计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20200750
57(12):2501-2513,2020

人机混合的知识图谱主动搜索

王 萌 王靖婷 江胤霖 漆桂林
(东南大学计算机科学与工程学院 南京 211189)
(计算机网络和信息集成教育部重点实验室(东南大学) 南京 211189)
(meng.wang@seu.edu.cn)

Hybrid Human-Machine Active Search over Knowledge Graph

Wang Meng, Wang Jingting, Jiang Yinlin, and Qi Guilin
(*School of Computer Science and Engineering, Southeast University, Nanjing* 211189)
(*Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, Nanjing* 211189)

Abstract Effective search over knowledge graphs can provide support for applications such as question answering and semantic search. However, when the user cannot give a clear query, accurately capturing the user's interest and finding the answer are difficult for machines. Hybrid human-machine active search provides a pathway to bridge the gap between users and machines. Hybrid human-machine active search is a kind of interactive search, and it is originated from the thought of active learning in machine learning field. The core idea is to let the machine issue questions to the user, to obtain information from the user feedback, and then based on this information to eventually capture user intent and return answers. In this paper, we stand on recent advances in knowledge graph representation learning techniques and propose a hybrid human-machine active search in the vector space of a knowledge graph. Specifically, the knowledge graph is first embedded into the low-dimensional vector space, which quantizes the characteristics of entities and relationships, and at the same time, the user's interests and preferences are embedded into the same space. Then, the machine actively proposes questions to the user, and gets the feedback information by asking the user to rate the specific entity, thus updating the user preference positioning in the vector space. We design an evaluation method to measure the user's interest in a specific entity based on the Euclidean distance between the preference point and other entities, and finally find the final target entity to return to the user after multiple turns of human-machine interaction. In the experiment part, we conduct experiments on the knowledge graph embedding and the active search respectively, and the experimental results show that the proposed method is effective.

Key words hybrid human-machine intelligence; knowledge graph; representation learning; active search; semantic search

摘 要 在知识图谱进行有效的搜索可以为智能问答、语义检索等智能应用提供有效支撑.然而,当用户不能给出明确的查询意图时,一个搜索系统要如何精准捕获用户的兴趣并找到对应的查询目标是项难题.人机混合的主动搜索为缓解用户和机器之间的理解鸿沟提供了桥梁.人机混合的主动搜索核心在于让机器主动地向用户提出相关的问题,从用户的反馈中获取信息,再基于这些信息对检索候选项进行

收稿日期:2020-09-15;**修回日期**:2020-10-23
基金项目:国家自然科学基金项目(61906037);CCF-腾讯犀牛鸟基金项目
This work was supported by the National Natural Science Foundation of China (61906037) and CCF-Tencent Open Fund.

搜索,形成人机混合的回路,最终精准定位用户意图并返回查询结果.在知识图谱表示学习技术的基础上,将知识图谱的搜索任务建模成向量空间中人机混合的主动搜索任务.具体来说,首先将知识图谱和用户的兴趣偏好嵌入到同一低维向量空间.然后,机器主动向用户提问,通过让用户对具体实体进行打分的方式获取相应的反馈信息,进而更新用户偏好在向量空间中的定位.设计了一种评价方式,基于偏好点与其他实体之间的欧氏距离来度量用户对某个实体的兴趣,最终在人机多轮交互后找到对应的目标实体返回给用户.在实验部分,对知识图谱的嵌入过程和主动搜索的过程分别进行了实验,实验结果显示,所提出的方法具有一定的效果.

关键词 人机混合智能;知识图谱;表示学习;主动搜索 ;语义搜索

中图法分类号 TP182

自 2012 年谷歌发布了 Google Knowledge Graph[1]后,亚马逊[2]、脸书[3]等多家大型互联网公司先后公布其在知识图谱方面的研究进展和应用.于此同时,学术界有关知识图谱的研究工作也日渐增多[4-6],知识图谱正越来越多地被学术界和工业界所重视.知识图谱的核心思想是使用〈实体-关系-实体〉形式的三元组来表示事实或知识,多个三元组构成的集合形成的图即知识图谱.相比于传统的关系模型[7]或 NoSQL[8],知识图谱这种基于图的数据表示为多个领域提供了简洁直观的数据抽象,现实世界中物体的属性、类型和不同物体之间的联系可以直观地被抽象成为实体和关系,构成一个领域知识图谱,如百科图谱、社交网络中人物画像图谱、生物数据中不同蛋白质图谱等[9].

面向给定的知识图谱,当用户清楚地知道知识图谱的底层数据模式以及自身的查询需求时,很容易基于标准的 SPARQL 查询获取知识图谱中的目标实体和相关信息[10].然而,在很多情况下,用户的查询意图并不明确并且很难完全熟悉知识图谱的底层数据模式,直接给出一个精准复杂的标准 SPARQL 查询可能相对困难.传统的基于知识图谱的探索式搜索[11]可以帮助用户在模糊查询的基础上得到更准确的搜索结果,但依然存在 2 点不足:1)知识图谱可能存在数据不完整的情况,系统很多情况无法给出完全匹配用户意图的答案,需要提供近似候选项,而传统的探索式搜索依然依赖精准匹配策略;2)系统无法在用户不主动输入查询的情况下向用户推荐实体,尤其是当用户信息需求不明确时,往往在搜索过程中偏向于被动的指引,而探索式搜索无法应对这种情形.

人机混合的主动搜索[12]是预测用户兴趣偏好的一个行之有效的方法,主要应用于信息检索和推荐系统领域.人机混合的主动搜索是交互式搜索的一种.普通的交互式搜索系统通过与用户进行交互,如:让用户对返回结果进行评分、手动添加或去除搜索结果、自动将用户基于自然语言的限定附加条件转换为标准查询语言等各种不同的方法,提高搜索的准确率或用户的满意度.即,一般的交互式搜索系统利用用户的反馈对原有的搜索结果进行优化.而人机混合的主动搜索系统则不同,人机混合的主动搜索系统基于用户对机器所提问题的反馈产生搜索结果.即,主动搜索系统并不直接给出搜索结果,而是通过机器主动地向用户提出相关的问题,从用户处获取信息,其搜索结果基于这些信息得到.机器提出的问题包括:1)比较型:“这 2 件物品你更偏好哪一件?”;2)评分型:“请给你对演员约翰尼·德普的感兴趣程度评分.”等.在用户具体需求模糊的情况下,人机混合的主动搜索方法可以通过用户的反馈来挖掘并预测用户的查询需求,进而给用户推荐搜索结果.

最近人机混合的主动搜索研究开始与表示学习技术结合,如文献[13]和文献[14]将检索或推荐对象(item)嵌入到一个低维向量空间中,并试图通过人机交互的过程预测出用户兴趣在此向量空间中分布,即用户兴趣偏好点也被表示为一个空间中的向量.这种方法的优势是可以在一定程度上克服数据缺失带来的不利影响,提高系统对于数据不完整和数据噪音的鲁棒性.与此同时,随着 TransE[15]和 RESCAL[16]等知识图谱表示学习方法的出现,我们已经可以将知识图谱嵌入到低维空间中,在保留知识图谱本身信息的同时,提升知识图谱上的计算操作并进行知识补全.这意味着我们可以将人机混合的主动搜索与知识图谱表示学习技术结合,设计人机混合的知识图谱主动搜索方法,进而实现智能增强的知识图谱搜索体验.

