基于超平面时间维的演进知识图谱链接预测

摘要

知识表示和推理是知识图谱领域的研究热点。人类知识一直是增加、删除、更新的动态变化过程，知识图谱也因此不是一个静止的状态，而是会随着知识的变化改变。然而大多数现有的知识图谱方法在表示和推理过程中忽视了时间维度的重要影响。时间具有演进性，所以时间不能作为实体存在于三元组中。基于此，本文提出T2LP模型，它通过建立一个增加时间维的四元组，明确地将时间信息添加到表示和推理的过程中，同时量化了时间对事件带来的影响，基于相似性度量，最大似然方法和优化贝叶斯概率方法，对节点间依赖程度进行多样性研究和定量分析，并基于优化的贝叶斯网络实现高效链接预测，通过提取带有时间信息数据集的实验，进一步验证了本方法的有效性。

关键词 贝叶斯优化；时间演进；知识图谱；概率推理；链接预测；影响函数

1 引言

随着近年来人们对信息检索的积极探索，知识数据语义化成为一个研究热点。作为知识的一种语义化的表达方式，语义网[9,10]引发了广泛的关注。语义网概念的提出使大量HTML数据实现了语义化，为具有计算机可理解的语义提供了逻辑支持，对现实世界的事物及其相互关系进行形式化的描述[1]。但越来越多的数据堆积，开始出现结构松散、语义复杂等问题，为了使符号数据可直接参与运算，形成易于统计和应用的知识，以语义Web知识架构为基础的研究逐渐深入。2012年5月，Google首次提出知识图谱(KG，Knowledge Graph)的概念，并广泛用于语义搜索、问答系统、智能客服等应用，在智慧医疗[36]-[39]、人脉路径查询、企业社交、信贷审核反欺诈等特定领域也发挥了重要作用。知识图谱[7]是图形结构的知识库，具有强大的语义表达处理能力和图结构化展示的优势。链接预测是知识图谱推理领域的重要研究方向之一[11]，是机器学习领域的一个新兴课题[14]。它是为自然语言中的词找到一个映射或者函数，将源数据生成一个在新的连续低维向量空间上，从而找出未知或异常的关系对KG补全。为了实现这个目的，很多研究人员做了一系列的努力。基于能量函数的表示方法中，A Bordes等[15]提出了一种将实体与关系嵌入到低维向量空间中的简单模型TransE，弥补了传统方法训练复杂、不易拓展的缺点。但是TransE过于简单，很难对多关系建模。为了平衡模型复杂度和建模效果，Zhen Wang等[16]提出了TransH，解决了多关系建模的难题，还设计了复杂取样的办法用于训练，但是同一关系可能代表不同含义，因此TransD模型[17]建立了动态映射矩阵，适用于规模很大的KG。TransG[19]根据贝叶斯非参数无限混合模型和三元组的特定语义得到当中的最佳部分，对一个关系生成多个表示。KG2E[20]模型提出了用Gaussian Distribution的协方差来表示实体和关系的不确定度的新思想，提升了已有模型在Link Prediction等问题上的准确率。基于张量分解方法以Nickel等[23]提出的RESCAL为代表，将三元组对应张量分解为核心张量和因子矩阵的乘积，通过最小化结构风险或边界误差学习隐藏向量，但是关系数目较多时，分解过程的计算量会非常大。文献[13]旨在介绍使用时间点过程框架的强大数学工具进行时间推理而不是动态演化的知识图，除了预测事实的发生之外，还能够预测事实可能发生的时间。目前大多数方法研究已经对知识图谱表示学习和推理工作进行了有效推进。但是现实世界中的知识并不是静态存在的，而是带有时间标签，会随时间发生变化的动态知识，需要能够表征和推理时间演变的方法。但是目前大多数研究工作还没有将时间信息完全考虑进去，文献[12]提出了一种动态知识图谱来动态摄取传入新知识，并根据最近的关系和行为更新实体嵌入，但缺乏对时间维度上变化带来额外信息的详细刻画，造成信息的遗漏和损失、错误信息率高等问题，对于预测推理的效率也有待提高。文献[26]专注于预测未注释边的时间有效性的任务。HyTE模型[22]用了类似TransH模型的思路来整合时间维度的信息，表示事件发生的时间段，但是大多事件发生都是针对某一时间点的。因此，本文解决了时间对知识表示推理过程产生影响的问题，将时间信息加入到刻画实体关系的过程中，我们的策略是利用这个额外的时间维度来增强链接。主要研究内容可总结为以下几点：

