e)中，我们将相连的两条边作为关系推断的条件，单独的一条边作为推断的结论. 在进行关系推断时，若已知在三个节点，他们的类型满足图 3(e)中 A-P-A 的要求，且节点类型为 A-P 和 P-A、的节点对之间的关系分别为 c 和 w，则我们可推出两节点间存在 c 关系. 例如在图 2(c)中，我们已知 Tom 和 Lily 类型为作者，paper1 的类型为论文，且已知 Tom 引用了 paper1，Lily 写了 paper1，则根据图 3(e)中的 LE 模式，我们可推断 Tom 和 Lily 之间存在引用关系. 需要指出的是，在网络中利用这些模式进行推断的结果并非全部正确，例如图 1(b)所代表的模式的含义是，某一位作者写了两篇文章，可得出这两篇文章之间存在引用关系，而事实上，这个引用关系可能不存在，因此，对于网络中包含的所有的模式构成的集合，我们需要利用背包问题的思想，从中选出置信度较高且涵盖关系类型更广泛的模式子集，并用子集中的所有模式进行关系推断.

2 关系推断方法

2.1 基于混合背包问题的关系推断方法(KP-LIM)

为了实现基于某一演化知识网络 $G_{T,S}$ 上的关系推断，首先需要找出网络中所有可能存在的可用于关系推断的模式，通过混合背包问题求最优解的思想，对不同模式进行选择.

下面我们先简要介绍一下背包问题. 背包问题（Knapsack problem）是一种组合优化的 NP 完全问题，问题可以描述为：给定一组物品，每种物品都有自己的重量和价值，在限定的总重量 M 内，我们如何选择才能使物品的总价值最高.

这里我们将不同的模式看作背包问题中要装进背包中的物品，因此每个模式需要有相应的重量 *Weight* 和价值 *Value* 两个参数. 我们从不同 LE 模式在网络中匹配的实例个数以及正确的实例个数的角度，给 LE 模式的两个参数重量 *Weight* 和价值 *Value* 做了以下定义.

定义 3. 模式价值 *Value*. 已知演化知识网络 $G_{T,S}$，某一 LE 模式 *le* 的所有实例的集合 $H=\{h_1,...h_n\}$，其中包含 n 个不同实例，且每个实例中的每条边包含有该边产生的时间信息 t，则 *le* 的价值 $V(le)$ 为

$$V(le)=\sum_{i=1}^{n}\sum_{j=1}^{3}\omega_i(t_j)$$

这里 $\omega_i(t)$ 代表某一条边产生的相对时间，即 $\omega_i(t)=t-t_0$，t_0 为整个网络中边产生的最早时间. 因此，LE 模式的价值代表了 LE 模式对应所有实例间关系时间的和，即当 LE 模式对应的实例越多，且相应的实例发生的时间越靠后，该模式的价值越大.

定义 4. 模式重量 *Weight*. 已知一个 LE 模式中包含的三个关系为 E_1,E_2,E_3，若我们要用该 LE 模式对关系 E_3 进行推断，首先通过遍历全图得到所有满足 LE 模式中节点类型和关系 E_1,E_2 要求的所有实例的个数假设为 N，且这些实例中又满足 LE 模式中关系 E_3 要求的实例的个数为 n，则 *le* 的重量 $W(le)$ 为

$$W(le)=1-n/N$$

例如，由图 1(a)表示的 LE 模式，若用它来推断作者间的引用关系，它在全图中所有匹配上另外两条关系规则的实例数量有 100，其中两作者间存在引用关系的有 99 个实例，则该 LE 模式的重量为 0.01. 下面我们介绍如何将 LE 模式的选择问题建模成一个混合背包问题.