综上,在本文中我们提出一种人机混合的知识

图谱主动搜索方法,通过机器主动要求用户给一个实体打分,同时利用知识图谱中关系以及类型信息,最后结合用户历史搜索结果来估计用户的查询意图,输出最终的搜索答案.以图 1 的场景为例:

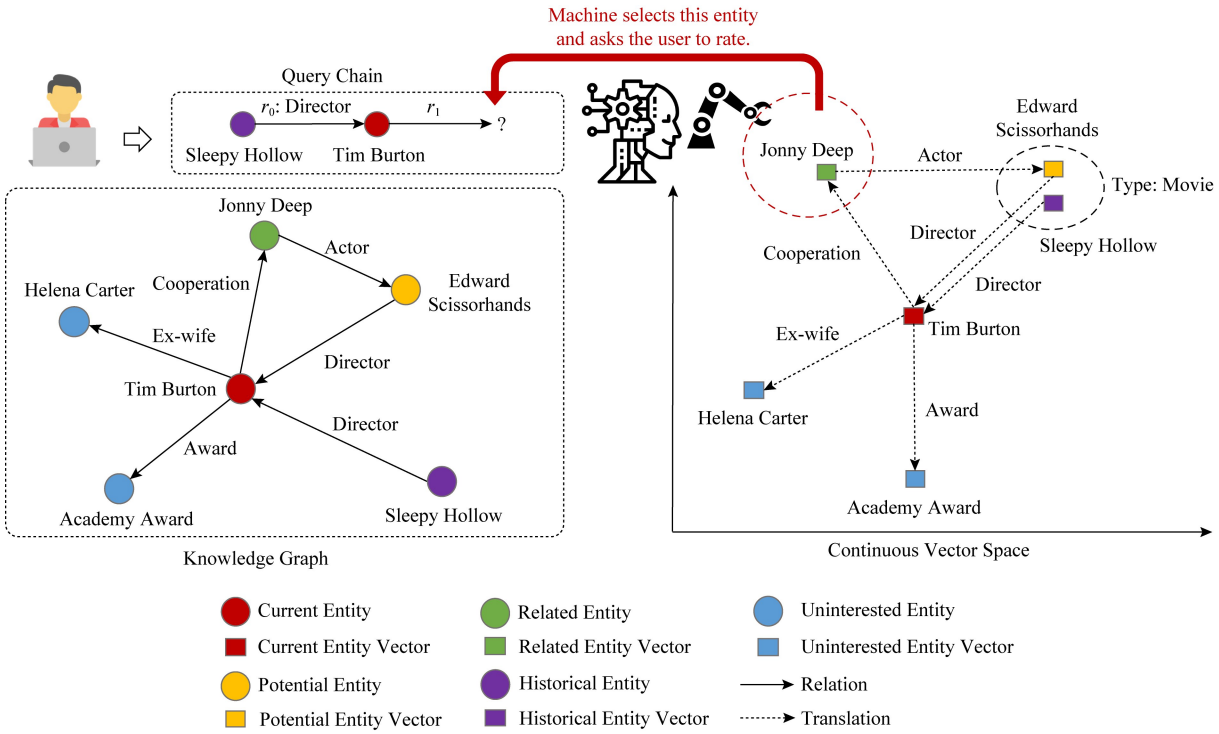


Fig. 1　Example of hybrid human-machine active search over knowledge graph

图 1　人机混合的知识图谱主动搜索实例

想象一个用户从知识图谱中搜索电影《断头谷》的有关信息,在找到它的导演蒂姆·伯顿后,想要了解蒂姆·伯顿的更多相关信息.如图 1 所示,知识图谱中给出了这个导演的相关实体,包括许多方面:1)他的电影生涯.包括他导演或参与的电影,比如《剪刀手爱德华》,他获得的奖项,比如奥斯卡奖等.2)人际关系.包括与他合作的演员 Jonny Deep,以及他的前妻 Helena Carter.在当前场景下,机器对于用户的目标实体可能很难直接预测.假设在我们示例中用户的最终目标实体是另一部电影《剪刀手爱德华》.但请注意,在主动搜索过程开始时,用户并不知道知识图谱中存在什么样的信息,也不知道什么样的信息可以引起其兴趣.现在机器需要不停地向用户提问,进而通过用户的反馈在所有相邻实体中找到用户潜在的感兴趣实体,即目标实体《剪刀手爱德华》.在该例子中,一些实体可以提供信息:实体 Jonny Deep 可能因其与目标实体之间存在关系而位于目标附近,用户可能倾向于给它更高的分数,这意味着用户偏好点可能更接近该实体;而实体 Helena Carter 与目标实体没有关系,偏好点可能离该实体较远.历史查询实体也可以提供信息:用户的历史查询中包含电影《断头谷》,因此用户可能对其他电影类型的实体有更多的兴趣.我们的目标是让机器最大限度地利用已知信息向用户提出问题,最终将用户偏好点在向量空间中尽可能预测在目标实体附近.以上人机混合的主动搜索过程,可以在用户本身搜索意图不明确时,在较短时间内将用户意图定位,实现答案的定位和输出.

1　相关工作

本节主要对本课题相关研究领域的国内外研究现状进行概述,主要包括:知识图谱上的传统搜索方法、人机混合的主动搜索,以及本文所用到的知识图谱表示学习技术.

1.1　知识图谱搜索方法

知识图谱的标准查询是基于结构化的查询语言 SPARQL[①],类似于在关系数据库中使用 SQL 查询

① https://www.w3.org/TR/rdf-sparql-query/

数据.在此基础上,研究人员提出了多种查询范式[17],和本文关系密切的是知识图谱的探索式搜索.其中,基于自动操作的搜索系统尝试不同的方法来自动地对结果进行优化.基于词法分析的系统[11]可以自动放宽查询条件并提高召回率.Yahya 等人提出的 TriniT 模型[18]允许用户在查询之前使用自然语言给出一个详细描述,并根据该描述重新构建查询语句或放宽该语句的条件.交互式搜索系统一般在返回搜索结果后与用户交互来优化查询本身或搜索结果.DEANNA[19]和 IMPROVE-QA[20]系统将用户的自然语言查询转换为标准的 SPARQL 查询,并且在转换的过程中,针对可能出现的歧义问题,系统通过收集用户对先前结果的反馈来改进标准 SPARQL 查询.

1.2 人机混合的主动搜索

当用户需求不明确时,人机混合的主动搜索模型[12]让机器主动向用户提问,从用户反馈中学习,帮助预测用户兴趣目标,这里用户的兴趣目标通常被表示为一个点或向量.主动搜索的思想由主动学习而来,背后的关键思想是,如果允许机器学习算法从其学习的目标(即用户处)获得训练数据,那么它可以在较少的标记训练实例的情况下获得更高的精度[16].当应用在传统的数据库如图像数据库时,这些方法通过向用户提出对比型问题或相关型问题来主动地获取用户反馈信息.Cox 等人[21]要求用户从 2 个图像中选择一个更相关的图像以寻找单个目标.SVM-active[22]通过选择包含最多信息的项(图像)并询问用户它们是否与用户目标相关,学习了一个支持向量机模型,最终从数据库中返回一个图像集合.在这种思想的指引下,研究人员开始将用户的偏好预测建模成向量空间中的人机混合主动搜索过程[13-14].在此过程中,模型将待检索条目和用户偏好同时嵌入欧几里德空间(Euclidean space)之后,利用用户和条目之间的欧几里得距离(Euclidean distance)来反映相似性.在此基础上,Canal 等人[13]基于机器主动提出成对比较问题的方法,对包含有噪声的系统响应进行了建模.MF-ASC[14]采用多级置信的方法,在获得用户的真实目标的同时降低了时间复杂度.在已有的主动搜索系统中,检索条目在多数情况下都是单独的个体.它们之间最紧密的联系仅仅是它们所共享的属性,而包含了大量信息的物体之间的关系并没有被考虑.而知识图谱包含实体类和实体间关系,在主动搜索这一过程中并没有被考虑.