（1）为了刻画知识的时间标签及随时间改变的动态特征，本文提出一种新的模型T2LP，在原有的实体-关系维基础上，添加了时间维，从时间维度出发考虑知识；

（2）针对事件发生时间的先后对实体或关系的影响程度不同，本文提出影响结果因子的概念以及影响函数，并给出建模过程及其规则，量化了时间对事件带来的影响程度；

（3）为了实现T2LP的高效链接预测，本文基于贝叶斯网络并进行优化，主要通过基于针对相邻节点计算关系相似性度量的特征学习方法来研究，另一方面，定义打分函数测量关系成立的可能性。可有效地分析不确定性知识，缺失的链接不被视为错误，隐式关系发现，提高了链接预测的准确性；

（4）本文基于图特征模型Neo4j，实验并测试了预测隐含关系，对已知关系进行概率推理验证。在Freebase、WordNet两个数据集上进行对比，我们的模型实现了最先进的结果，进一步表明本文方法的有效性。

本文组织结构如下：

第1节提出研究背景及意义，以及国内外研究现状，指出局限性；第2节对T2LP建模过程和方法进行描述；第3节设计T2LP的实现流程；第4节通过横向对比实验和性能分析，验证本文算法的有效性；第5节总结与未来工作展望。

2 四元组链接预测

2.1 知识嵌入

对于传统知识图谱，通用表达方式是三元组结构，即。表示头实体，表示尾实体，表示头尾实体之间的关系。三元组成立满足

|  |  |  |
| --- | --- | --- |
|  |  | (2.1) |

但是事实一般都是具备时间属性的，现实中的图结构是随时间改变的（如图1），传统三元组没有能够表示时间的信息。因此我们在传统知识图谱三元组基础上加入时间参量，形成时间演进知识图谱四元组。其中是时间维度信息。定义实体集，关系集，时间集，其中，