首先，根据要推断关系的类型的不同，将所有的 LE 模式分成 k 类，因此这里的 k 等于关系类型的总数. 假设每一类中包含 $N_i(i=1,2,...,k)$ 个不同 LE 模式. 为了满足所有关系类型都可以被推断出来，我们在应在每个分类中至少选出一个 LE 模式. 最终该问题可归纳为一个多重混合背包问题：

$$\max mize=\sum_{i=1}^{k}\sum_{j=1}^{N_i}v_{ij}x_{ij}$$

$$subject\quad to\ \sum_{i=1}^{k}\sum_{j=1}^{N_j}w_{ij}x_{ij}\le M$$

$$\sum_{j\in N_i}x_{ij}\ge 1,\quad x_{ij}\in\{0,1\}$$

这里 x_{ij} 表示对于第个 i 分类中的第 j 个 LE 模式， $x_{ij}=1$ 表示选择该模式， $x_{ij}=0$ 表示不选择该模式， v_{ij} 表示该模式的价值， w_{ij} 表示重量.

为了求解上面的混合背包问题，我们将问题拆解成两个 0-1 背包问题：1）多重选择背包问题，即在约束为耗费小于 $M^*(M^*<M)$ 条件下，从每个分类中选出一个结果，这个步骤主要是保证每个分类中都有一个 LE 模式被选择出来，即在后期做推断时，每种关系都可能被推断出来；2）常规的背包问题，在耗费小于 $M-M^*$ 约束下我们可以从剩下的所有模式中选择更多意义的模式，提高推断的召回率.

2.2算法及分析

实现关系推断的过程可主要分为以下几个步骤：1）构造可能存在的 LE 模式. 已知演化知识网络 $G_{T,S}$ 上的边的所有类型，则任意三种类型的边的组合可构成一个候选 LE 模式. 2)背包问题实现模式选择. 遍历全图，找出不同模式相应的所有实例，计算得到不同模式的 *Weight* 和 *Value*，通过混合背包问题的求解选出有意义的 LE 模式. 3)利用选出的模式在网络中进行匹配得到推断结果. 下面算法一中给出我们提出的关系推断算法的实现.

算法 1. 基于链接延展模式的关系推断

输入：演化知识网络 $G_{T,S}$，待推断的实体对集合 $\{(n,n')_1 \ldots (n'',n''')_k\}$

输出：推断出的关系集合 *result*

① 初始化 Le 模式集合 *LE{}*，结果集合 *result{}*
② for $r_1 \in R$
③ for $r_2 \in R$
④ for $r_3 \in R$
⑤ 构造一个 LE 模式 *le*
⑥ 将 *le* 加入 *LE{}* 中
⑦ end for
⑧ end for
⑨ end for
⑩ 构造混合背包问题
⑪ 调用算法 2($G_{T,S}$, $LE\{le_1,\ldots,le_m\}$)计算背包问题参数
⑫ 求解背包问题得到新的模式集合 $LE'\{le_1,\ldots,le_n\}$
⑬ for $le_i \in LE'$
⑭ for $j=1$TO k
⑮ 针对节点对$(n,n')_i$匹配模式 le_j
⑯ if 匹配上该模式 then
⑰ 将推断结果添加到 *result{}*
⑱ end if
⑲ end for
⑳ end for

由于在网络中实体间关系的类型的数量较少，因此LE模式的构造部分不是主要耗时的部分. 分析可知整个关系推断过程的计算量集中在步骤 2）中的全图遍历找与模式向匹配的实例，即上述算法中调用的算法 2. 在算法二中我们采用一些优化技巧以提高效率，适应大数据环境的要求. 下面我们给出算法 2 具体过程.

算法 2. 实例匹配与查找算法.