1.3 知识图谱表示学习

知识图谱表示学习,也称知识图谱嵌入,旨在将图谱中的实体和关系映射到一个低维连续向量空间中,并使用低维的向量/矩阵等表示它们,其中,实体被表示为向量,而关系一般被表示为转移向量(translation)或矩阵.这样做可以在保留知识图谱本身信息的同时,简化知识图谱上的计算,这一过程也被称为知识图谱嵌入.嵌入的过程也是实体和关系数值化的过程,便于相关问题的建模.常见的知识图谱嵌入模型分为翻译距离模型和语义匹配模型:1)翻译距离模型使用基于距离的评分函数,其中最具有代表性的是 Bordes 等人提出的 TransE[15].在该模型中,实体和关系都被表示为同一空间的向量,对于每个三元组,模型假设头实体向量与关系向量之和等于尾实体向量,通过减小二者之差对图谱的表示进行学习.TransH[23],TransR[24]模型分别引入了关系平面、关系空间的概念,通过将实体向量映射后再进行运算,以便提升翻译距离模型的适用性和效果.2)语义匹配模型则主要使用基于相似度的评分函数.Nickel 等人提出 RESCAL[16]模型,将关系建模为矩阵,使用双线性函数作为评分函数,相较于翻译模型包含了更多语义信息,复杂度也更高.DistMult[25]和 HolE[26]试图简化 RESCAL 的复杂度:DistMult[25]将关系矩阵定义为对角矩阵,在降低复杂度的同时减少了信息;HolE[26]则通过循环乘积的方法,在信息总量不变的情况下降低了关系矩阵的规模.更多知识图谱表示学习方法可参见综述[27].

2 本文动机和关键问题

文献[28-29]将知识图谱的表示学习技术应用于搜索任务中,主要用以解决缺失信息引起的查询结果空集问题.这类方法将用户的搜索目标建模为低维空间中的一个点,通过返回该点周围的实体,可以得到符合用户预期的近似结果.在此基础上,本文将知识图谱表示学习技术和人机混合的主动搜索模型结合,为预测用户偏好和寻找潜在目标信息提供了一种思路.然而,要想实现此目标还面临 3 个主要问题:

1) 在主动搜索场景下,应当使用哪种合适的方法嵌入知识图谱.

2) 我们需要让用户在一个实体上打分进行反馈.在给定的情况下,我们应该如何设计机器的选择机制,对实体进行选择提问,以及系统在收到用户的反馈后应该如何反应.

3) 实体、实体类型和查询历史之间的关系中包含了大量有用的信息,我们如何在主动搜索的过程

中充分考虑这些信息进而提升模型效果.

为此,在本文所提框架中,我们首先将用户的意图表示为同一个低维空间中的低维向量,通过对用户主动提问获取信息,基于用户反馈来推断并预测用户的兴趣意图,从而实现基于知识图谱嵌入的人机混合主动搜索的问题定义;在此基础上,定义了用户查询链的概念来表示用户的历史查询,然后根据实体的特征、关系、类型和用户查询链,引入一个加权的实体相关向量来选择问题实体,提高了推荐的准确率.

3　基于知识图谱嵌入的主动搜索模型

人机混合的主动搜索系统目标是让机器和人进行交互,这里的"主动"指机器"主动提问"的含义.核心思想是通过"机器提问-用户回答"的方式,利用一个用户的历史查询链在知识图谱向量表示空间中逐渐地找出用户查询目标.为此,我们提出了一个基于监督学习的模型,它通过要求用户给一个实体打分来主动获取信息.为了保证机器在提问过程中选择包含更多信息的问题实体,整个主动搜索框架设置了与关系类型、实体类型和历史查询链信息相关的可训练参数.此外,我们还为本框架设计了一个规则,该规则主要用于决定用户对实体打分后用户偏好点的移动路径.本节主要分为3个部分,分别介绍本课题相关的基本定义与概念、所构建模型的理论分析和算法设计以及通过监督学习的方法对该模型进行训练的过程.

3.1　基本问题定义与相关概念

本节将介绍人机混合的知识图谱主动搜索基本问题定义,以及一些相关术语和基本概念.

1) 知识图谱.考虑一个知识图谱 $\mathcal{G}=(\mathcal{E},\mathcal{R})$,其中 $\mathcal{E}$ 是知识图谱中所有实体的集合,$\mathcal{R}$ 是能够连接 $\mathcal{E}$ 中的实体的所有关系的集合,$\mathcal{G}$ 是一组由〈实体-关系-实体〉三元组组成的集合.一条三元组可以表示为(e_h,r,e_t),其中 $e_h,e_t\in\mathcal{E},r\in\mathcal{R}$.一条三元组代表的意义为,在头实体 e_h 和尾实体 e_t 之间,存在关系 r 将二者连接.当2个实体同时出现在同一条三元组中时,我们称这2个实体互为邻居,在本文中,我们使用集合 $\mathcal{N}_e$ 来表示实体 e 的所有邻居实体的集合.

2) 用户查询链.当用户查询完 n 个实体之后,我们使用一条用户查询链 $\mathcal{C}=(\mathcal{E}_\mathcal{C},\mathcal{R}_\mathcal{C})$来表示用户的查询历史.其中,$\mathcal{E}_\mathcal{C}=\{e_0,e_1,\cdots,e_n\}$,$\mathcal{R}_\mathcal{C}=\{r_0,r_1,\cdots,r_{n-1}\}$.对于任意的 $i\in[0,n-1]$,实体 e_i,e_{i+1} 和关系 r_i 可以组成一个三元组,即三元组(e_i,r_i,e_{i+1})或三元组(e_{i+1},r_i,e_i)成立.我们通过保存用户查询路径上的实体和关系来记录其查询链.

3) 知识图谱嵌入和用户偏好嵌入.通过知识图谱的表示学习算法,我们可以将知识图谱嵌入到一个低维的向量空间.对于知识图谱中的任一实体 e,在该空间中被表示为一个点,我们使用粗体 $\boldsymbol{e}$ 来表示原点到 e 所对应的点的向量.我们使用 u 表示用户兴趣偏好的对应点,而向量 $\boldsymbol{u}$ 来表示该空间中原点到用户的兴趣偏好嵌入的向量.用户对特定实体 e 的感兴趣程度可以通过 u 和 e 的欧几里得距离来测量,也就是说,当用户偏好点与某个实体所对应点的距离越近,用户就越可能对该实体感兴趣.

4) 人机混合的主动搜索问题.给定一个用户查询链 $\mathcal{C}$,人机混合的主动搜索的目标是,在用户不再给出新的查询关键字的情况下,通过机器主动对用户进行提问,获取反馈进而找到用户的下一个目标 e_t.因此,本框架中机器应当主动地向用户提出一系列的问题以得到更多的信息来准确定位查询目标.我们定义一个问题 q 包含了一个实体 e_q,用户需要基于其本身的兴趣对该问题实体 e_q 给出评分 $rate$,其中 $rate\in\{-2,-1,1,2\}$,从小到大分别代表"厌恶""不感兴趣""感兴趣"和"非常感兴趣".提问过程基于 e_t 的邻居集合,也即,问题实体 $e_q\in\mathcal{N}_{e_n}$,且目标实体 $e_t\in\mathcal{N}_{e_n}$.基于用户给出的评分,我们的框架相应地对用户的偏好嵌入向量 $\boldsymbol{u}$ 进行调整,使之更接近目标节点的嵌入.从用户处获得了足够的信息之后,框架会返回一个答案列表 $\mathcal{L}$,$\mathcal{L}$ 中包括最有可能为目标实体的 k 个答案.具体工作流程如图2所示.

3.2　理论分析与算法设计

3.2.1　知识图谱嵌入过程

本文中的知识图谱嵌入基于 TransE 算法,该算法将一个知识图谱嵌入到低维向量空间中.在 TransE 模型的实现里,实体之间的关系在低维向量空间中被表示为翻译的过程,即该算法基于如下先决假设:如果三元组(e_h,r,e_t)成立,那么头实体 e_h 的嵌入与某个与关系 r 相关的向量相加之和应当与尾实体 e_t 的嵌入相近;反之,若该三元组不成立,那么头实体嵌入与关系向量相加之和应当尽可能远离尾实体嵌入.

