计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2017.20170640
54(12):2687-2697,2017

一种面向临床领域时序知识图谱的链接预测模型

陈德华[1] 殷苏娜[1] 乐嘉锦[1] 王 梅[1] 潘 乔[1] 朱立峰[2]

[1](东华大学计算机科学与技术学院 上海 201600)

[2](上海交通大学医学院附属瑞金医院 上海 200025)

(chendehua@dhu.edu.cn)

A Link Prediction Model for Clinical Temporal Knowledge Graph

Chen Dehua[1], Yin Suna[1], Le Jiajin[1], Wang Mei[1], Pan Qiao[1], Zhu Lifeng[2]

[1](*College of Computer Science and Technology, Donghua University, Shanghai* 201600)

[2](*Ruijin Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai* 200025)

Abstract Link prediction on knowledge graph is the main task of knowledge base completion, predicting whether a relationship existing in the knowledge base is likely to be true. However, traditional knowledge link prediction models are only appropriate for static data rather than temporal knowledge base. Temporal knowledge base exists on various fields. Take medical medicine field as example, diabetes is a typical chronic disease which evolves slowly. Thus, link prediction on clinical knowledge base such as diabetic complication requires the analysis on temporal characteristic of temporal knowledge base, which is a great challenge for traditional link prediction models. Thus, to address the prediction of temporal knowledge base, this paper proposes a long short-term memory (LSTM) based model for temporal knowledge base. The proposed model adopts memory cells of LSTM for sequential learning, and then builds incremental learning layer. Afterwards, timing characteristics can be extracted by the way of end-to-end, which realizes the prediction on temporal knowledge base. In experiments, the proposed model in clinical temporal knowledge base shows significant improvements compared with baselines including Rescal, NTN, TransE, TransH, TransR and DNN.

Key words temporal knowledge graph; knowledge graph link prediction; translation model TransR; long short term memory (LSTM) networks; incremental learning

摘 要 知识图谱(knowledge graph)链接预测可以解决知识图谱中缺失信息的发现和还原,是目前知识图谱领域的研究热点.传统的知识图谱链接预测方法大多面向静态的数据,并不适用于具有动态变化特性的时序知识图谱.时序知识图谱广泛存在于不同领域中,以临床医学领域为例,糖尿病作为一种典型的慢性病,其病程是一个疾病缓慢发展演化的过程.因此,在临床医学时序知识图谱上进行临床意义的链接预测,比如预测糖尿病的并发症,则需要考虑糖尿病病程发展随时间变化的时序特性,这也为传统的知识图谱链接预测方法带来巨大挑战.为此,结合临床医学事实知识的时序特性,提出一种基于

收稿日期:2017-09-01;修回日期:2017-10-06

基金项目:上海市科技创新行动计划项目(15511106900);上海市科技发展基金项目(16JC1400802);上海市信息化发展专项基金项目(XX-XXFZ-01-14-6349)

This work was supported by the Shanghai Innovation Action Project of Science and Technology (15511106900), the Science and Technology Development Foundation of Shanghai (16JC1400802), and the Shanghai Specific Fund Project for Information Development (XX-XXFZ-01-14-6349).

LSTM 序列增量学习的临床领域时序知识图谱链接预测模型. 该模型结合 LSTM 长短期记忆单元递归神经网络在序列学习上的优势,通过构建基于 LSTM 的序列增量学习层,以端到端的方式提取时序知识图谱中的三元组时序特征,从而实现对时序知识图谱的链接预测. 通过在糖尿病时序知识图谱上的实验,验证了模型的高效性、可用性及稳定性.

关键词 时序知识图谱;知识图谱链接预测;转换模型 TransR;长短期记忆网络;增量学习

中图法分类号 TP391

知识图谱(knowledge graph)是表示知识的一种新方法,属于语义网络范畴,用于描述真实世界中存在的各种实体和概念以及这些实体、概念之间的关联关系,捕捉并呈现特定领域概念之间的语义关系[1]. 近年来知识图谱在医学领域也逐渐得到重视和关注,国内外均开展了医学领域知识图谱相关研究. 国外有牛津大学创建的用于药学的 LynxKB 知识图谱[2]以及由日本东北大学将传统关系型数据库融合于知识图谱中进行基因研究[3];与此同时,国内医疗信息学领域也提出了多种医学知识图谱,包括中国中医科学院中医药信息研究所构建的中医药知识图谱[4]、基于知识图谱的基因组流行病学可视化分析[5]等. 然而,这些医学知识图谱的知识来源主要是公开的医学文献,较少涉及医院的实际电子病历(electronic medical record, EMR)数据.

众所周知,EMR 电子病历[6]记录着患者在医疗活动中产生的各种临床事实数据,蕴含着丰富的临床事实知识,主要体现为各种医学实体如患者实体、药物实体、诊断实体等,以及医学实体之间存在的各种联系. 本文利用知识图谱表示 EMR 中临床事实知识,构建基于 EMR 的临床领域时序知识图谱,刻画临床数据中存在的实体和概念,提供具体且丰富的语义和时序信息,以便更准确地揭示实体之间的内在联系,从而避免来自不同数据源信息的语义异构.

知识图谱链接预测[7]是知识图谱学习与推理的重要应用,其主要任务是对知识图谱中实体间可能存在的关系进行预测,实现知识图谱中缺失信息的发现和还原[8]. 由于实际电子病历数据普遍存在数据质量不高的特点,使得基于 EMR 的临床领域知识图谱中可能存在着一些医学实体以及实体间关系的缺失,或者实体间存在错误的关系. 通过对临床领域知识图谱的链接预测,能够将这些关系所补全或者纠正出错误的关系,从而获得更为完整和真实的知识图谱.

目前成熟的知识图谱链接预测包含张量分解模型[9]、NTN 神经网络[10]、转换模型[11]等. 然而这些预测模型都只在通用知识图谱上取得了不错的效果. 通用知识图谱中大多为常识性知识,并不随时间而改变. 与此相反,在临床领域中一般疾病的病程发展是一个缓慢演变的过程,可见临床事实知识具有时效性,包含大量时序知识. 以糖尿病为例,在糖尿病患者临床诊治过程中,每次的血糖检查、糖化血红蛋白检查、用药情况、并发症诊断结果等均有时间的标记. 这些临床事实知识可按照时间前后顺序,转换形成具有时序特性的临床领域时序知识图谱. 但是,现有的知识图谱链接预测模型大多针对静态的数据,而未考虑到时序知识图谱中蕴含大量时序信息,无法对时序知识图谱做出准确的预测.