输入：演化知识网络 $G_{T,S}$，LE 模式集 $LE\{le_1,\ldots,le_m\}$

输出：集合中所有模式的 *cost*，*strength*

① 初始化实例正例、反例映射表 *map1(le,num)*，*map2(le,num)*
② for $node_i$ in $G_{T,S}$
③ for $relation_j$ in 所有与 $node_i$ 连接的边
④ for $relation_l$ in 所有与 $node_i$ 连接的边 and $relation_l$的另一端点为 $node_k$，$relation_j$的另一端点为 $node_h$
⑤ if $relation_j$ and $relation_l$类型满足 le_i 要求
⑥ if $node_k$ 和 $node_h$ 之间满足 le_i 要求
⑦ *map1* 中 le_i对应的数值加 1
⑧ else
⑨ *map2* 中 le_i对应的数值加 1
⑩ end if
⑪ end if
⑫ end for
⑬ end for
⑭ end for

一般的遍历全图找不同模式的实例的方法是，对于每一个模式，遍历图中的所有节点，对于某一个节点，遍历它所有的边，如果满足模式要求，则以该边的另一个端点为起点遍历其它所有的边，看是否满足模式要求，依此类推. 假设共有 m 个不同实例，全图有 n 个节点，且图中节点的平均出度入度和为 r，则该运算的复杂度为 $O(mnr^3)$. 以上方法虽然容易实现，却运算复杂度很高，效率低下，因此在我们的算法 2 中利用了一些技巧，降低了算法的时间复杂度.

首先，我们不针对每个模式遍历一遍全图来找

实例，而是做一个映射表，在这个映射表中不同模式对应的值为到当前为止该模式匹配上的实例个数，因此只需遍历一遍全图即可得到所有模式的实例个数. 由于这里我们采用的模式均为三元模式，基于三角形的特殊构造，对于每个节点的具体匹配过程，我们不需要从一个节点出发，以广度优先的思想遍历三层关系，只需要从一个节点出发，找出它自身的所有关系，任意两个关系为一组，若这两个关系对的终点之间也存在关系则可构成一个 LE 模式，将映射表中的相应模式的 value 值加 1 即可. 该算法经过优化后的时间复杂度为 $O(nr^2)$.

3 实验

在这部分中，我们将详细介绍相关的实验结果，证明我们提出的基于混合背包问题的关系推断方法 KP-LIM 的合理性与有效性，以及 KP-LIM 方法对于大数据环境的适应性.

3.1 实验数据集及参数选择

我们采用来自不同领域的数据构建成两个演化知识网络进行关系推断实验. 这两个演化知识网络的数据分别来自于学术领域和电影领域，均包含了多种不同类型的节点和关系. 其中学术网中的数据是从知名的学术文章网站 Soscholar①上爬取得到的，在这个网络中包含：10, 000,000 个作者(A)，7,000,000 篇论文(P)，500,000 个杂志(M)，14,000 个会议(C)和 50,000 个关键词(K). 我们选用论文的发表时间以及会议的召开时间作为网络中时间信息，选机构所在地构成网络中的空间信息.

对于电影网，它的数据主要来自于对知名电影网站 IMDB 上的电影信息的爬取. 该网主要包括：演员(AC)和导演(D)共 4,530,159 个，电影(M) 2,132,383 部，我们选取电影上映时间，拍摄地等信息作为电影知识网络的时间与空间信息. 在表 1 和表 2 中分别列举了学术网和电影网中所有的实体间的直接关系. 需要指出的是，下表中罗列的是从元数据中抽取的直接关系.

Table 1 Direct relations in the scholar network

表 1 学术网络直接关系表

实体类型	关系类型
A,P	写作，被写作
C,P	发表，被发表
O,P	拥有，被拥有
P,K	提到，被提到
A,O	雇佣，被雇佣
P,P	引用，被引用

① http://soscholar.com/

Table 2 Direct relations in the film network

表 2 电影网络直接关系表

实体类型	关系类型
AC,M	出演，由…出演
D,M	导演，由…执导

表 3 中列举了一些学术网络中的常见的 LE 模式，其中第三列是 LE 模式需要从网络中匹配实例的规则，第二列是待推断关系的类型，第一列是待推断关系中两端节点实体的类型. 在电影网中生成规则的方法与学术网相同，这里不再一一列举相应的 LE 规则.