Xavier 初始化方法[30]是一种很有效的神经网络初始化方法.其思想是使每一层网络输出的方差都尽量相等,也就是说,正向传播时,激活值的方差

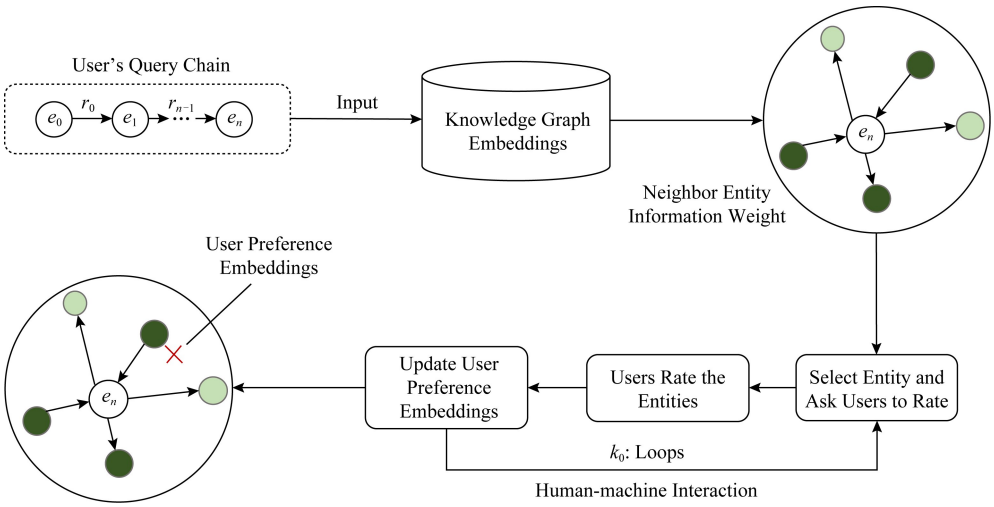


Fig. 2 Overview of our proposed framework

图 2 模型整体框架流程图

保持不变;反向传播时,关于状态值的梯度的方差保持不变.基于以上思想,最终推导出一个均匀分布来初始化参数,这样可以使网络中信息更好地流动.

给定一个由三元组(e_h,r,e_t)组成的训练集 S,其中 $e_h,e_t\in\mathcal{E},r\in\mathcal{R}$.假设训练时的批尺寸为 bs,学习率为 lr,嵌入的维度为 d,e_h,e_t 的嵌入为 $\boldsymbol{e}_h$ 和 $\boldsymbol{e}_t$,r 的嵌入为 $\boldsymbol{r}$,则有 $\boldsymbol{e}_h,\boldsymbol{e}_t,\boldsymbol{r}\in\mathbb{R}^d$.接下来的具体过程为:

1) 使用 Xavier 初始化方法对实体和关系的嵌入执行初始化操作.

2) 在当前 batch 中,从 S 中取得大小为 bs 的正样本三元组集合 P_{batch},对每个三元组$(e_h,r,e_t)\in P_{batch}$,随机更换其头实体或尾实体,获得一个新的负样本三元组(e_{nh},r,e_{nt}),使得$(e_{nh},r,e_{nt})=(e_h,r,e'_t)$或$(e_{nh},r,e_{nt})=(e'_h,r,e_t)$,将所有负样本三元组组成负样本集合 N_{batch},令 $T_{batch}=P_{batch}\cup N_{batch}$,组成当前批次的训练数据.

3) 定义其损失函数为:

$$loss=\sum_{(e_h,r,e_t)\in P_{batch}}[\gamma+dist(\boldsymbol{e}_h+\boldsymbol{r},\boldsymbol{e}_t)-dist(\boldsymbol{e}'_h+\boldsymbol{r},\boldsymbol{e}'_t)],$$

其中 γ 为常数,$dist$ 为欧几里得距离函数.计算出损失函数后,基于反向传播算法更新实体嵌入和关系嵌入.

4) 循环步骤 2,3 直至训练结束.

基于过程 1)~4),我们通过训练即可得到知识图谱的低维向量空间嵌入.语义上,每个实体都由低维空间中的一个点来表示.在之后,我们将用户偏好同样用该空间的一个点表示,采用欧几里得距离来度量用户偏好与某特定实体之间的相近程度.

3.2.2 人机混合的主动搜索过程

由于用户不会提供任何新的查询条件,因此机器只能通过主动提问的方式来获取新的信息.此步骤的核心关键思想是,机器需要学会利用现有的信息(即,用户查询链和知识图谱嵌入情况),通过向用户提出尽可能少的问题来获取尽可能多的信息.因此,整个主动搜索的解决过程可以被拆分为 2 个主要问题:

1) 在提问之前,如何选择合适的实体进行提问.

2) 在收到用户的答案之后,系统如何调整用户偏好点的嵌入.

接下来将主要描述以上 2 个问题的解决方法:

1) 选择合适实体进行提问.为了最大限度地提高提问效率,模型需要选择在用户给出答案后能够获取更多信息的提问实体.基于常识和经验,我们对可能包含更多信息的实体的一般特征做出了一些假设,系统根据假设对问题实体进行选择.

假设 1. 在某一个具体的主动学习场景下,具有以下特征的实体可能包含有更多的信息:①拥有一个更大的度.一个拥有更多邻居的实体意味着它更有可能成为用户的搜索目标.以蒂姆·伯顿为例,当用户搜索到电影《剪刀手爱德华》之后,蒂姆·伯顿作为该电影的导演,相比于该电影的助理导演弗朗西斯·康威更有可能成为用户的下一个目标,因为前者所参与导演的电影要比后者多得多,用户也更有可能对他感兴趣.②属于一个更好的类.当某个类型所包含的实体越热门,该类型就越可能是一个热门类,

用户对其感兴趣的可能性会更高.以电影演员和电影版号的区别为例,在用户搜索一部电影的相关信息时,他们显然更愿意知道该电影的演职员信息而非电影的版号之类的数字信息.③与历史查询链相关.这个假设中的相关,包括了"该实体与此前查询过的实体属于同一类型""该实体与起点实体之间的关系属于此前查询过的关系之一".显然地,当用户曾经搜索过一名导演之后,其会更倾向于搜索另外一名导演,因为该用户的搜索历史反应其很可能对导演这一职业感兴趣.

基于该假设,我们分别针对以上3条问题实体的特征,定义了3个可训练的参数,它们可以结合起来共同表示一个实体中所包含的信息总量.

定义1. 实体信息总量. 给定一个用户查询链 $\mathcal{C}=(\mathcal{E}_{\mathcal{C}},\mathcal{R}_{\mathcal{C}})$,其中 $\mathcal{E}_{\mathcal{C}}=\{e_0,e_1,\cdots,e_n\}$,$\mathcal{R}_{\mathcal{C}}=\{r_0,r_1,\cdots,r_{n-1}\}$,我们定义 $\mathcal{T}_{\mathcal{C}}=\{t_0,t_1,\cdots,t_n\}$,对任意的 $i\in[0,n]$,t_i 是实体 e_i 所属的类别.对于 e_n 的一个邻居实体 e_m,$e_m\in\mathcal{N}_{e_n}$,假设其从属于类型 t_m,且通过关系 r_m 与实体 e_n 相连接,我们如下定义 e_m 所包含的信息总量:

$$\mathcal{INFO}_{(\mathcal{C},e_m)}=W_{\mathrm{e}}(e_m)+W_{\mathrm{t}}(\mathcal{C},e_m)+W_{\mathrm{r}}(\mathcal{C},e_m),\tag{1}$$

其中,W_{e},W_{t},W_{r} 分别代表该实体 e_m 的度、所属类别以及与 e_n 相连接的关系所带来的影响.这3条权重分别可以被表示为:

$$W_{\mathrm{e}}(e_m)=w_{\mathrm{e}}[e_m],\tag{2}$$

$$W_{\mathrm{t}}(\mathcal{C},e_m)=(p_{\mathrm{t}}*g_{\mathrm{t}}(\mathcal{C},t_m)+(1-p_{\mathrm{t}}))*w_{\mathrm{t}}[t_m],\tag{3}$$

$$W_{\mathrm{r}}(\mathcal{C},r_m)=(p_{\mathrm{r}}*g_{\mathrm{r}}(\mathcal{C},r_m)+(1-p_{\mathrm{r}}))*w_{\mathrm{r}}[r_m],\tag{4}$$