为解决上述问题,本文从医院实际的 EMR 数据出发,结合临床医生的经验与知识,建立临床领域时序知识图谱,并且提出了一种基于 LSTM 序列增量学习的临床领域时序知识图谱链接预测模型. 该模型采取 LSTM 长短期记忆单元的递归神经网络的序列学习能力,并创建序列增量学习层对临床事实知识时序特征进行提取,同时通过端到端(end-to-end)的方式进行知识图谱三元组序列的增量学习过程,从而实现对临床领域时序知识图谱的链接预测. 本文通过多种对比实验,从准确度、召回率和精准度等方面对增量 LSTM 新模型进行了评估验证;同时分析了新模型的时间复杂度,最终验证了新模型在时序知识图谱链接预测上具有较好的性能.

1 相关工作

近年来,业界陆续提出了多种不同的通用知识图谱,比如 2012 年 Google 公司推出 Google Knowledge Graph[12],之后又提出了多种通用知识图谱,如 FreeBase[13], DBPedia[14], WordNet[15]等. 而对临床领域,国外较为流行的临床领域知识库有 MorphoCol[16], Nursing KB[17]等,并基于此进行了临床决策诊断支持工作[18]. 国内探索了中医药知识

图谱构建[4]方法以及基于医药知识图谱推理的辅助开药[19].之后,有关时间信息在知识图谱中的重要性被逐渐关注,比如对含有时间的知识进行知识图谱模型建立[20].

链接预测一直是知识图谱学习和推理的热点问题,许多研究者提出了不同的链接预测模型,用于学习和预测实体间存在的关系.现有成熟的知识图谱链接预测方法可分为3类:1)基于张量分解的知识图谱链接预测方法,包括Rescal、神经张量网络(NTN)等;2)基于向量转换模型的知识图谱链接预测方法,比如说TransE[21],TransH[22],TransR[23]等;3)以深度学习[24]为代表的知识图谱链接预测方法异军突起.比如文献[25]尝试使用深度神经网络进行通用知识图谱的链接预测.然而这些方法目前只适用于知识图谱的静态数据,还不适用于具有动态变化特性的时序知识图谱链接预测.

关于时间信息在知识图谱链接预测中的应用,文献[26]提出了基于TransE转换模型改进的TransE-TAE模型,通过对知识图谱中的时间信息分析,对关系作出了预测;文献[27]进而研究由时间导致的不确定性知识图谱上的预测模型.然而,大多数工作都集中在具有时效性的通用知识图谱中,缺乏对临床领域知识图谱中尤为突出的时序特征做研究.

2 临床领域时序知识图谱及其链接预测

2.1 临床领域时序知识图谱相关定义

临床领域时序知识图谱基于实际EMR数据构建而成,其形式化定义如下.

定义1. 临床领域时序知识图谱G.临床领域时序知识图谱为一张有向标签图$G_t=(t_0,t_e,E,R,\tau)$,其中E为知识图谱的顶点集,用于表示实体集合;R为知识图谱的边集,用于表示事实关系集合;τ为$E\times E\rightarrow R|k$的函数,表示知识图谱中的所有元组.k表示在时间段$[t_0,t_e]$按照时间前后排序的知识图谱三元组列表中,两实体之间存在第k次的关系R.举例来说,现有一个2015—2016年的临床领域时序知识图谱$G_t=(2015\text{-}01\text{-}01,2016\text{-}01\text{-}01,E,R,\tau)$,则病人张三的血糖检测三元组序列$L(t_0,t_e,\tau)=\{$(张三,血糖检查|第1次检查,正常),(张三,血糖检查|第2次检查,异常偏高),(张三,血糖检查|第3次检查,异常偏高),…$\}$.

图1所示为临床领域时序知识图谱.由图1可见,临床领域时序知识图谱由概念层和实例层两部分组成.

其中概念层包括患者、疾病、检验指标、并发症及药品等实体类型.概念层中,不同实体类型之间存在概念层之间的关系.在图1所示的知识图谱中,实体概念层次可分为3个层级:一级实体概念包括基本信息、患者、检验报告、诊断和用药等;二级实体概念包括糖尿病诊断和并发症诊断;其他为三级实体概念.

在实例层中,每个实体都含有自己的属性以及属性值.例如"张三"是属于患者实体类,因此,将"张三"实例化为患者实体的姓名属性值,张三患者实体的医疗卡号"113"作为实例则对应于患者中的医疗卡号属性值.

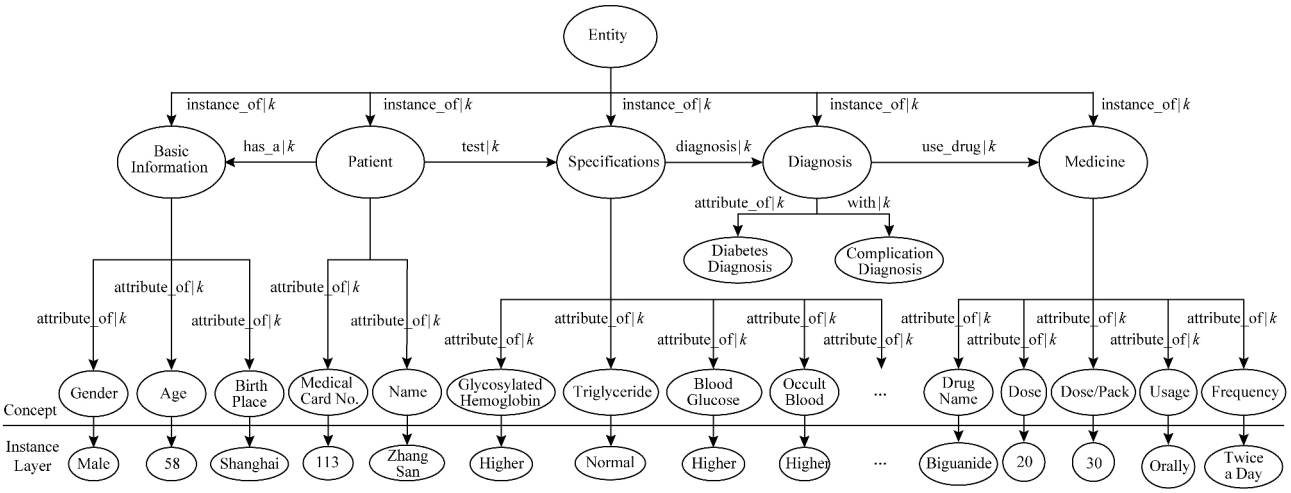


Fig. 1 Clinical domain temporal knowledge graph model

图1 临床领域时序知识图谱模型

2.2 时序链接预测相关定义

定义 2. 时序链接预测. 时序链接预测是指在临床领域时序知识图谱 G 中，通过对已知信息的分析，对图谱中 2 个实体 E_1 和 E_2，预测出二者之间可能存在临床意义的关系 R.