Table 3 General LE-Rules in the scholar network

表 3 学术网络中常见的 LE 规则

实体类型	关系类型	规则
A,A	coauthor	A←(write)P→(write)A
A,A	colleague	A←(work)O→(work)A
A,C	contribute	A→(work)O→(contribute)C
A,P	cite	A→(write)P→(cite)P

对于学术网，已知边的类型数，则根据 LE 模式的特点可以找出所有可能存在的 LE 模型，去掉一些不可能存在的模式，最终，对于学术网我们可得到了 124 个不同的模式. 同理对于电影网，我们共得到 18 种模式.

实验过程中，我们在图中随机隐藏掉部分关系，并通过剩下的网络中的信息对隐藏掉的关系进行推断. 为了证明我们提出的方法的有效性，在实验中我们将 KP-LIM 方法与两种最近比较流行的关系推断方法进行比较，一个是经典的有监督算法的代表逻辑回归方法[14]（简称 Logistic），一个是论文[8]中提出的方法 M-CN+ANC+OAWpress（简称 CN），它是一种典型的无监督学习方法. 对于逻辑回归方法，我们是将每一个 LE 模式作为一维特征，构造训练数据，通过逻辑回归算法进行学习，学出每个特征的系数，在进行关系推断时，给出一个实体对，

匹配不同的模式，如果没有模式可以匹配上，则推断两节点间没有关系，如果有模式匹配上则不同的关系可以得到一个分数，当分数大于某一阈值时，我们推断这两节点间存在这种关系，实验可得，当阈值选择为 0.6 时，逻辑回归的推断效果最好，因此在后面的比较实验中，选择阈值为 0.6. 对于 M-CN+ANC+OAWpress 方法，在给出实体对后，针对每种关系可计算出一个得分，同理我们选择一个阈值，当得分大于该阈值时，推断两实体间存在某一关系，实验可知当阈值选择 0.8 时效果最好.

对于 KP-LIM 模型，在背包问题中有两个参数需要确定，背包问题的总体约束值 $M>0$ 和多重选择背包问题的约束条件 M^*，对于 M^* 有一个宽泛的约束是 $0<M^*<M$，而事实上我们对 M^* 给出一个更为严格的约束即对于每一个模式分类，找出一个耗费最小的模式，则 M^* 应大于所有分类中最小耗费值之和，用公式可表示为：

$$\sum_{i=0}^{k} c_i^{\min} < M^* < M$$

对于 M 的值，也有一个更为严格的约束，即所有模式的耗费之和，对于学术网 M 值的上限 $M_{\max}$ 是 10.2，对于电影网的 $M_{\max}$ 是 4.15.

为了确定 M^* 和 M 的具体值，我们首先固定 M 的值，变更 M^* 的值，来研究 M^* 的值变化对于关系推断效果的影响，找出推断结果最优时 M^* 的值，然后固定 M^* 的值，在 $(M^*, M_{\max})$ 的范围内调节 M 的值，找出推断结果最优的 M 值，最终可通过实验确定出两个参数的值. 图 4 中给出了对学术网的进行参数调节的过程，其中(a)，(b)，(c)，(d)分别代表 $M=6,7,8,10$ 时对 M^* 的值进行调节得到的测试结果，从图 4 中可知，当 $M=7$，$M^*=5$ 时，对于学术网上的关系推断结果最好，同理我们对电影网上的数据进行测试，得到当 $M=3$，$M^*=2$ 时结果最优.

3.2 实验结果及分析

在这部分我们会通过实验进行两方面的比较，一是比较 KP-LIM 与 Logistic 和 CN 方法间的关系推断正确率，二是 KP-LIM 算法对于大数据的适应性的性能测试.

3.2.1 不同关系推断方法的推断效果比较

在进行不同关系推断方法的比较试验时，我们选择准确率作为评价指标，主要是因为基于大规模知识网络的关系推断，推断结果的正确与否具有决定性作用，推断结果作为知识必须要保证期准确性，因此这里我们选准确率作为指标.