其中,参数 p 是一个常数,它决定用户查询链对信息总量影响的占比.一个越大的 p 意味着我们为历史查询赋予更大的权重.在实验中,我们默认设置 $p_{\mathrm{t}}=p_{\mathrm{r}}=0.5$.函数 g 为布尔函数,它代表着实体 e_m 和用户历史查询链 $\mathcal{C}$ 之间的关系,如果 e_m 的所属类别 t_m 或 e_m 与 e_n 相连接的关系 r_m 曾经在历史查询链中出现过,则该项置为1,若未曾出现,则该项置为0.该函数将推荐实体和用户历史查询相关联,该实体与用户的查询历史关联越大,则其被作为问题提问的可能性越大.函数 g 用数学式表示为:

$$g_{\mathrm{t}}(\mathcal{C},t_m)=\begin{cases}1, & t\in\mathcal{T}_{\mathcal{C}},\\ 0, & t\notin\mathcal{T}_{\mathcal{C}},\end{cases}\tag{5}$$

$$g_{\mathrm{t}}(\mathcal{C},r_m)=\begin{cases}1, & r\in\mathcal{R}_{\mathcal{C}},\\ 0, & r\notin\mathcal{R}_{\mathcal{C}}.\end{cases}\tag{6}$$

基于定义1,对于集合 $\mathcal{N}_{e_n}$ 中的任意一个实体 e_m,我们都能够计算其信息权重向量 $\boldsymbol{I}_{\mathcal{C}}$ 了,计算公式为:

$$\boldsymbol{I}_{\mathcal{C}}=softmax([\mathcal{INFO}_{(\mathcal{C},e_m)}\mid\forall e_m\in\mathcal{N}_{e_n}]).\tag{7}$$

在获取问题实体时,我们可以基于该权重向量 $\boldsymbol{I}_{\mathcal{C}}$ 进行随机加权采样,以在用户对该实体进行评分时获取尽可能多的信息.显然地,该信息权重与假设中3种实体特征呈正相关,即:越符合假设中3种特征的实体,其信息权重越高,同时其被选中作为问题实体的可能性也就越大.我们称这种采样方式为标准采样.

2) 更新用户偏好的嵌入表示.在用户对被选中的问题实体进行评分并反馈之后,模型显然获得了更多的信息,基于这些信息,模型可以更新用户的偏好嵌入点.在本模型中,这一更新过程的规则基于一个朴素的假设,即:用户的偏好嵌入应当更接近其感兴趣的实体,而更远离其不感兴趣的实体.因此,在设计的偏好嵌入更新规则中:用户每回答一个问题,如果呈现出正向的情感,即给出了正分数,则用户的偏好嵌入向该实体靠近;反之,如果呈现出负面的情感,则用户的偏好嵌入离该实体更远.该模式下最理想的情况是,用户的目标 e_{t} 在第1次提问时就被推荐,而在第1次提问过后,用户的偏好嵌入 $\boldsymbol{u}$ 直接朝目标 $\boldsymbol{e}_{\mathrm{t}}$ 方向移动一倍的 $dist(\boldsymbol{u},\boldsymbol{e}_{\mathrm{t}})$ 的距离.但实际情况下,显然这是不可能成立的,因为可能会出现2种情况:

① 由于问题实体是以一个信息权重向量为基础进行随机加权采样的,如果提问次数过少(如1次或2次),则有可能会抽取到包含信息较少的实体;

② 若增加提问的次数,在进行了多轮提问之后,用户偏好很可能已经靠近目标实体,此时若与第1轮移动相同的幅度,有可能会造成反作用,即用户偏好嵌入反而会离目标实体越来越远.

因此,我们分别考虑了避免这2种情况的方式:为了防止单次提问后用户偏好嵌入误差较大的现象,我们在训练和测试时设置了多轮次的提问模式来提高这一过程的容错率;为了防止用户偏好嵌入移动幅度过大反而远离目标点,我们在每次提问时移动的距离量为当前用户嵌入到问题实体距离的 $1/k_0$,其中 k_0 为提问的总次数.

在3.1节中曾经提到,用户对一个问题实体的评分被按照其对该实体的兴趣划分为4种.假设该问题实体为 e_{q},其在低维空间中的嵌入为 $\boldsymbol{e}_{\mathrm{q}}$,且用户对其评分为 s.将用户的偏好嵌入从 $\boldsymbol{u}$ 更新为 $\boldsymbol{u}'$,

我们提供了对这一过程的实现算式：

$$\boldsymbol{u}'=\boldsymbol{u}+p\times\frac{s}{k_0}(\boldsymbol{e}_q-\boldsymbol{u}), \tag{8}$$

其中 p 是一个噪声常数，可以用不同的方法实现，列举如下：

常数： $$p=p_0, \tag{9}$$

距离正则： $$p=\frac{dist(\boldsymbol{e}_q,\boldsymbol{u})}{dist(\boldsymbol{e}_q,e_n)}\times p_0. \tag{10}$$

基于式(8)规则，我们可以在每次从用户处获取对问题实体的反馈之后，对用户偏好向量进行移动，以使其在理论上更靠近目标实体.

3) 生成答案列表.将选择问题实体、向用户主动提问、更新用户偏好嵌入这一过程迭代 k_0 次之后，用户偏好嵌入可以被调整到一个相对准确的位置.我们可以通过执行最近邻搜索算法来找到排名前 k 位次的实体作为答案列表返回给用户.该搜索过程可以用简单的 KD 树算法或者局部敏感 Hash 等高效算法来提升效率.

3.2.3 算法复杂度分析与对比

基于以上给出的信息，我们可以计算出本模型所采用的算法的时间复杂度.用户从查询链开始到查询结束共分为 3 个过程，即问题实体选择、用户偏好嵌入更新和答案列表生成.在问题实体选择阶段，对于每个邻居实体都需要计算其相应的权重，在计算权重时，每个实体都要与历史查询链中的关系、实体和实体类型进行比对，因此该过程的时间复杂度为 $O(len(\mathcal{C})\times|\mathcal{N}_c|)$，其中 len 表示用户查询链的长度，而 $|\mathcal{N}_c|$ 则表示邻居实体的个数；在用户偏好嵌入更新阶段，用户原本的偏好向量需要与它到问题实体向量的某个倍数相加，因此该过程的时间复杂度为 $O(d)$，其中 d 为知识图谱嵌入的维度.以上 2 个阶段共需要循环 k_0 次，因此该过程所需的时间复杂度为 $O(k_0\times(len(\mathcal{C})\times|\mathcal{N}_c|+d))$.在生成答案列表时，需要利用 KNN 算法，在所有邻居实体中选出距离最近的 n 个，该过程的时间复杂度为 $O(d^2\times|\mathcal{N}_c|)$.整个过程的时间复杂度为二者之和，即 $O(d^2\times|\mathcal{N}_c|+k_0\times len(\mathcal{C})\times|\mathcal{N}_c|)$.

文献[13]与本文相似，它通过机器提出的成对比较问题解决了在低维空间上的主动搜索问题.在用户给出反馈并调整用户偏好嵌入这一步骤，它预测了用户偏好的后验分布，因此其时间复杂度与嵌入维度 d 的立方相关，这导致它每预测一个用户偏好项，其模型内部计算耗费的时间以分钟为量级，等待时间很长，与现实使用场景并不符合.本文通过更改具体训练过程，进而优化了模型内部计算的时间复杂度.