例 1. 以图 1 所示的时序知识图谱为例，给出时序链接预测实例：由于糖尿病患者在就诊过程中有多次指标检测等，因此有如表 1 所示的糖尿病时序知识图谱中的三元组序列 $X^{(i)}$ 作为输入，经过链接预测，可以预测出该患者实体与眼病实体之间是否具有患有关系 Y，即为输出. 整个预测过程可以表示为

$$P=\{X,Y\}=\{X^{(1)},X^{(2)},\cdots,X^{(n)},Y\}, \quad (1)$$

其中，$X^{(i)}$ 描述有关该病人实体的各类关系所对应的三元组序列，可视为病人各方面的属性：

$$X^{(i)}=\{X_1^i,X_2^i,\cdots,X_n^i\}, \quad (2)$$

Table 1 Temporal Knowledge Base Link Prediction

表 1 临床领域时序知识图谱链接预测

Variables	Description	Values
$X^{(1)}$	Sequence of Blood Sugar Testing	$(patient,blood\_sugar\_test\|1,value)$ $(patient,blood\_sugar\_test\|2,value)$ ⋮
$X^{(2)}$	Sequence of Insulin Testing	$(patient,insulin\_test\|1,value)$ $(patient,insulin\_test\|2,value)$ ⋮
$X^{(3)}$	Sequence of leukocyte Testing	$(patient,leukocyte\_test\|1,value)$ $(patient,leukocyte\_test\|2,value)$ …
⋮	⋮	⋮
$X^{(n)}$	Sequence of Drug Use	$(patient,use\|1,drug)$ $(patient,use\|2,drug)$ …
y	Complication with Eye Disease	(patient, has/no complication diagnosis, Eye Disease)

3 基于 LSTM 序列增量学习的时序链接预测

本节首先阐述临床领域时序知识图谱链接预测模型整体框架，然后具体阐述模型的细节，最后给出模型的训练过程.

3.1 模型的整体框架

本文提出基于 LSTM 序列增量学习的临床领域时序知识图谱链接预测模型，用来推理预测带时间的临床领域知识图谱中各个实体之间的链接. 图 2 所示为模型结构图，一共包括 4 层：三元组向量化层(输入层)、LSTM 序列增量学习层、序列特征组合层以及分类层(输出层).

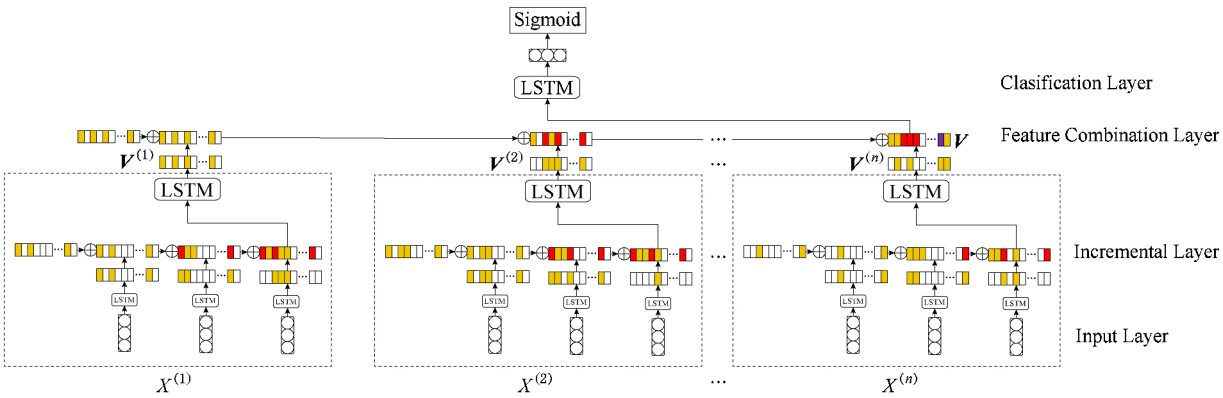


Fig. 2 Temporal link prediction model based on incremental LSTM sequence learning

图 2 基于 LSTM 序列增量学习的时序链接预测模型

模型的输入是临床领域时序知识图谱中的三元组序列，输出是实体之间的关系预测结果. 模型训练过程主要是：在得到对应的序列化三元组 $x^{(n)}$ 之后，首先用 TransR 进行向量化作为输入层；其次输入到 LSTM 增量学习层，将得到的向量做增量计算；再将增量之后的向量 $\boldsymbol{V}^{(n)}$ 输入到 LSTM 序列特征组合层计算得到 $\boldsymbol{V}$，最后进入分类层.

3.2 基于 TransR 的多语义三元组向量化

基于 TransR 的多语义三元组向量化层为时序链接预测模型的第 1 个层次. 本层主要是采用 TransR 转换模型将临床领域时序知识图谱 G 中的三元组$\langle E_i,R,E_j\rangle$嵌入到低维空间内. 在临床领域时序知识图谱中存在多对多且语义不同的关系，比如多个患者实体和多种不同的检查指标实体之间都是检查关系，同时这些检查关系里有的是超声检查关系，有的是穿刺检查关系. TransR 模型支持对不同实体拥有不同语义空间的处理，这符合临床数据中关系来自不同语义空间的特点；与此同时，在

TransR 中,首先将各个实体向量向关系空间中做投影,因此原来在实体空间中相似的实体就被区分开来了,从而在临床事实知识图谱中实现了对多对多关系两边不同实体的区分,并将实体和关系嵌入到低维向量.

TransR 模型为了考虑不同语义空间,对于多对多关系有更精确的向量化表示,在 TransE 模型的基础上对实体向量向关系空间中进行投影,然后建立从头实体到尾实体的转换关系.

图 3 为 TransR 翻译模型运用在本文临床事实知识图谱上的一个简单例子:张三和上海 2 个实体通过在_born_in 关系空间上做投影,从而嵌入到向量坐标中,两者间建立起被映射的_born_in 向量转换关系.