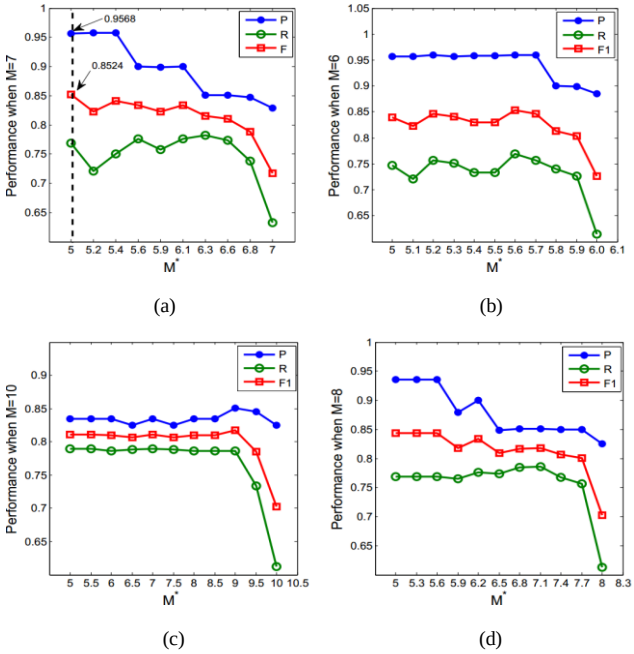


Fig. 4 Parameter tuning on the scholar network

图 4 学术网调参结果

实验结如表 4 所示，表中主要比较了三种不同方法的准确率，在这里 α 表示隐去的关系数量占全图关系总数量的比例 α，随着 α 的增加 KP-LIM 方法的表现均优于其他两种方法，与有监督算法 Logistic 相比，我们的方法在学术网上准确率获得了 2%-62%的提高，对于 CN 方法，我们的方法获得了 28%-70%的提高，平均提高量分别为 29%和 37%. 同理在电影网上，我们的方法分别获得了和 1%-22%和 20%-30%的提高，平均提高值分别为 8%和 24%.

综合以上结果可得出结论，KP-LIM 方法在不同的数据集上，比当前比较流行的两种方法均取得更好的推断结果. 其主要原因是，我们的方法将不同模式的特点以及全图的背景知识信息融入到背包问题中，通过背包问题的求解，从模式集合中选出高质量的模式进行关系推断，而 Logistic 采用了所有的模式，仅仅通过训练数据给不同的模式学习出不同的系数，一是未将全部的背景知识信息加以利用，模型的好坏完全受训练数据的好坏的影响，二是并未对模式进行筛选. 而 CN 方法只考虑了待推断的两个实体的相关关系，对图的结构信息利用不充分，因而它的正确率最低.

3.2.2 KP-LIM 的性能测试

由于在大数据环境下，数据量急剧增长，能否适应大数据的挑战，也是衡量一个算法好坏的重要方面，因此，下面我们对 KP-LIM 方法的计算性能进行测试. 首先分别构造不同大小的知识网络，测试在不同规模的网络上 KP-LIM 进行关系推断的时间消耗，这里我们将网络中的点的数量从 1，000,000 逐渐增加的到 10,000,000，并记录下推断时间的变化，如图 5 所示：

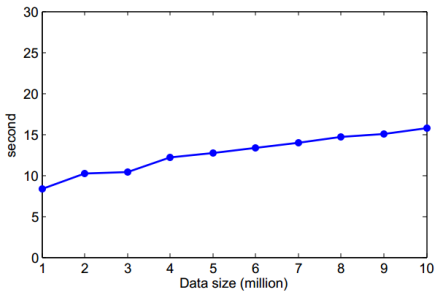


Fig. 5 The change of time consumption with the increase scale of relation network