3.3 模型的训练过程

在现实情况(即测试环境)下，模型在定位用户偏好嵌入之前，只能向用户提出有限次问题，否则由于用户解答某个问题所需要的时间以秒为单位，该提问的过程会耗费用户过多的时间.但在训练的过程中并没有时间的限制，因此我们设计了一种模拟评分模式，该模式在训练时能够基于用户的目标实体嵌入、起点实体嵌入以及问题实体嵌入三者之间的距离关系来模拟用户给出的分数.以图 1 为例，在训练时对每一个邻居进行考量，邻居如“海伦娜·卡特”和“学院奖”离目标较遥远，而邻居如“约翰尼·德普”“《剪刀手爱德华》”则离目标较近或者就是目标本身.对于前者，用户应当不感兴趣；对于后者，用户应当感兴趣.而即便对 2 个实体都感兴趣，其兴趣程度也不尽相同，如，当用户的目标恰为“《剪刀手爱德华》”时，对实体“约翰尼·德普”的兴趣程度显然会低于这部电影.因此在设计的评分模式中，我们假设用户对某实体的兴趣程度随着该实体与目标实体距离的减小而提高，如图 3 所示，黑色点表示起点实体，绿色点表示目标实体.以目标实体为中心，目标实体与起点实体之间的距离的 0.5,1,2 倍为半径，将空间划分为 4 个不同的区域，处于这四个区域的其他实体分别对应该区域的分数(−2,−1,1,2).

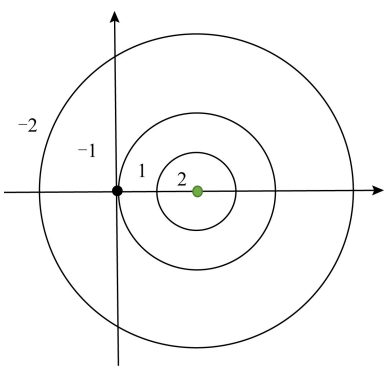


Fig. 3 Entity user ratings based on the Euclidean distance

图 3 基于欧几里得距离来模拟用户对实体评分程度

基于该评分模式，我们设置了一种训练机制，在该机制的作用下，训练过程中，每一个邻居实体都会被考虑到，训练过程也会更为充分.详细的说明与相关数学式为：

在 3.2 节中我们提到，该模型中共有 3 个可供训练的参数，分别是代表着实体本身特征权重的 w_e，

实体所属类型的权重 w_t 以及链接实体的关系的权重 w_r,它们共同作用决定了一个实体被选中作为问题实体的可能性,得到了权重向量 $\boldsymbol{I}_C$.也就是说,$\boldsymbol{I}_C$ 代表着每个邻居实体对用户偏好嵌入的移动所施加的影响.在训练的过程中,我们也基于权重向量 $\boldsymbol{I}_C$ 对用户偏好嵌入进行移动.对于实体 $e\in\mathcal{N}_{e_n}$,假设用户给出的评分为 s_e,其权重为 $\boldsymbol{I}_C[e]$,则其对用户偏好嵌入造成的影响为 $\boldsymbol{I}_C[e]\times s_e\times(\boldsymbol{e}-\boldsymbol{u})$.可见,在该训练模式中,单个邻居实体对用户偏好嵌入的影响与其本身的信息权重呈正比.将所有邻居实体造成的影响叠加,即得到用户偏好嵌入移动的公式为:

$$\boldsymbol{u}'=\boldsymbol{u}+p\times\sum_{e\in\mathcal{N}_{e_n}}\boldsymbol{I}_C[e]\times s_e\times(\boldsymbol{e}-\boldsymbol{u}). \tag{11}$$

由于在训练过程中没有用户反馈的参与,我们通过比较起点实体 e_n、问题实体 e_q 和目标实体 e_t 三者之间的欧几里得距离来估计用户的评分.当目标实体与问题实体之间的距离大于起点实体与问题实体之间的距离时,模拟用户为不感兴趣,给出负分;反之,当目标实体与问题实体之间的距离小于起点实体与问题实体之间的距离时,模拟用户为感兴趣,给出正分.而感兴趣的程度通过比较两距离之间的倍数得到.不妨设 $d=dist(\boldsymbol{e}_q,\boldsymbol{e}_t)$,$d_1=dist(\boldsymbol{e}_n,\boldsymbol{e}_t)$,设评分函数为 h,该函数用数学式表示如下:

$$h=\begin{cases}-2, & d\in[2\times d_1,\infty)\\ -1, & d\in[d_1,2\times d_1)\\ 1, & d\in\left[\dfrac{d_1}{2},d_1\right)\\ 2, & d\in\left[0,\dfrac{d_1}{2}\right)\end{cases}. \tag{12}$$

现在,我们在训练时可以得到更新后的用户偏好嵌入 $\boldsymbol{u}'$ 了.接着,我们设置损失函数为:

$$\mathcal{L}(\mathcal{C})=dist(\boldsymbol{u}',\boldsymbol{e}_t). \tag{13}$$

显然地,该损失函数与用户偏好嵌入和目标实体之间的距离成正比.在训练时,针对该函数值进行优化,我们可以使用户偏好嵌入在移动后与目标实体更加靠近.

3.4　小　结

本节给出了人机混合的知识图谱主动搜索框架的理论分析、框架设计、参数设计、算法设计、具体实现方法和详细的监督学习训练步骤等内容.我们提出了实体信息总量的基本概念和定义,利用多个可训练参数对其进行表示.基于实体信息总量,我们可以对向用户提问的实体进行采样.同时,我们提出了一种基于欧几里得距离的模拟用户偏好的方法,以模拟用户对某个具体实体的评分.我们设计了一种训练方法,对于每一条数据,该方法都能够训练到每一个邻居实体的权重,提高了训练的效率.利用 PyTorch 深度学习框架①,我们用 Python 语言对所提出的监督学习过程进行了实现和测试.

4　实验结果和分析

4.1　实验数据集及评测流程

我们基于 Freebase 提供的知识图谱数据集 FB15K 的数据进行实验.该数据集共包含有 14 951 个实体、1 345 种关系以及 571 种实体类型,其训练集包含 483 142 个三元组,验证集包含 50 000 个三元组,测试集包含 59 071 个三元组.其中,每一个三元组都是唯一的,且出现在验证集和测试集中的同义词集也出现在训练集中.FB15K 中的实体类型来自于 Xie 等人的 TKRL[31] 模型中提供的数据.基于该数据我们对用户历史查询链进行采样,首先随机采样一个实体 e_0,接着随机采样一个该实体的邻居实体 e_1,及二者之间的关系 r_0.在采样之后的实体时,考虑到用户可能更倾向于选取此前出现过的类别的实体,我们记录出现过的实体类,并对属于该类别的实体权重进行一定程度的提高.我们采样了总长度为 4 的用户搜索链,即假设用户正在搜索第 5 个实体,并进行了测试.所有本文的数据集及源代码我们已开源在 Github② 上以供参考.

4.2　相关实验数据展示与分析

4.2.1　知识图谱嵌入效果

本项目采用了基于 TransE 的算法对知识图谱进行嵌入.嵌入后的评价指标为,对一个三元组集合,隐藏其尾实体,并取头实体嵌入与关系对应向量之和,对该相加和与所有的实体嵌入之间的距离进行从小到大的排序,并记录尾实体的名次.我们统计了该名次的均值、排名前十的比例、排名前三的比例以及排名第一的比例.且又因为,在训练时会刻意破坏一些三元组,这些损坏的三元组可能在训练集和验证集中.在这种情况下,这些损坏的三元组的尾实体的排名可能排在测试三元组之上,但这不应被视为

① https://pytorch.org/

② https://github.com/seu-kse/KG_active_search

错误,因为这 2 个三元组都是真实存在的.因此,我们移除训练、验证或测试集中出现的损坏的三元组,以确保损坏的三元组不在数据集中.同样地,这样处理过后的数据,我们也统计了名次均值、排名前十(hit@10)、前三(hit@3)和第一(hit@1)的比例.详细数据见表 1 所示,在表 1 中,前者被标注为 Testing set(standard),而移除损坏三元组后的测试数据被标注为 Testing set(filtered).分析实验数据可知,该方法能较为准确地将知识图谱嵌入到低维向量空间中.