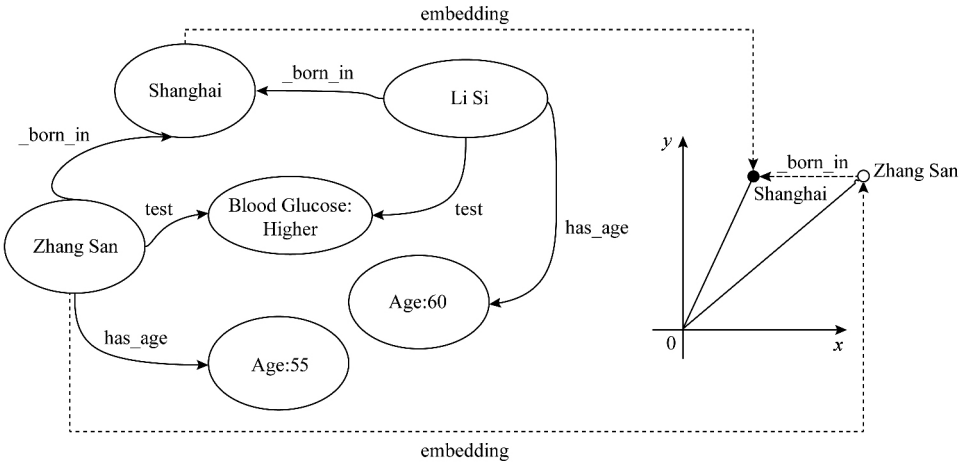


Fig. 3 Embeddings from Knowledge Graph

图 3 知识嵌入图

TransR 中对于每个关系,都定义了一个投影矩阵 $\boldsymbol{M}_r$,将实体向量从实体空间投影到关系 r 的子空间,$\boldsymbol{l}_{hr}$ 和 $\boldsymbol{l}_{ht}$ 表示为

$$\begin{aligned} l_{hr} &= l_h \boldsymbol{M}_r, \\ l_{tr} &= l_t \boldsymbol{M}_r. \end{aligned} \tag{3}$$

对应的损失函数为

$$f_r(h,t) = \| l_{hr} + l_r - l_{tr} \|_{L_1/L_2}. \tag{4}$$

因此,在 TransR 模型中,将每个实体都看作不同属性构成的,对于不同的关系,关注的是实体的不同属性;并且将有着不同语义空间的关系作出投影,从而区分多种语义,由此对临床领域知识图谱实现高效的向量化.

3.3 基于 LSTM 的三元组序列增量学习

基于 LSTM 的三元组序列增量学习层为时序链接预测模型的第 2 个层次. 递归神经网络模型 LSTM[28]是一种可以学习长期依赖信息的神经网络,其输入数据的形式记为 $f=\{Xn\}, n=t_1, t_2, \cdots, t_k$ 该输入数据是带序列性质的数据向量. 临床领域时序知识图谱通过 TransR 向量化后,输出为具有 2 个特点的三元组向量:1)临床数据在向量化后尽量不丢失原有的语义信息;2)向量化后输出的三元组按照时序排列,具有时序性. 因此,本文将 TransR 后的三元组向量送入 LSTM 中,不仅保持了原有语义,同时也具有 LSTM 输入数据 f 的时序特征,从而适用于 LSTM. 所以,经过 TransR 后的语义三元组与 LSTM 的叠加增强作用,记忆单元可利用序列中的历史信息,从而能较充分且准确地挖掘序列之间的依赖信息.

目前,使用最广泛的 LSTM 单元有 3 个门:输入门、输出门和遗忘门,以保存历史信息. 其中,输入门用于控制当前数据输入对记忆单元状态值的影响,遗忘门用于控制历史信息对当前记忆单元状态值的影响,通过计算得到记忆单元状态;输出门用于控制记忆单元状态值的输出.

针对临床领域时序知识图谱中存在于$\langle T_1, T_2\rangle$时间段的三元组序列作为输入,按照式(3)计算得到输出:

$$\begin{aligned} L_{\text{input}} &= \{\langle E_i, R, E_j, t\rangle, t \in (T_1, T_2)\}, \\ X_t &= \{\langle E_i, R, E_j, t'\rangle, t' = t\}, \\ f_t &= \sigma(W_f \times [h_{t-1}, x_t] + b_f), \\ i_t &= \sigma(W_i \times [h_{t-1}, x_t] + b_i), \\ C'_t &= \tanh(W_c \times [h_{t-1}, x_t] + b_c), \\ C_t &= f_t \times C_{t-1} + i_t \times C'_t, \\ o_t &= \sigma(W_o [h_{t-1}, x_t] + b_o), \\ h_t &= o_t \times \tanh(C_t), \end{aligned} \tag{5}$$

其中,X_t 作为输入,由遗忘门通过 σ 函数 sigmoid 来

控制,它会根据上一时刻的输出 h_{t-1} 和当前输入 X_t 来产生一个 0～1 的 f_t 值来决定是否让上一时刻学到的信息 C_{t-1} 通过或部分通过,即进行选择性的遗忘;i_t 用于决定输入信息的接受状态,接着由 tanh 层用来生成新的候选值 C'_t,它作为当前层产生的候选值可能会添加到状态中. 最后将这 2 部分产生的值结合起来进行更新.

在三元组$\langle h,R,t\rangle$输入到转换模型向量化之后,得到$\langle l_h,l_r,l_t\rangle$,然后将批量向量组成的矩阵输入到 LSTM 中,通过 LSTM 对上下文的前后分析,对三元组提取特征,输出更精确的特征向量$\langle V'_h,V'_r,V'_t\rangle$.

对于带时间序列的临床领域知识图谱,本文将 LSTM 改进为增量 LSTM. 该模型采用交叉熵函数作为损失函数,通过反向传播,对 LSTM 中的参数进行调整.

3.4 LSTM 序列特征增量组合

LSTM 序列特征增量组合为本文时序链接预测模型的第 3 个层次,在上下三元组序列中,采用增量形式代替简单的前后连接过程:当后一个向量与前一个向量在同一位置上有值时,即增量相加来强化特征,并将强化后的向量作为下一个的输入. 图 4 所示为 LSTM 序列特征增量组合示意图. 其中,每个长方形框对应向量中的每一位,颜色不同意味着每位的值不同. 如果是白色,代表此向量在该位上无数据;若是黄色,代表此位置上有数据;若为红色,代表该位置是经过了增量相加后的数据. 例如在时序数据中上一时间点的向量 $\boldsymbol{V}^{(1)}$ 和目前向量 $\boldsymbol{V}^{(2)}$ 的第 1 位都有数据,则需要对其进行增量相加,从而强化其特征,则该位置由黄色变成了红色. 算法 1 所示为具体的特征增量组合过程描述.