|  |  |  |
| --- | --- | --- |
|  |  | (2.2) |
|  |  | (2.3) |
|  |  | (2.4) |

所以，在中，显式四元组集合可表示为。

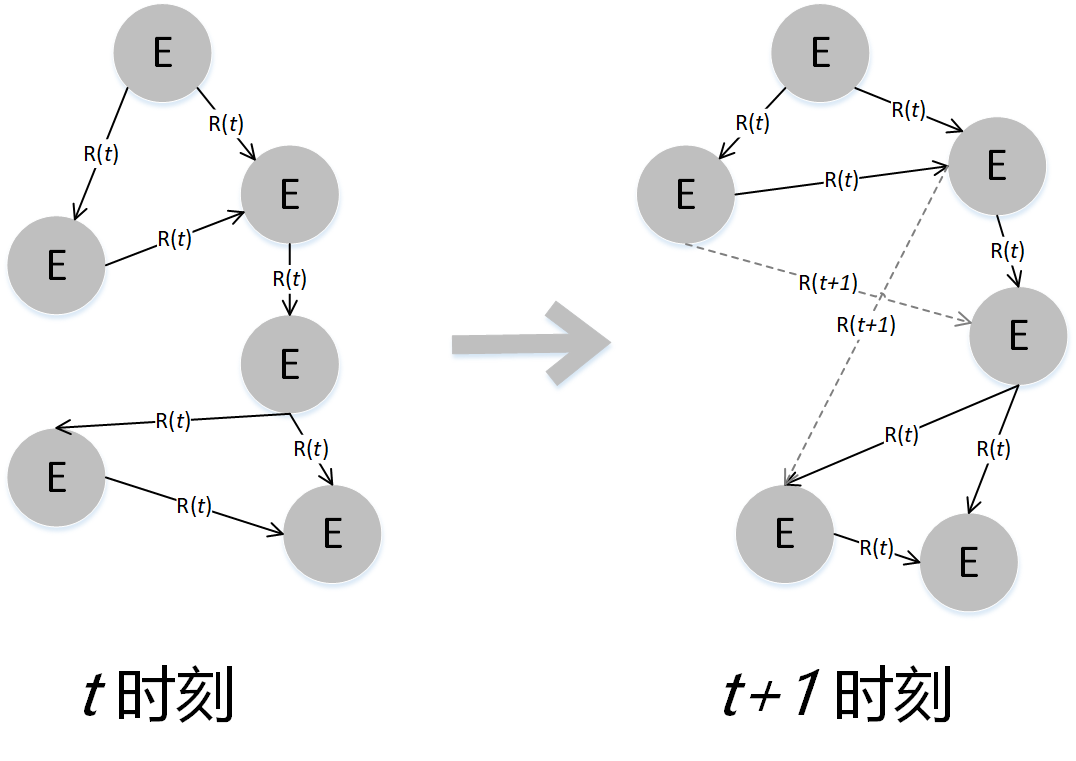


图1 KG动态变化示意图

另外，为了知识嵌入方便，此处引入四元组成立概率。四元组成立的概率记作。根据条件随机场模型求解的方法，当前事件成立的概率可以近似看作是过去已发生近似事件概率的乘积，具体地，

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

式中为平滑参数，其中。即当事件A与事件B的头实体与尾实体的分布相近，关系相似时，事件A成立的概率越大。

要判断关系相似性，就要考虑到估算样本间的相似性度量，这时采用的方法就是计算两个样本之间的距离。通过计算样本嵌入成向量的夹角余弦值来评估它们之间的相似度，余弦相似系数用符号表示：

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

其中，和表示两个样本映射所得向量。，用来判断相似性，结果越接近0，相似度就越高。该方法同样适用于高维向量空间中的相似度判断。

贝叶斯网常被用来处理不完整和带有噪声的数据集，进而也使得使用贝叶斯进行推理和预测变得相对容易实现。其理论基础是贝叶斯法则，通用公式为：

|  |  |  |
| --- | --- | --- |
|  |  | () |

其中，A是事件，B是事件A的划分，。

将隐式变量置为随机值然后进行概率推理。

对于以上定义的事件成立概率及其四元组，我们假设满足定义的函数具备以下条件：

**R1**：（归一性）对于两组成立的三元组

|  |  |  |
| --- | --- | --- |
|  |  | (2.5) |
|  |  | (2.6) |

其中，，，，无限制类型，属于时间类型信息。联合(2.5)(2.6)式可得：

|  |  |  |
| --- | --- | --- |
|  |  | (2.7) |

(2.7)式归一化为一组带时间维的四元组表示形式。

**R2**：（传递性）存在四元组及，满足

|  |  |  |
| --- | --- | --- |
|  |  | (2.8) |

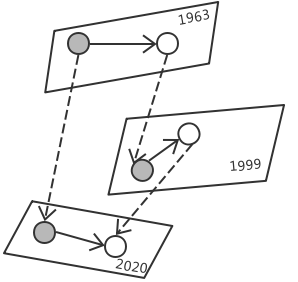


图2 传递性示意图

**R3**：（一致性）

1. 时刻到时刻无新事件发生，则

|  |  |  |
| --- | --- | --- |
|  |  | (2.9) |

1. 为两个事件，且,，其中符号表示递进，即发生在之前，那么

|  |  |  |
| --- | --- | --- |
|  |  | (2.10) |

**R4**：（自反性）对一组成立的四元组，如在时刻，头实体和尾实体不同但关系可以互换，那么

|  |  |  |
| --- | --- | --- |
|  |  | (2.11) |
|  |  | (2.12) |

例如，Tom和Jerry在2019年成为朋友，用四元组表示：

|  |  |  |
| --- | --- | --- |
|  | (Tom, friend\_of, Jerry, 2019) = (Tom, friend\_of, Jerry, 2019) | (2.13) |

是成立的。

**R5**：（反自反性）对一组成立的四元组，如在时刻，头实体和尾实体不同且关系不可以互换，那么

|  |  |  |
| --- | --- | --- |
|  |  | (2.14) |

例如，Tom和Jerry在1995年成立父子关系，用四元组表示：

|  |  |  |
| --- | --- | --- |
|  | (Tom, father\_of, Jerry, 1995) (Tom, father\_of, Jerry, 1995) | (2.13) |

**R6**：（衰减性）为事件集，为其子事件，为子事件对当前产生的影响。结合衰减性定义及贝叶斯原理[]：

1. ，则

|  |  |  |
| --- | --- | --- |
|  |  | (2.16) |
|  |  | (2.17) |

1. ，则

|  |  |  |
| --- | --- | --- |
|  |  | (2.18) |

过去每次事件的发生都会对现在的事件产生影响。一般认为，随着时间的演进，事件的影响力会在传递过程中发生衰减，即历史发生的近似事件会随着时间的变化对当前事件产生的影响力逐渐降低，我们将这一过程定义为影响衰减过程，并用符号表示激活函数：

|  |  |  |
| --- | --- | --- |
|  |  | (2.15) |

其中，是用于平滑处理的超参数，可以调节衰减速率，。公式（2.15）是满足以上定义及条件特性的改进sigmoid函数，保留其收敛特性的同时解决了饱和性缺陷[用到sigmoid函数的论文]。函数变化趋势如图2：