Table 1 The Result of Knowledge Graph Embedding

表 1 知识图谱嵌入过程实验结果

Testing Set	Average Ranking	hit@10/%	hit@3/%	hit@1/%
Testing Set (standard)	366.66	32.29	17.79	9.61
Testing Set (filtered)	226.58	47.75	32.80	19.31

4.2.2 主动搜索效果

实验分别统计了在按照 3.2.2 节的 1)来选择问题实体的采样方法(标准采样)与随机选择问题实体的采样方法(随机采样)中目标实体在所有邻居实体中的排名比例.该比例数值的计算方法为:

统计用户偏好嵌入到每个邻居实体的距离并进行从小到大排序,假设用户偏好嵌入到目标实体的距离排名第 rt 位,则该排名比例为 $\frac{rt}{|\mathcal{N}|}\times 100\%$,该数值越低,用户偏好的定位就越准确.实验结果如图 4 所示,显然地,在实验结果中,该比例随着横轴搜索次数的增加而降低.按照我们所设定的问题实体的标准采样方法,最终结果的平均排名百分比相比于随机采样平均要高约 2%.分析结果可知,本模型中,对问题实体的标准采样方法取得了一定的效果,相比于随机采样有一定的提升;但对于问题反馈的处理并不太理想,虽然随着搜索次数的增加,用户偏好嵌入的定位也越精准,但总体的平均水平相对不高,不能够高效地满足用户实时准确查询的需求.

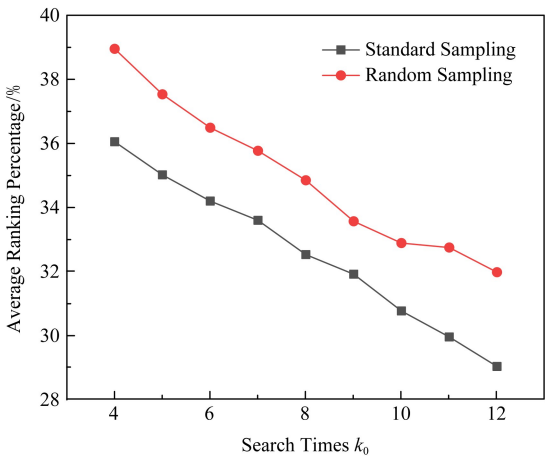


Fig. 4 Results of active search

图 4 主动搜索实验结果展示

4.3 缺陷分析与改进方法

4.3.1 缺陷分析

经过总结,本工作的不足之处主要有 3 点:

1) 所使用的知识图谱数据集在类别相关上的特点不具有普遍性.本工作使用的 FB15K 数据集,虽然是知识图谱领域的一个经典的通用数据集,但其本身的实体类别并不统一,在大多数情况下一个实体同时对应多个类别,而其每种关系的头尾实体类别并未明确标注,这给我们的工作带来了一定的困难.如在数据集中,某个演员所对应的实体同时属于“film/actor”和“tv/tv_actor”,而我们无法判定它在某个三元组里所属的具体类别.因此最终,对于一个属于多个类别的实体,我们采用了出现频率最大的类别对其进行标注.而在某个具体的上下文语境中,这个实体很可能并不属于当前类别,因此会影响结果的准确程度.

2) 用户偏好嵌入更新的步骤准确程度有所欠缺.在 3.1 节中也提到,用户偏好嵌入每次被更新时,直接在原嵌入的基础之上增加了用户偏好嵌入和问题实体所连接的向量的倍数,即用户偏好嵌入直接朝问题实体嵌入方向或相反方向移动.当更新次数达到一定值时,该嵌入值会趋于稳定,但考虑到实际情况,总体更新次数有限的情况下,该模型所输出的最终的用户偏好嵌入可能与真实的用户偏好存在一定的偏差.

3) 直接基于欧几里得距离模拟用户的评分可能并不准确.在真实情况中,用户可能对多个而非一个相关目标感兴趣.比如在我们所给出的例子中,当用户同时对演员约翰尼·德普以及电影《剪刀手爱德华》感兴趣时,本文中所提出的模拟评分机制就很难生效了.总的来说,本工作在实际应用方面的考虑有所欠缺,提出问题所基于的假设太过理想化,可能导致出现一系列的偏差或问题.

4.3.2 针对缺陷的改进方法

针对以上 3 点缺陷,分别有如下 3 种方法可以在本文的基础之上进行直接或间接提升:

1) 选取更偏重专业领域的数据集进行实验.FB15K 作为通用数据集,其涉及的领域较广,因而

实体所属的类型更为多样.而在更专业的领域,实体往往是一个专有名词,其所代表的类型也相对更少.如在专业的生物数据集中,一个蛋白质实体种类只有蛋白质而无其他,在本课题中类似这样的数据集相对更具有优势.

2) 改进现有的基于概率分布的用户偏好预测方法作为用户偏好嵌入更新步骤.在相关工作中我们介绍,目前现有的工作如文献[13]和[14],都是在用户反馈信息之后,基于概率分布来更新用户偏好.但在复现的过程中我们发现,这类模型本身运算时间长、反馈速度慢,且需要大量地向用户提问之后才有明显效果.如文献[13],该工作的实验中,其向用户提问的次数最高达到 80 次,不便于实际应用.因此在设计本课题的实现方法时,我们尝试将其更换成时间复杂度相对更小的方法,也即第 2 节中的方法,但最终效果也并不理想.因此可以通过对现有的基于概率分布的模型进行复杂度上的优化来作为本实验的算法.

3) 通过采用实体可见的数据集,我们可以模拟出具体搜索场景开发一个 Web 服务系统,通过网络招募志愿者进行人工测试来获取更精准的数据.要求志愿者选定其兴趣实体后,再对具体实体进行兴趣程度评分,系统可以通过这种方式收集更真实的数据作为测试集以提升模型准确率和可用性.

5　总结与展望

5.1　总　结

本文提出了一种人机混合的知识图谱主动搜索方法,基于知识图谱嵌入和用户在当前查询会话中的历史记录来解决模糊查询问题.在对知识图谱进行嵌入时,我们采用了 TransE 算法.在设计的主动搜索方法中,系统通过主动向用户提问,让其对某个实体进行兴趣程度的评价来获取信息以明确用户的目标.在提问的过程中,系统所选择的实体基于该实体本身的受欢迎程度、所属类别、与当前实体所连接的关系、与历史查询记录之间的关系等特征,通过加权的方式得到.在获取一定的信息之后,系统通过一种特定的移动方法更新用户的兴趣偏好嵌入,最终通过度量用户兴趣偏好嵌入与其他实体的欧几里得距离来度量用户对某个实体的感兴趣程度排名.在此过程中,我们设计了一种训练方法,对于每一条数据,该方法都能够训练到每一个邻居实体的权重,提高了训练的效率.在训练和测试时,我们提出了一种基于欧几里得距离的方法以模拟用户对某特定实体的评价.

在实验部分,我们对于知识图谱嵌入的效果和主动搜索过程的效果分别进行了测试.实验结果显示,知识图谱嵌入的效果良好,而关于主动搜索过程的效果值得商榷.在主动搜索过程的第 1 步骤,即问题实体推荐过程上,实验结果显示我们设计的基于有监督学习的模型效果尚可,基于该方法选择出来的实体其效果相比于随机选择有一定程度的提升,这也说明其所包含的信息量要高出所有实体的平均值;但该过程的第 2 步骤,也即用户偏好嵌入更新的过程效果欠佳,模型所得出的最终结果显示,该方法无法有效地在有限次数的提问内将采样的答案定位在所有邻居实体的前 20%,也就意味着无法有效地查询到用户理想的结果.该问题出现的可能原因是我们所设计的移动规则过于简单,每次更新直接移动了用户偏好嵌入和问题实体所连接的向量的倍数,而并没有在概率上考虑到用户偏好点的分布情况.因此该步骤的具体方法还需要进一步分析、考量和改进.