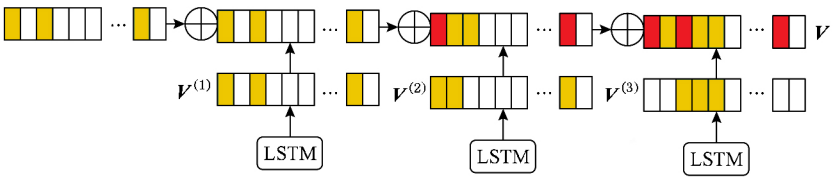


Fig. 4 Feature combination by incremental learning

图 4 通过增量进行特征组合

算法 1. LSTM 序列特征增量组合过程.

输入:LSTM 的输出向量 $\boldsymbol{V}^{(m)}$;

输出:序列增量组合后的特征向量 $\boldsymbol{V}$.

$\boldsymbol{V}$:特征组合层后的向量,$\boldsymbol{V}^{(m)}$:序列增量层中第 m 个 LSTM 产生的向量.

① $\boldsymbol{V}^{(0)}=\boldsymbol{V}^{(1)}$;

② for $i=0$ to n

③ $\quad\boldsymbol{V}^{(i)}=\boldsymbol{V}^{(i-1)}+\boldsymbol{V}^{(i)}$;

④ endfor

⑤ $\boldsymbol{V}=\boldsymbol{V}^{(n)}$.

3.5 时序链接预测模型的训练算法

经过上述 3 个层次处理之后,对每一类的三元组都通过增量计算进行上述的特征组合提取,直到得到最终的特征向量作为分类器的输入,整个训练过程采取反向传播进行调参. 具体的时序链接预测模型的训练算法如算法 2 所示.

算法 2. 时序链接预测模型训练过程.

输入:按照时间从前往后排序的三元组序列;

输出:0 或 1,1 代表该三元组为正确事实,0 则反之.

N_{it}:使用批数据集训练的次数;

D^{tr}/D^{te}:训练集/测试集;

M:序列增量层中 LSTM 的数量;

E:分类层中的损失值;

${}^{A}E_m$:序列增量层中第 m 个 LSTM 的损失值;

S:批数据集的数量;

${}^{A}\theta_m^i/{}^{C}\theta^i$:第 i 次训练中:序列增量层中第 m 个 LSTM 的参数/分类层中 LSTM 的参数.

① $({}^{A}\theta_1^0,\cdots,{}^{A}\theta_M^0,{}^{C}\theta^0)\leftarrow Initialize(M)$;

② for $i=1$ to N_{it}

③ $\quad$for $j=1$ to S

④ $\quad\quad D^{tr}\leftarrow GetMiniBatch(D)$;

⑤ $\quad\quad$for $m=1$ to M

⑥ $\quad\quad\quad\boldsymbol{V}^{(m)}\leftarrow LSTM(D^{tr},{}^{A}\theta_m^{i-1})$;

⑦ $\quad\quad\quad\boldsymbol{V}\leftarrow Incremental(\boldsymbol{V},\boldsymbol{V}^{(m)})$;

⑧ $\quad\quad$endfor

⑨ $\quad\quad E\leftarrow CLASSIFICATION\_LSTM(\boldsymbol{V},{}^{C}\theta^{i-1})$;

⑩ $\quad\quad({}^{C}\theta^i,{}^{A}E)\leftarrow BACKWARD(E,{}^{C}\theta^{i-1})$;

⑪ $\quad\quad$for $m=1$ to M

⑫ $\quad\quad\quad{}^{A}E_m=Separate({}^{A}E,m)$;

⑬ $\quad\quad\quad{}^{A}\theta_m^i=BACKWARD({}^{A}E_m,{}^{A}\theta_m^{i-1})$;

⑭ $\quad\quad$endfor

⑮ $\quad$endfor

⑯ $EvaluateModel(D^{te},{}^{A}\theta_1^0,\cdots,{}^{A}\theta_M^0,{}^{C}\theta^0)$;

⑰ endfor

4 实验结果与分析

4.1 实验数据

本文所用的数据来源于上海交通大学医学院附属瑞金医院实际 EMR 数据,该知识图谱由该院内分泌科近 10 年的 EMR 数据抽取而成,共有 61 000 个实体数、53 种不同的关系类型、训练集规模为 345 549 个三元组.

为了验证本文所提的三元组序列增量学习模型预测的效果,本文在上述临床领域时序知识图谱数据集上采用 10 折交叉验证方法进行评估,故验证集和测试集规模均随机选取 34 554 个三元组.

4.2 对比模型

本文采取知识图谱中常用的 6 种推理预测模型作为对比模型:

1) 张量分解 Rescal 模型. Luo 等人[29]采用张量分解的方法对临床健康数据进行分析,挖掘其隐藏知识. 主要是将知识图谱中的三元组转换成张量 $\boldsymbol{Y}$,如果三元组$\langle h,R,t\rangle$存在,则 $Y_{hrt}=1$,否则 $Y_{hrt}=0$. Rescal 算法将分解为实体和关系表示,以此得到低维向量表示,通过矩阵分解计算对张量进行分解得到预测结果.

2) 张量神经网络(neural tensor network, NTN)模型. 采用双线性张量取代传统神经网络中的线性变换层,通过将头、尾实体向量在不同的维度下的联系进行实体间预测. 其中,对于每组三元组$\langle h,R,t\rangle$,NTN 对其都有一个评分函数:

$$f_r(h,t)=\boldsymbol{u}_r^{\mathrm{T}}g(l_h\boldsymbol{M}_r l_t+\boldsymbol{M}_{r,1}l_h+\boldsymbol{M}_{r,2}l_t+b_r),\quad(6)$$

其中,$\boldsymbol{u}_r^{\mathrm{T}}$ 是一个与关系相关的线性层;函数 $g()$ 是 tanh 函数;$\boldsymbol{M}_r$ 是一个三阶向量,$\boldsymbol{M}_{r,1}$ 和 $\boldsymbol{M}_{r,2}$ 是与关系 r 有关的投影矩阵.

3) 转换模型 TransE. TransE 模型为转换模型的代表. 将知识库中的关系看作是实体之间的某种平移向量,对于三元组$\langle h,R,t\rangle$,TransE 模型将 $\boldsymbol{l}_r$ 表示关系 r 的向量,将 $\boldsymbol{l}_h$ 和 $\boldsymbol{l}_t$ 分别作为 h 和 t 的向量,$\boldsymbol{l}_r$ 可以作为 $\boldsymbol{l}_h$ 和 $\boldsymbol{l}_t$ 向量之间的平移,即将 $\boldsymbol{l}_r$ 看作是 $\boldsymbol{l}_h$ 和 $\boldsymbol{l}_t$ 的转换. 关系向量可以作为实体向量之间的平移,由此推测三元组的正确性.