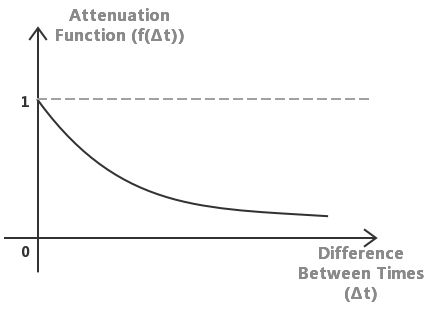


图2 影响衰减函数

对于时间演进的动态知识图谱来说，当新实体被引入并且新链接通过原来的实体产生相应的关系而出现时，实体和关系的集合都随时间而变化。

2.2. 链接预测

当前的知识图谱中仅包含了现实世界中很小一部分事实，还存在大量有待发现的实体之间的关系，并且这些关系不是静止的，而是会随着时间的演进时刻动态变化的。因此，链接预测问题是理解大型知识库结构的关键。考虑到时间参数的影响，旨在给定实体之间的现有链接的情况下，预测某一时刻知识图谱中隐式链接以及新链接。

考虑到有时间标签的知识在更高维度的空间中能够更加完整的表示，为了增强模型的表现力，我们提出T2LP（TimeLine Based Link Prediction）模型，在实体-关系维基础上增加时间参量作为新维度，新增了时间维超平面。T2LP的三维架构如图3所示：

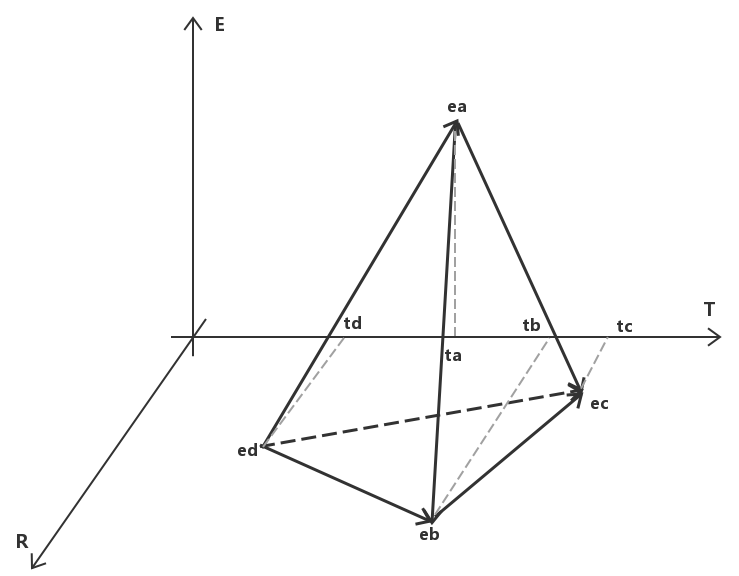


图3 关系-实体-时间三维架构

图3中的所有四元组可以分别表示为（以A，B，C，D，E标记）：

|  |  |  |
| --- | --- | --- |
|  |  | () |

一个先前发生的事件对后发生事件的影响力可以通过动态链接预测中的影响结果因子表示，用来描述事件X对事件Y产生的影响，这些都是可见关系之间影响力，用于关系之间重要程度的度量。

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  | () |

根据计算出的影响结果因子可以获得分布矩阵。接下来可以根据事件之间不同的关系影响力得到一个新的有向加权图，从而进行概率推理。

现实存在的事实很多是动态变化的，因此具有时效性，比如两个实体之间的关系可能随着时间的变化而改变，过去存在的关系可能会在现在或未来的某一时刻解除，过去不存在的关系也可能会在现在或未来的某一时刻产生新的联系，实体间彼此存在的关系中可能会有着隐式存在的关系互联，也具有时效性。因此，我们考虑了关系具有时效性的特点，提高链接预测中受时间影响较大的关系预测的准确性。在时间线上，可以看到取出固定时间片，实体及其关系在不同时间作用下都是不同的，如图4：