5.2　展　望

本节提出了一些与本工作相关的未来可能研究方向,具体如下:1)基于复杂拓扑结构信息的主动搜索系统.在实际使用基于 SPARQL 查询的系统时,复杂查询占所有查询的很大一部分.复杂的 SPARQL 查询包括了一些集合操作如取交、取并、取反等等.在低维空间中,这些不同的查询链形成了不同的拓扑结构,如链状、树状(有向无环图)、环状以及复杂状等,文献[32]对不同的基本拓扑结构进行了分类列举.这些拓扑结构包含其本身的结构信息,目前已有的利用结构信息优化该类型查询的研究如文献[28],通过对同类型的构成有向无环图的查询对应的嵌入进行训练,来提高该类查询的准确率.若将底层拓扑结构的相关信息应用于主动搜索系统,即,找到搜索过程(或单个搜索会话)中的同类结构或普遍结构的拓扑特性,则可以构造监督学习模型来学习类似结构的特征,优化相应的主动搜索过程.2)利用 Hash 算法提高搜索过程效率的主动搜索系统.如文献[13]采用的方法,现有的基于连续向量的主动搜索方法需要计算用户兴趣嵌入的后验分布,而这一计算过程基于 Markov Chain,通过 Monte Carlo 抽样法进行.显然,计算该后验分布是复杂且低效的,

因而面临着很大的计算挑战.而 Hash 学习的方法从高维的输入数据中学习压缩的二进制码,它通过测量海明距离(Hamming distance)而不是欧氏距离或点积来提高效率.在海明空间中,每个节点都被 Hash 化,即由原本的实数编码转换为了 0-1 编码.如果将前述主动学习算法应用于嵌入到海明空间而非低维向量空间的背景下,其运算速度在理论上将会大大提升,能够在损失有限的精确度的前提下,大幅度地提高运算的效率,降低运行时间.

参 考 文 献

[1] Singhal A. Introducing the knowledge Graph: Things, not strings [EB/OL]. [2019-01-01]. https://www.blog.google/products/search/introducing-knowledge-graph-things-not/

[2] Krishnan A. Making search easier: How Amazon's product graph is helping customers find products more easily [EB/OL]. [2019-01-01]. https://blog.aboutamazon.com/innovation/making-search-easier

[3] Noy N, Gao Y Q, Jain A, et al. Industry-scale knowledge graphs: Lessons and challenges [J]. Communications of the ACM, 2019, 62(8): 36-43

[4] Paulheim H. Knowledge graph refinement: A survey of approaches and evaluation methods [J]. Semantic Web, 2016, 8(3): 489-508

[5] Pujara J, Miao Hui, Getoor L, et al. Knowledge graph identification [C] //Proc of the Int Semantic Web Conf. Piscataway, NJ: IEEE, 2013: 542-557

[6] Liu Qiao, Li Yang, Duan Hong, et al. Knowledge graph construction techniques [J]. Journal of Computer Reaserch and Development, 2016, 53(3): 582-600 (in Chinese)
(刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述[J]. 计算机研究与发展, 2016, 53(3): 582-600)

[7] Codd E F. The Relational Model for Database Management [M]. Version 2. Boston: Addison-Wesley, 1990

[8] Han Jing, Haihong E, Le Guan, et al. Survey on NoSQL database [C] //Proc of the 6th Int Conf on Pervasive Computing and Applications. Piscataway, NJ: IEEE, 2011: 363-366

[9] Angles R, Gutierrez C. Survey of graph database models [J]. ACM Computing Surveys, 2008, 40(1): 1-39

[10] Bizer C, Schultz A. The Berlin SPARQL benchmark [J]. International Journal on Semantic Web and Information Systems, 2009, 5(2): 1-24

[11] Elbassuoni S, Ramanath M, Weikum G. Query relaxation for entity-relationship search [C] //Proc of the Extended Semantic Web Conf. Amsterdam: Elsevier, 2011: 62-76

[12] Garnett R, Krishnamurthy Y, Xiong Xuehan, et al. Bayesian optimal active search and surveying [C] //Proc of the 29th Int Conf on Machine Learning. New York: ACM, 2012: 843-850

[13] Canal G, Massimino A K, Davenport M A, et al. Active embedding search via noisy paired comparisons [C] //Proc of the 36th Int Conf on Machine Learning. New York: ACM, 2019: 902-911

[14] Klyuchnikov N, Mottin D, Koutrika G, et al. Figuring out the user in a few steps: Bayesian multifidelity active search with cokriging [C] //Proc of the 25th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2019: 686-695

[15] Bordes A, Usunier N, Garciaduran A, et al. Translating embeddings for modeling multi-relational data [C] //Proc of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2013: 2787-2795

[16] Nickel M, Tresp V, Kriegel H. A three-way model for collective learning on multi-relational data [C] //Proc of the Int Conf on Machine Learning. New York: ACM, 2011: 809-816

[17] Hogan A, Harth A, Umbrich J, et al. Searching and browsing linked data with SWSE: The semantic Web search engine [J]. Journal of Web Semantics, 2011, 9(4): 365-401

[18] Yahya M, Berberich K, Ramanath M, et al. Exploratory querying of extended knowledge graphs [J]. The VLDB Endowment, 2016, 9(13): 1521-1524

[19] Yahya M, Berberich K, Elbassuoni S, et al. Robust question answering over the Web of linked data [C] //Proc of the 22nd ACM Int Conf on Information & Knowledge Management. New York: ACM, 2013: 1107-1116

[20] Zhang Xinbo, Zou Lei, Hu Sen. An interactive mechanism to improve question answering systems via feedback [C] //Proc of the 28th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2019: 1381-1390

[21] Cox Ingemar J, Miller Matthew L, Minka T, et al. An optimized interaction strategy for Bayesian relevance feedback [C] //Proc of the 1998 IEEE Computer Society Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 1998: 553-558

[22] Tong S, Chang E Y. Support vector machine active learning for image retrieval [C] //Proc of the 9th ACM Int Conf on Multimedia. New York: ACM, 2001: 107-118

[23] 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

[24] 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. Menlo Park, CA: AAAI, 2015: 2181-2187

[25] Yang Bishan, Yih W T, He Xiaodong, et al. Embedding entities and relations for learning and inference in knowledge bases [C] //Proc of the Int Conf on Learning Representations. Cambridge, MA: MIT Press, 2015

[26] Nickel M, Rosasco L, Poggio T. Holographic embeddings of knowledge graphs [C] //Proc of the 30th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2016: 1955-1961

[27] Liu Zhiyuan, Sun Maosong, Lin Yankai, et al. Knowledge representation learning: A review [J]. Journal of Computer Reaserch and Development, 2016, 53(2): 247-261 (in Chinese)
(刘知远, 孙茂松, 林衍凯, 等. 知识表示学习研究进展[J]. 计算机研究与发展, 2016, 53(2): 247-261)

[28] Hamilton W L, Bajaj P, Zitnik M, et al. Embedding logical queries on knowledge graphs [C] //Proc of the Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2018: 2026-2037

[29] Wang Meng, Wang Ruijie, Liu Jun, et al. Towards empty answers in SPARQL: Approximating querying with RDF embedding [C] //Proc of the Int Semantic Web Conf. Piscataway, NJ: IEEE, 2018: 513-529

[30] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks [C] //Proc of the 13th Int Conf on Artificial Intelligence and Statistics. Cambridge MA: JMLR, 2010: 249-256

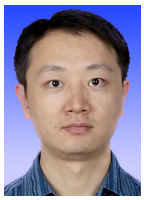
[31] Xie Ruobing, Liu Zhiyuan, Sun Maosong. Representation learning of knowledge graphs with hierarchical types [C] //Proc of the Int Joint Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2016: 2965-2971

[32] Wu Buwen, Zhou Yongluan, Yuan Pingpeng, et al. Scalable SPARQL querying using path partitioning [C] //Proc of IEEE 31st Int Conf on Data Engineering. Piscataway, NJ: IEEE, 2015: 795-806

Wang Meng, born in 1989. PhD, assistant professor. Member of CCF. His main research interests include knowledge graph, semantic search, and machine learning.

Wang Jingting, born in 2000. Master candidate. Her main research interests include knowledge graph, question answering, and natural language processing. (jtwang@seu.edu.cn)

Jiang Yinlin, born in 1999. Master candidate. His main research interests include knowledge graph, semantic search, and question answering. (yljiang@seu.edu.cn)



Qi Guilin, born in 1977. PhD, professor, PhD supervisor. Member of CCF. His main research interests include knowledge representation, logical reasoning, and natural language processing. (gqi@seu.edu.cn)