图 5 随着网络规模的增加关系推断的时间消耗

由图 5 可知，随着网络规模的迅速增加，我们提出的 KP-LIM 推断方法的时间消耗的增长缓慢. 当网络规模扩大了 10 倍，推断耗时仅从 8.372s 增加到 15.726s. 在前面的算法部分我们也对 KP-LIM 方法的时间复杂度进行了分析，该算法的主要时间消耗集中在遍历全图找不同模式实例的过程，复杂度为 $O(nr^2)$，这里 n 代表网络中节点的个数，r 图中的节点的出入度，因此随着网络规模的增加 r 基本的变化较小，整个算法的时间消耗主要受到网络中节点的个数 n 的影响，因而随着网络规模的增加算法的时间消耗呈线性增长。综上所述，我们提出的关系 KP-LIM 不仅在推断准确率上取得了较好的结果，其计算成本也并没有对着网络规模的扩大而呈指数增长，因此 KP-LIM 的计算性能也能够满足大规模知识网络对关系推断性能与效率的要求.

4 研究展望与总结

本文首先提出了一个融合了时间与空间信息的演化知识网络，基于该网络提出了一种关系推断方法. 实验证明我们的方法比当前流行的一些关系推断方法取得了更高的准确率，且对于大数据的环境下依然拥有较好适应性.

关于这个工作，仍有以下几个需要研究的方向：1）KP-LIM方法对知识网络中已有的关系的数量的依赖性较强，当面对冷启动问题时，如何保证推断的正确率有待进一步研究；2）对于演化知识网络中的空间信息的利用有限，下一步可研究如何更充分的利用网络中的时空信息，进一步提高推断效果.

Table 4 precision obtained by using different inference methods

表 4 不同关系推断方法正确率值的比较

数据集	方法	正确率(%)									
		=0.01	α =0.03	α =0.05	α =0.01	α =0.07	α =0.1	α =0.3	α =0.5	α =0.7	α =0.9
Movie	CN	68.44	71.71	71.79	67.65	67.65	67.65	42.74	25.00	39.01	51.97
	Logistic	97.28	92.66	87.73	77.01	78.72	77.01	44.01	48.26	25.31	32.63
	KP-LIM	**100**	**100**	**99.82**	**99.69**	**99.76**	**99.29**	**89.89**	**95.90**	**87.45**	**80.89**
Soschool	CN	78.12	76.51	76.38	76.37	75.33	75.92	69.90	55.20	49.96	46.60
	Logistic	94.64	95.37	94.76	94.80	90.18	89.88	76.80	63.78	72.18	74.53
	KP-LIM	**99.65**	**98.21**	**97.81**	**96.46**	**95.68**	**95.68**	**91.03**	**85.90**	**84.76**	**82.94**

参考文献

[1] Wang Yuanzhuo, Jin Xiaolong, Cheng Xueqi. Network big data: Present and future [J]. Chinese Journal of Computers, 2013, 36(6): 1125-1138 (in Chinese)
(王元卓, 靳小龙, 程学旗. 网络大数据：现状与挑战[J]. 计算机学报，2013,36（6）：1125-1138)

[2] Lujun F, Anish D S, Cong Y, et al. Rex: Explaining relationships between entity pairs [J]. VLDB Endowment, 2011, 5(3): 241–252

[3] Suchanek F M, Kasneci G, Weikum G. Yago: A large ontology from

wikipedia and wordnet[J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2008, 6(3): 203-217

[4] Hoffart J, Suchanek F M, Berberich K, et al. YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia[J]. Artificial Intelligence, 2013, 194: 28-61

[5] Jia Yantao, Wang Yuanzhuo, Cheng Xxueqi, et al. OpenKN: An open knowledge computational engine for network big data [C] //Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on. BeiJing: IEEE, 2014: 657-664

[6] Hailun Lin, Yantao Jia, Yuanzhuo Wang, et al. Populating Knowledge Base With Collective Entity Mentions: A Graph-based Approach. [C] //Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on. BeiJing: IEEE, 2014:504-611

[7] Cheng Xueqi, Jin Xiaolong, Wang Yuanzhuo. Survey on Big Data System and Analytic Technology [J]. Journal of Software, 2014, 25(9):1889-1908 (in Chinese)
(程学旗, 靳小龙, 王元卓, 等. 大数据系统和分析技术综述[J]. 软件学报, 2014, 25(9):1889-1908)

[8] Lee J B, Adorna H. Link prediction in a modified heterogeneous bibliographic network[C] //Proc of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012). Washington, DC：IEEE Computer Society, 2012: 442-449

[9] Rossetti G, Berlingerio M, Giannotti F. Scalable link prediction on multidimensional networks[C] //Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on. Vancouver, BC: IEEE, 2011: 979-986.