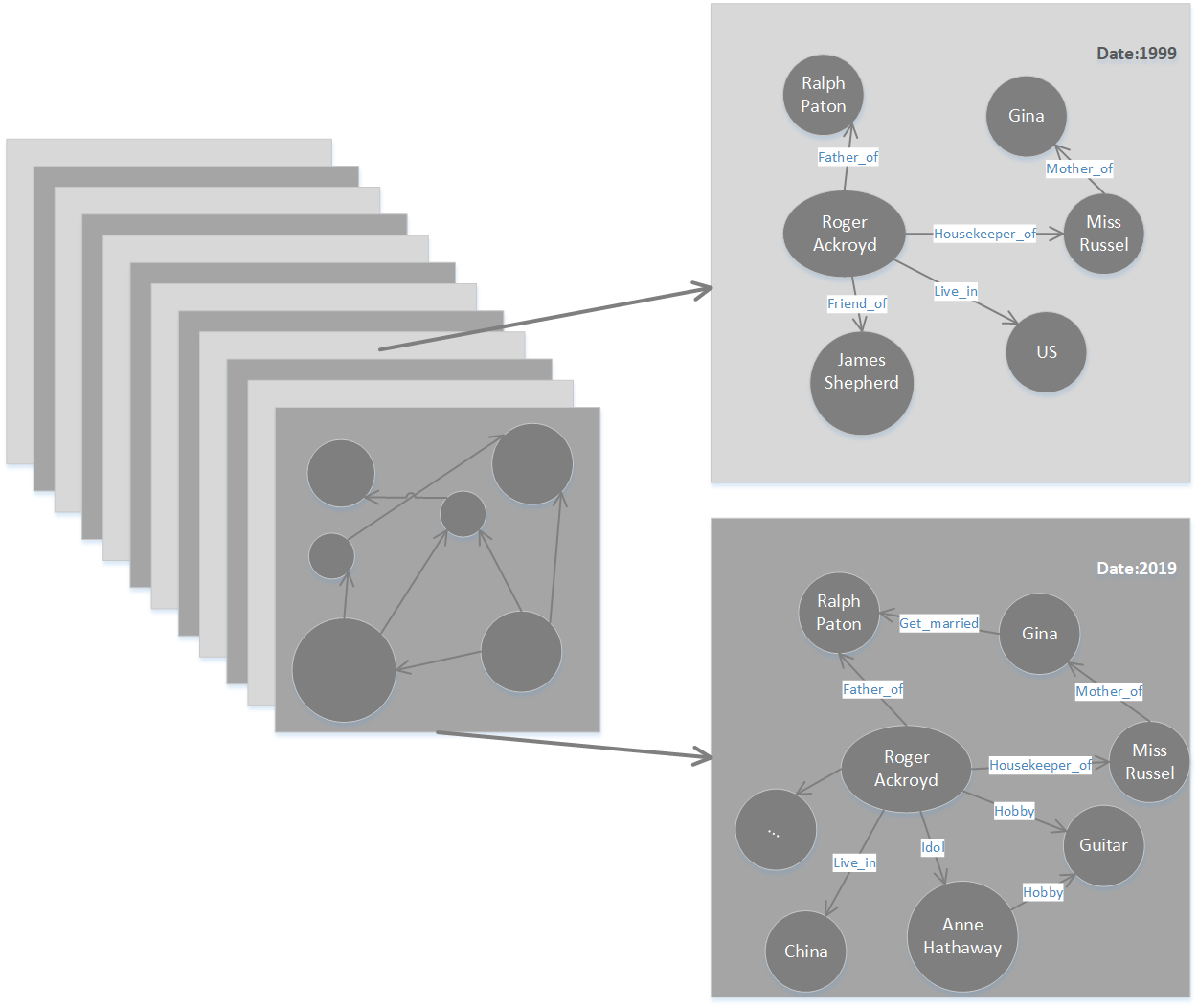


图4 链接的时效性

在TKG中，边真实存在的四元组概率值叫做显式影响力。然而，在实体—关系的组合中也会存在一种不可见的隐含关系，可以由已知关系推出。我们将不可观察到的变量定义为随机值，然后进行概率推理。当我们需要模型的概率分布的时候，为了符合参数估计的判断标准，最大似然估计是获得一致性最好的方法。我们基于优化后的贝叶斯结合最大似然法统计模型，计算存在的隐含关系预测的概率。