4) 转换模型 TransH. TransE 模型在处理 1-N,N-1,N-N 复杂关系时,有着一定的局限性,TransH 模型提出来对复杂关系局限性的解决方案. TransH 模型将采用平移向量 $\boldsymbol{l}_r$ 和超平面的法向量 $\boldsymbol{W}_r$ 来表示,从而进行复杂关系的推理预测.

5) 转换模型 TransR. 对不同语义空间中实体与关系的推理预测模型. 具体请参见 3.2 节中的详细介绍.

6) 深度神经网络模型 DNN. 目前由 Taheri 等人提出的深度学习模型[25]成熟应用于 ConceptNet 知识库[30],该模型通过 Word2Vec 对知识库做向量化,接着通过双向 LSTM 联系上下文对向量做修正,最后通过深度神经网络(deep neural network, DNN)模型做分类预测.

4.3 评价指标

本文采用准确度(accuracy, AUC)、召回率(recall, R)和精确度(precision, P)作为模型评估的指标.

准确率指的是对于给定的测试数据集,模型正确预测的样本数与总样本数之比:

$$AUC=\frac{TP+TN}{TP+FP+FN+TN},\quad(7)$$

其中,TP 是指被模型预测为正的正样本,TN 是被模型预测为负的负样本,FP 是被模型预测为正的负样本,FN 是被模型预测为负的正样本.

同时,为了反映被正确判定的正例占总正例的比重,本文采取了召回率 R 作为评价指标,体现了模型对正样本的预测能力,召回率 R 越高,说明模型对正样本识别能力越强:

$$R=\frac{TP}{TP+TN}.\quad(8)$$

此外,精确度 P 体现了模型对负样本的区分能力,P 越高,说明对负样本的区分能力越强. P 指正确预测的正样本数占总正样本的比例:

$$P=\frac{TP}{TP+FP}.\quad(9)$$

因此,本文还采用 Precision-Recall(PR)图评估正负样本区分能力. 同时,$F1$ 分数可以看作是准确率和召回率的加权平均.

本文分别对 6 种参考模型和新模型做 AUC,R,$F1$-Score 进行分类性能对比,对 ROC,PR 图进行图对比,同时也分析了本文增量 LSTM 模型中的参数选择对比以及各模型的时间复杂度.

4.4 结果分析与讨论

本节阐述对比实验结果以及对结果的分析. 实验中将本文增量 LSTM 模型与其他参考模型相比较.

4.4.1 模型分类性能对比

首先,本文将本文模型与其他 6 种参考模型方法运用在临床领域时序知识图谱中,表 2 为这些方法的性能对比.

Table 2 Performance Comparison on Different Models

表 2 不同模型的性能对比

Model	*AUC*	*R*	*F*1-Score	Time Complexity
Rescal	0.55	0.58	0.57	$O(n_e\times k+n_r\times k^2)$
NTN	0.6	0.58	0.59	$O(n_e^2\times k_b+n_r\times(k_b+k_a)+2n_rkk_a+n_ek)$
TransE	0.56	0.52	0.53	$O(n_e\times k+n_r\times k)$
TransH	0.62	0.62	0.62	$O(n_e\times k+2\times n_r\times k)$
TransR	0.73	0.75	0.74	$O(P+n_e\times k+2\times n_r\times k)$
DNN	0.81	0.80	0.80	$O(\log V+n_0^3+k+n_1\times n_2+n_2\times n_3+\cdots+n_{r-1}\times n_r)$
Incremental LSTM	**0.85**	**0.87**	**0.86**	$(\boldsymbol{n_r+n+1})\times\boldsymbol{O(n^3)}$

Note: The bold words mean best choice.

从表 2 中可以看出,通过该实验,翻译模型中的 TransR 模型,在本文的临床事实知识图谱中,准确度要比 Rescal 算法和 NTN 算法高. 翻译模型中 TransE 和 TransH 模型,与 Rescal 和 NTN 相比,并无明显优势. 而深度学习模型 DNN 通过对隐藏层的控制表达,对比传统链接预测方法取得了较大的优势. 可见,针对临床领域时序知识图谱中语义丰富、时序性等特征,本文提出的增量 LSTM 模型取得了最高的准确度、召回率和 *F*1-Score,分类性能有了明显提升.

4.4.2 模型 ROC 图对比

ROC 曲线图以真阳性率 *TP* 作为纵轴,以假阳性率 *FP* 作为横轴. 由图 5 可知,本文提出的增量 LSTM 模型 ROC 曲线下的面积最大,同时最凸,最靠近左上点,表明在这 3 类模型中新模型的诊断价值最大,准确性最高,利用价值大.

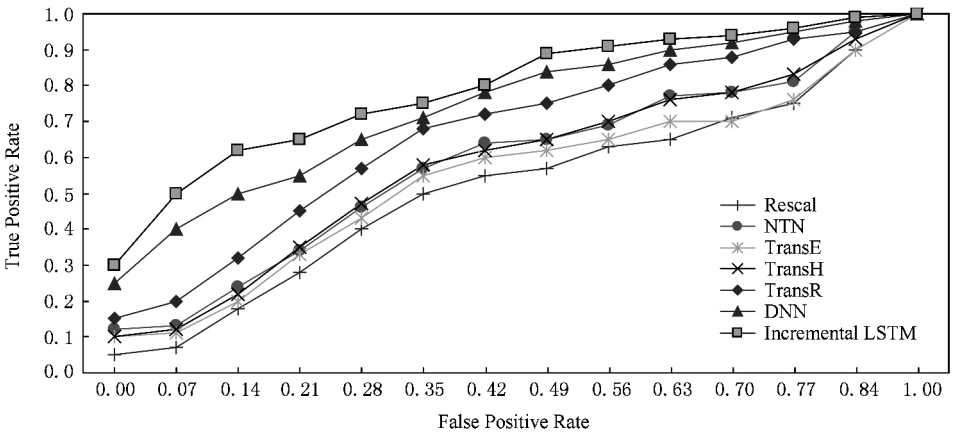


Fig. 5 ROC results on different models

图 5 ROC 对比图

4.4.3 模型 PR 图对比

本文对这 3 个模型还做出了 PR 曲线图的对比. 在 PR 曲线中,以召回率 *R* 为横坐标,*P* 为纵坐标. 召回率表明的是查全率,精确度表明的是查准率,两者不可同时兼得,一般 *R* 高,*P* 就低,反之 *R* 低,*P* 就高. 因此,往往通过 PR 曲线去看它们之间的关系和权衡点. 与 ROC 曲线左上凸不同的是,PR 曲线中,越右上凸的曲线,说明该模型的效果越好.