[10] Cho E, Myers S A, Leskovec J. Friendship and mobility: user movement in location-based social networks[C] //Proc of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY: ACM, 2011: 1082-1090

[11] Wang Dashun, Pedreschi D, Song C, et al. Human mobility, social ties, and link prediction[C] //Proc of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY ACM, 2011: 1100-1108

[12] Crandall D J, Backstrom L, Cosley D, et al. Inferring social ties from geographic coincidences [J]. National Academy of Sciences, 2010, 107(52): 22436-22441

[13] Popescul A, Ungar L H. Statistical relational learning for link prediction[C] //IJCAI workshop on learning statistical models from relational data. 2003, 2003

[14] Sun Yizhou, Barber R, Gupta M, et al. Co-author relationship prediction in heterogeneous bibliographic networks[C] //Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on. Kaohsiung: IEEE, 2011: 121-128

[15] Sun Yizhou, Han Jiawei, Aggarwal C C, et al. When will it happen?: relationship prediction in heterogeneous information networks[C] //Proc of the fifth ACM international conference on Web search and data mining. New York, NY: ACM, 2012: 663-672

[16] Tang Jie, Lou Tiancheng, Kleinberg J. Inferring social ties across heterogenous networks[C] //Proc of the fifth ACM international conference on Web search and data mining. New York, NY: ACM, 2012: 743-752

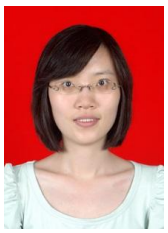
[17] Jia Yantao, Wang Yuanzhuo, Li Jingyuan, et al. Structural-interaction link prediction in microblogs[C] //Proc of the 22nd international conference on World Wide Web companion. Switzerland: ACM, 2013: 193-194

[18] Davis D, Lichtenwalter R, Chawla N V. Multi-relational link prediction in heterogeneous information networks[C] //Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on. Kaohsiung: IEEE, 2011: 281-288

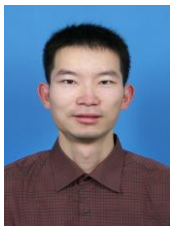
[19] Liu Dawei, Wang Yuanzhuo, Jia Yantao. LSDH: a Hashing Approach for Large-Scale Link Prediction in Microblogs[C] //Proc of the 28th AAAI'14. Québec City, Québec, 2014

[20] Zeya Zhao, Yantao Jia, Yuanzhuo Wang. Content-structural relation inference in knowledge base[C] //Proc of the 28th AAAI'14. Québec City, Québec, 2014

[21] Yang Yang, Chawla N V, Sun Yizhou, et al. Predicting Links in Multi-relational and Heterogeneous Networks[C] //ICDM. 2012, 12: 755-764.



Zhao Zeya，born in 1990. Master student. Her main research interests include open knowledge engineering, data mining, social computing.



Jia Yantao，born in 1983. PhD, Assistant professor. His main research interests include open knowledge network, social computing, combinatorial algorithms.

Wang Yuanzhuo， born in 1978. PhD, Associate professor. He is a IEEE member and a senior member of China Computer Federation. His current research interests include social computing, open knowledge network, network security analysis, stochastic game model, etc.

Jin Xiaolong, born in 1976. Associate professor, PhD supervisor. His research interests include social computing, multi-agent systems, performance modelling and evaluation, etc. He has published over 85 papers.

Cheng Xueqi, born in 1971. Professor, PhD supervisor. His research interests include network science, websearch & data mining.

校对负责人：赵泽亚，18311091893，E-mail: zhaozeya@software.ict.ac.cn