在数据集中，可表示为变量集合，其中变量取值集合表示为。含有四元组数目为，第个样本里有个属性，其属性矩阵为。根据公式（）可得显式链接成立的概率，含有的四元组数目为，则隐式链接成立概率为，含有四元组数目为。对于正确四元组预测概率为

P()=

随着时间的动态变化，新证据不断涌现，我们需要不断更新预测结果。当采用线性假设映射时，我们在得分函数方法[27]的基础上，定义一个改进的贝叶斯得分函数，对每个四元组进行可信度测量，以及实体关系的时间四元组进行预测，衡量链接的可能性，作为链接是否成立的评价指标：

|  |  |  |
| --- | --- | --- |
|  |  | () |

其中表示知识图谱在迭代过程中的一个中间过渡状态，使用得分函数是链接预测过程中的一次更新依据。每一次链接预测的迭代过程都将得分函数作为评价标准，找到正确的链接路径实现知识图谱补全。

3 实验

我们将T2LP算法应用在链接预测上，该模型的目的是完成对于缺失了head、tail或rel的四元组<head, rel, tail, time>的预测出对应的entity and relation。

3.1 数据集

我们使用了HyTE提取好的YAGO和Wiki数据，并将数据<head, rel, tail, start\_time, end\_time>处理成<head, rel, tail, start\_time>, < head, rel, tail, start\_time+1>,…, < head, rel, tail, end\_time>的形式，对于不完整数据我们没有采取自动补齐策略，以此来排除噪声数据的干扰。

整理后，在YAGO数据集中共10623个实体和10种关系，抽取得到了82550个训练集和8046个验证集；在Wiki数据中共有12554个实体和24种关系，最终得到了201100个训练样本和18208个验证集，如表1所示。

|  |  |  |
| --- | --- | --- |
|  | **wiki** | **yago** |
| **entity** | 12554 | 10623 |
| **relation** | 24 | 10 |
| **test** | 15115 | 5791 |
| **train** | 201100 | 82550 |
| **triple2id** | 235066 | 96387 |
| **valid** | 18208 | 8046 |

Table 1: 数据数量表

3.2 方法

本文实验的主要思想是首先是进行实体和关系抽取，构造出带有时间维的数据四元组。通过公式(2里的几)的余弦相似函数判断关系的相似度，从而将两个关系相近但只有其中一组含有时间信息的三元组构造成一组有时间维的四元组；然后根据数据集中的四元组概率值及关系的影响因子判断关系的先后顺序，通过word2vec将其映射成向量。对同一实体，与当前时间相比，时间早的关系值小，时间晚的关系值大。

3.3 评价指标

MeanRank，Hits@10标准普遍被用于KG预测准确度的衡量，这两个下降表示准确度的提高。

准确率：；召回率：

3.4 实验结果与分析

文中主要基于WordNet、Freebase两个数据集进行了验证，

模型改进技术包括batch normalization、weight decay、dropout等防止过拟合，增加模型的泛化能力。变化学习率通过在训练过程中递减学习率，使得模型能够更好的收敛，增加模型的拟合能力。通过加深模型层数和解决梯度衰减问题，增加模型的拟合能力。这些改进方法使网络的拟合能力和泛化能力越来越强，最终获得更高的准确率。

4 总结和展望

知识图谱旨在描述客观世界概念、实体、事件及其之间关系，具备可解释性，而且可以用于解决复杂决策问题。本文针对动态变化的知识图谱中结构不完整、链接随时间变化的问题，提出了四元组和T2LP模型，结合衰减函数和得分函数的统计过程，通过判断相似性实现动态的链接预测，并解决隐藏链接预测未来关系，在已有关系的基础上进行相关推理，实时更新动态环境下的时间演进知识图谱。实验结果表明，用T2LP模型进行时间上的动态链接预测是快速且有效的。下一步工作将重点研究基于分布式本文方法，并将我们的发现应用于其他方式的链接预测中，如在本文基础上添加空间相关实体约束，挖掘时间或空间维度上给事件带来的额外信息。

Reference

[1]Xu ZL, Sheng YP, He LR, Wang YF. Review on knowledge graph techniques. Journal of University of Electronic Science and Technology of China, 2016,45(4):589-606.

[2]Quillan MR. Semantic Memory. Bolt Beranek and Newman Inc., 1966.

[3]Wang YQ. Principles and Methods of Artificial Intelligence. Xi’an: Xi’an Jiaotong University Press, 1998 (in Chinese).