由图 6 的各个模型的 PR 图对比,可以看出,相比于其他模型,增量 LSTM 模型在 PR 图上的表现有了比较明显的提高,右凸程度比原先大了很多,说明该模型的效果有了提升,是 7 种模型中右凸程度最明显的,因此,本文提出的增量 LSTM 效果最佳.

通过从以上准确度 *AUC*、召回率 *R* 以及 ROC 曲线和 PR 曲线对比图方面进行分析,综合得出本文提出的增量模型相比原先的模型,在各个方面都

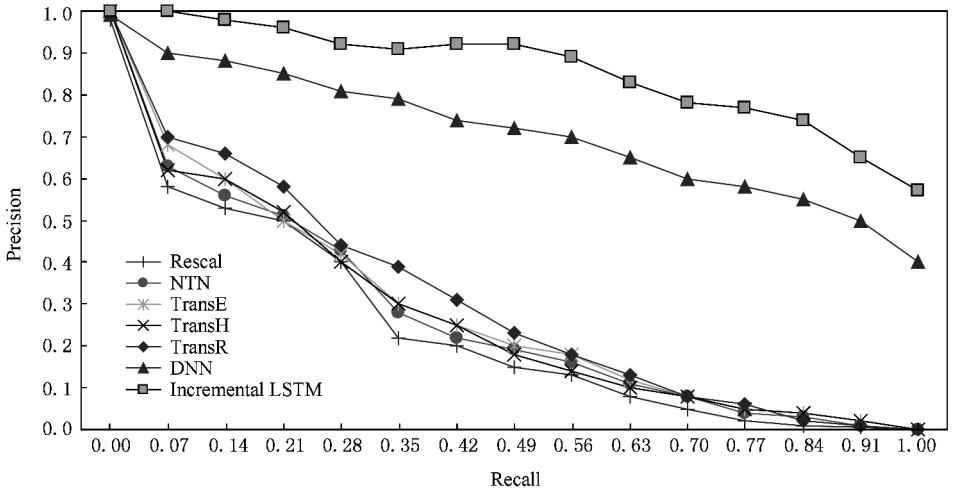


Fig. 6 PR results on different models

图 6 PR 对比图

有了显著的提升.

4.4.4 模型中网络结构对比

本文对增量 LSTM 模型中 LSTM 神经元个数进行了不同参数对比,表 3 为不同参数对最终性能的影响对比表.

Table 3 Comparison on Different Network Parameters

表 3 不同网络参数的对比

Numbers of Neurons in LSTM	*AUC*	*R*	*F*1-Score
8	0.58	0.58	0.57
16	0.62	0.64	0.63
32	0.69	0.67	0.68
64	0.74	0.72	0.73
128	**0.85**	**0.87**	**0.86**
256	0.80	0.81	0.81

Note: The bold words mean best choice.

随着 LSTM 中神经元数量的增加,增量 LSTM 模型在准确率、召回率以及 *F*1 分数上有了明显提高.然而,当神经元增加到 256 时,性能反而有所下降,其中的原因可能是因为过拟合.因此,本文在进行了模型参数对比实验后,最终选取 128 个 LSTM 神经元个数.

4.4.5 模型时间复杂度分析

表 2 中包含本文提出模型和其他模型的复杂度对比.其中 n_e 指实体的数量,n_r 指关系的数量,k 指向量化中向量的维度,K_a,k_b,k_c 分别对于 NTN 中的第 a,b,c 层的大小,n_r 是第 r 层神经网络中的神经元个数,P 代表 TransR 中做投影的时间,在 DNN 模型中,V 是通过 One-hot 方式初始化稀疏向量的维度,n_0 是双向 LSTM 中输入层的序列数,n_r 是第 r 层神经网络中的神经元个数,n 表示增量 LSTM 中输入的三元组序列数.

从表 2 中可见,Rescal 的复杂度最高,需要的时间消耗最大.NTN 方法的时间消耗也过长,若运行在更大的数据量中,时间限制则会加剧.而基于转换模型的时间复杂度大大降低,适于大数据量.增量 LSTM 模型的时间与输入量联系紧密,知识图谱中关系种类越多,输入的三元组序列数越大,时间复杂度即越高.

5 总 结

本文提出了一种基于深度学习的临床领域时序知识图谱链接预测模型.该模型用于以医院内部实际的 EMR 记录为基础,所创建具有时序特性的临床领域时序知识图谱.该模型选取适合大规模数据的 TransR 转换模型,在包含不同语义的关系空间中做实体投影,从而对图谱中的实体和复杂语义关系向量化.然后,采用 LSTM 递归神经网络,加入了图谱中的上下关联信息,进行序列化学习.接着对时序信息做增量计算,对时序信息提取更精准的特征向量.最后,不断通过增量计算和 LSTM 递归网络进行深层学习,提高预测准确度.实验表明:增量 LSTM 模型突出临床事实中隐含的语义和时序信息,有效地利用序列化学习挖掘其前后依赖信息,弥补了传统链接预测模型导致对时效性知识图谱预测准确度较低的不足.在未来的工作中,考虑将本文所提框架下的 LSTM 替换为其他 LSTM 变种方法,

进一步集成其他深度学习方法,从而优化增量LSTM模型.

参 考 文 献

[1] Sequeda J F. Integrating relational databases with the semantic Web: A reflection [C] //Proc of the 13th Int Summer School. Berlin: Springer, 2017: 68-120

[2] Sulakhe D, Balasubramanian S, Xie Bingqing, et al. Lynx: A database and knowledge extraction engine for integrative medicine [J]. Nucleic Acids Research, 2014, 42(D1): 1007-1012

[3] Ogishima S, Takai T, Shimokawa K. Integrated database and knowledge base for genomic prospective cohort study in tohoku medical megabank toward personalized prevention and medicine [C] //Proc of the 15th World Congress on Health and Biomedical Informatics. Amsterdam: IOS Press, 2015: 1057-1057

[4] Jia Lirong, Liu Jing, Yu Tong, et al. Construction of traditional Chinese medicine knowledge graph [J]. Journal of Medical Informatics, 2015, 36(8): 51-55 (in Chinese)
(贾李蓉, 刘静, 于彤, 等. 中医药知识图谱构建[J]. 医学信息学杂志, 2015, 36(8): 51-55)

[5] Wang Qiao, Wang Wei. Papers on genome epidemiology in the world: A knowledge map-based visual analysis [J]. Chinese Journal of Medical Library and Information Science, 2013, 22(4): 2-9 (in Chinese)
(王俏, 王伟. 基于知识图谱的国际基因组流行病学可视化分析[J]. 中华医学图书情报杂志, 2013, 22 (4): 2-9)