[4]Handbook of Knowledge Representation. Elsevier, 2008.

[5]Post EL. Formal reductions of the general combinatorial decision problem. American Journal of Mathematics, 1943,65(2):197-215.

[6]Ehrlinger L, Wöß W. Towards a definition of knowledge graphs. In: Proc. of the Semantics. 2016.

[7]Liu Qiao，Li Yang，Duan Hong，et a1. A survey of knowledge graph construciton technology[J].Computer Research and Development,2016,53(3):582-600.

[8]徐增林, 盛泳潘, 贺丽荣, 等. 知识图谱技术综述[J]. 电子科技大学学报, 2016, 45(4):  589-595.

[9]Decker S , Melnik S , Van Harmelen F , et al. The Semantic Web: the roles of XML and RDF[J]. IEEE Internet Computing, 2000, 4(5):63-73.

[10]Berners-Lee, Tim et al. “The Semantic Web" in Scientific American.” ,2001.

[11]Bordes A, Gabrilovich E. Constructing and mining Web-scale knowledge graphs: KDD 2014 tutorial[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, Aug 24-27, 2014. New York: ACM, 2014:1967-1967.

[12]Trivedi R , Dai H , Wang Y , et al. Know-Evolve: Deep Temporal Reasoning for Dynamic Knowledge Graphs[J]. 2017.

[13]Mutlu E C , Oghaz T A . Review on Graph Feature Learning and Feature Extraction Techniques for Link Prediction[J]. 2019.

[14]Nickel M, Murphy K, Tresp V, et al. A review of relational machine learning for knowledge graphs[J]. Proceedings of the IEEE, 2016, 104(1): 11-33.

[15]Bordes, Antoine et al. Translating Embeddings for Modeling Multi-relational Data[J]. NIPS, 2013.

[16]Wang Z , Zhang J , Feng J , et al. Knowledge Graph Embedding by Translating on Hyperplanes[C]// Twenty-eighth Aaai Conference on Artificial Intelligence. AAAI Press, 2014.

[17]Guoliang Ji, Shizhu He , et al. Knowledge Graph Embedding via Dynamic Mapping Matrix[J]. ACL, 2015.

[18]Xiao H , Huang M , Hao Y , et al. TransA: An Adaptive Approach for Knowledge Graph Embedding[J]. Computer Science, 2015.

[19]Xiao H , Huang M , Hao Y , et al. TransG : A Generative Mixture Model for Knowledge Graph Embedding[J]. Computer Science, 2016.

[20]He S , Liu K , Ji G , et al. Learning to Represent Knowledge Graphs with Gaussian Embedding[C]// the 24th ACM International. ACM, 2015.

[21]Ji G , Liu K , He S , et al. Knowledge Graph Completion with Adaptive Sparse Transfer Matrix[C]// Thirtieth Aaai Conference on Artificial Intelligence. AAAI Press, 2016.

[22]Shib, et al. HyTE: Hyperplane-based Temporally aware Knowledge Graph Embedding[J]. Conference on Empirical Methods in Natural Language Processing, 2018: 2001-2011.

[23]Nickel M , Murphy K , Tresp V , et al. A Review of Relational Machine Learning for Knowledge Graphs[J]. Proceedings of the IEEE, 2016, 104(1):11-33.

[24]Miller, George A. WordNet: a lexical database for English[J]. Communications of the Association for Computing Machinery, 1995, 38(11):39-41.

[25]Bollacker K D, Evans C, Paritosh P, et al. Freebase: A collaboratively created graph database for structuring human knowledge[C]// Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008. ACM, 2008.

[26] Deriving Validity Time in Knowledge Graph

[27] 翟社平, 郭琳, 高山, et al. 一种采用贝叶斯推理的知识图谱补全方法[J]. 小型微型计算机系统, 2018, v.39(05):133-137.

[36]袁凯琦, 邓扬, 陈道源, 等. 医学知识图谱构建技术与研究进展[J/OL]. 计算机应用研究,  2018(07): 1-11.

[37]中医药知识图谱构建与应用[J]. 医学信息学杂志, 2016, 37 (4): 8-13.

[38]SNOMED International[EB/OL]. http://www.snomed.org/, 2015-05-06/2018-04-17.

[39]IBM[EB/OL].http://www935.ibm.com/industries/healthcare/index.html, 2015-04-10/2018-04-17.