[6] Liu Danhong, Luo Xiaonan, Xu Yongyong. Overview of electronic medical records and its application [J]. Chinese Health Quality Management, 2010, 17(4): 1-5 (in Chinese)
(刘丹红, 罗小楠, 徐勇勇. 电子病历及其应用概述 [J]. 中国卫生质量管理, 2010, 17(4): 1-5)

[7] Taskar B, Fai W M, Abbeel P, et al. Link prediction in relational data [C] //Proc of the 17th Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2003: 659-666

[8] Baader F, Sertkaya B. Usability issues in description logic knowledge base completion [C] //Proc of the 7th Int Conf on Formal Concept Analysis. Berlin: Springer, 2009: 1-21

[9] Sahebi S, Yu-Ru L, Brusilovsky P. Tensor factorization for student modeling and performance prediction in unstructured domain [C] //Proc of the 9th Int Conf on Educational Data Mining. Berlin: Springer, 2016: 502-506

[10] Chang Kaiwei, Yih W, Yang Bishan, et al. Typed tensor decomposition of knowledge bases for relation extraction [C]//Proc of the 2014 Conf on Empirical Methods in Natural Language Processing. Stroundsburg, PA: ACL, 2014: 1568-1579

[11] Morales C, Collarana D, Vidal D E, et al. MateTee: A semantic similarity metric based on translation embeddings for knowledge graphs [C] //Proc of the 17th Int Conf on Web Engineering. Berlin: Springer, 2017: 246-263

[12] Vang K J. Ethics of Google's knowledge graph: Some considerations [J]. Journal of Information, Communication and Ethics in Society, 2013, 11(4): 245-260

[13] Bollacker K D, Evans C, Paritosh P. Freebase: A collaboratively created graph database for structuring human knowledge [C] //Proc of the 39th ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2008: 1247-1250

[14] Bizer C, Lehmann J, Kobilarov G, et al. DBpedia—A crystallization point for the Web of data [J]. Journal of Web Semantics, 2009, 7(3): 154-165

[15] Miller A, WordNet G. A lexical database for English [J]. Communications of the ACM, 1995, 38(11): 39-41

[16] Sousa A M, Pereira O M, et al. MorphoCol: An ontology-based knowledgebase for the characterisation of clinically significant bacterial colony morphologies [J]. Journal of Biomedical Informatics, 2015, 55: 55-63

[17] Abraham I L, Buckwalter K C. Geropsychiatric nursing: A clinical knowledge base in community and institutional settings [J]. Journal of Psychosocial Nursing and Mental Health Services, 1994, 32(4): 20-26

[18] Vives-Boix V, Fernández D R, et al. A knowledge-based clinical decision support system for monitoring chronic patients [C] //Proc of the 7th Int Work-Conf on the Interplay Between Natural and Artificial Computation. Berlin: Springer, 2017: 435-443

[19] Ruan Tong, Sun Chenglin, Wang Haofeng, et al. Construction of traditional Chinese medicine knowledge graph and its application [J]. Journal of Medical Informatics, 2016, 37(4): 8-13 (in Chinese)
(阮彤, 孙程琳, 王昊奋, 等. 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)

[20] Xiang Y, Poh K L. A knowledge-based modeling system for time-critical dynamic decision-making [C] //Proc of the 9th Pacific Rim Int Conf on Artificial Intelligence. Berlin: Springer, 2006: 212-221

[21] Bordes A, Usunier N, García-Durán A, et al. Translating embeddings for modeling multi-relational data [C] //Proc of the 27th Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2013: 2787-2795

[22] Wang Zhen, Zhang Jianwen, Feng Jianlin, et al. Knowledge graph embedding by translating on hyperplanes [C] //Proc of the 28th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2014: 1112-1119

[23] Lin Yankai, Liu Zhiyuan, Sun Maosong, et al. Learning entity and relation embeddings for knowledge graph completion [C] //Proc of the 29th AAAI Conf on Artificial Intelligence Learning. Menlo Park, CA: AAAI, 2015: 2181-2187

[24] Xie Danfeng, Zhang Lei, Bai Li. Deep learning in visual computing and signal Pprocessing [J]. Applied Computational Intelligence and Soft Computing, 2017, 2017: Article ID 1320780

[25] Li Xiang, Taheri A, Tu Lifu. Commonsense knowledge base completion [C] //Proc of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2016: 1445-1455

[26] Jiang Tingsong, Liu Tianyu, Ge Tao, et al. Towards time-aware knowledge graph completion [C] //Proc of the 26th Int Conf on Computational Linguistics. Stroudsburg, PA: ACL, 2016: 1715-1724

[27] Melisachew W C, Giuseppe P, Joerg S, et al. Marrying uncertainty and time in knowledge graphs [C] //Proc of the 31th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2017: 88-94

[28] Sepp H, Jürgen S. LSTM can solve hard long time lag problems [C] //Proc of the 10th Neural Information Processing Systems. Cambridge, MA: MIT Press, 1996: 473-479

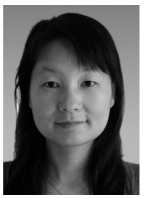
[29] Luo Yuan, Ahmad F S, Shah J S. Tensor factorization for precision medicine in heart failure with preserved ejection fraction [J]. Journal of Cardiovascular Translational Research, 2017, 10(3): 305-312

[30] Robert S, Catherine H. Representing general relational knowledge in conceptNet 5 [C] //Proc of the 8th Int Conf on Language Resources and Evaluation. Marrakech, Morocco: ELRA, 2012: 3679-3686

Chen Dehua, born in 1976. PhD and associate professor. His main research interests include database, data warehouse, big data and deep learning.

Yin Suna, born in 1994. Master candidate. Her main research interest is data mining (yinsuna312@126.com).

Le Jiajin, born in 1951. Professor and PhD supervisor. Member of CCF. His main research interests include database and data warehouse, software engineering theory and practice (lejiajin@dhu.edu.cn).



Wang Mei, born in 1980. PhD and professor. Member of CCF. Her main research interests include database, image semantic analysis and information retrieval (wangmei@dhu.edu.cn).

Pan Qiao, born in 1977. Associate professor and deputy director of department. His main research interests include big data and cloud computing, machine learning (panqiao@dhu.edu.cn).

Zhu Lifeng, born in 1976. PhD candidate at Donghua University, and senior engineer at Ruijin Hospital. His main research interests include medical information management and medicaldata (zlf@rjh.com.cn).