专利提案表（2018版）

**一、基本信息（加粗信息必填）**

|  |  |
| --- | --- |
| 提案名称：一种基于实体、属性和关系路径推理的知识图谱问答系统 | |
| 发明人: 黄勇、邱雪涛、佘萧寒、张琦、王宇、万四爽、费志军 | |
| 承办人：黄勇 | 所属单位：电子支付研究院 |
| 座机: 20631812 | 手机: 13161788676 |
| 第一发明人: 黄勇 | 第一发明人身份证: 420321199206023812 |
| 涉及的项目（如有）： | 项目上线时间: |
| 共同专利申请人（如有）: | |

**★务必如实填写发明人信息,《专利法》规定，发明人（设计人）：指发明创造的实质性特点作出创造性贡献的人**

**二、构想的来源**

（什么原因引发了本专利提案的构思。如：（1）在完成项目的过程中，发现技术难点，为了克服技术难点。（2）在使用产品过程中发现产品的缺陷，为了克服缺陷。（3）对竞争对手或行业热点技术检索分析后进行的技术补充或者完善。（4）与技术标准相关的技术方案。）

随着互联网和人工智能技术的快速发展，越来越多的公司、科研机构构建了大规模知识图谱，如谷歌公司构建的Knowledge Graph，微软公司提出的ConceptGraph，百度知识图谱，阿里电商知识图谱等。这些公司利用这些构建的知识图谱构建智能问答系统等，苹果的Siri、微软小娜、IBM的Waston等都在问答系统中集成知识图谱技术。

在客服自动问答、公司业务知识图谱等系统建设中，构建基于知识图谱的问答系统，不仅能在公司内部为业务人员服务，帮助他们快速了解业务知识，也可以用在客服问答系统中，为需要帮助的顾客找到关键的领域知识，解答客户的疑问。

当前构建基于知识图谱的问答系统存在很多的问题，第一个面临的挑战是如何构建结合通用常识、公司内部业务的知识图谱，并如何使用这些知识图谱提供各种服务，如为自动问答、知识推理等应用提供服务；其次是如何能对用户提问的逻辑进行推理，回答常识性、具有复杂逻辑的提问。对于给定的问题“苹果最新的价格是多少？”，“张雨绮的老公的拍过什么电影？”等，则需要结合用户信息、问题上下文、知识图谱中的隐藏逻辑进行推理，从知识图谱中获取准确的答案。其中难点之一是如何构建基于知识图谱推理的问答系统，处理复杂的问答推理系统是一个重要挑战。

**三、如何发现他人侵权**

（请说明我司可以通过什么方式发现其他公司使用了本专利提案的技术方案。如：公开渠道、技术分析、码流分析、反向工程）

目前自然语言处理领域中可投入业界使用的自动问答系统，很多都是基于知识图谱推理构建的问答系统，不同问答系统使用的推理方法具有很大的差异性。因此，若通过公开渠道发现相关的产品或方案类似的知识图谱推理特征宣传，则有可能侵权。

**四、现有技术方案**

1、现有技术实现的方法/现有技术架构

（对所有相关现有技术方案作详细的描述，主要描述最接近的现有技术的现状，可以引用行业标准、行业规范、论文、技术标准、专利文献和实际产品。）

1. 知识图谱推理

**知识图谱**是为描述客观世界中的概念、实体、事件等之间关系的一种知识表示形式。知识图谱可以用有向图进行描述，图中的“节点”描述现实中存在的概念或对象，图中“边”描述这些节点之间的关系。除了“节点”和“边”，知识图谱中也存在“属性（值）”，用来描述“节点”存在的属性。如图1展示了一个关于“中国银联”的知识图谱，知识图谱通常使用三元组进行物理存储，包括**实体关系型三元组**（中国银联，位于，上海）和**实体属性型三元组**（银联，成立于，2002年），实体关系型三元组描述实体之间的关系，实体属性型三元组描述实体具有的某些特征。

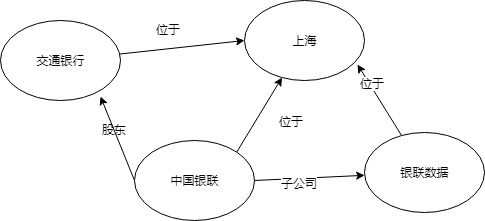


图 1 知识图谱的语义网络示意图

**知识图谱推理**指对实体和实体之间的关系进行预测。通常有两类知识图谱推理算法，一类是**基于关系路径的知识图谱推理算法**，这类方法利用实体和实体之间现有的关系路径，通过抽取实体和实体之间频繁出现的关系路径，预测实体之间可能存在的未知关系；另一类是**基于表示学习的知识图谱推理算法**，通过学习实体、关系的低维度向量表示，利用实体向量之间的数学关系，学习“头实体+关系≈尾实体”的向量表示方式，从而进行知识图谱推理。

1. 基于知识图谱的问答系统

**问答系统**输入完整和口语化的问句，通过自然语言解析、问题分类、命名实体识别和逻辑推理等复杂的内部处理系统，返回明确的问题答案或者高度精确的网页结果。问答系统不仅仅需要考虑提问者的问句，问题的上下文信息，问题的类型，需要结合问题分类、自然语言解析、知识推理等模块才能优化问答结果。除此之外，一项重要的研究是理解问题的逻辑，利用知识图谱推理技术，从知识图谱中计算出最佳的结果作为问题答案。

现有的问答系统从知识领域来说可以分为封闭领域、开放领域问答系统。**封闭领域的问答系统**专注于回答特定领域的问题，通常根据答案语料结构化信息，或问题上下文来检索正确答案。**开放领域问答系统**构建序列化的模型，利用自然语言生成等方法实现问答系统。一些问答系统也通过对问题进行分类，将可以检索获取的问题使用封闭领域问答系统获取答案，适合基于自然语言生成方法的使用开放领域问答系统进行回答。当前基于知识图谱推理的问答系统是一类封闭领域的问答系统，基于答案检索获取问题的最佳答案。依据推理方法的不同分为两类，一类是基于关系路径推理的问答系统，另一类是基于表示学习推理的问答系统。

2、现有技术存在的缺点或问题

（这里的缺点是本专利提案要解决的问题）

基于知识图谱关系路径推理的问答系统，通常利用简单的逻辑推理进行知识图谱问答，回答知识图谱中的问题。然而这些技术存在如下问题：（1）这类基于逻辑推理的问答系统通常并没有考虑问题的上下文信息，问题的提问者信息、问题的上下文环境等。如对于提问“苹果怎么样？”，这样一个问题，需要结合提问者的个人信息、提问的上下文环境确定，文中提出的苹果指的是“苹果（电影）”还是“苹果（手机）”，或者是“苹果（水果）”，才能回答出最佳的结果。（2）基于知识图谱表示学习推理的问答系统，通过将知识图谱中的实体、关系映射成为低纬度的向量表示进行推理，这类向量表示模型也很少考虑实体和实体之间存在的长的关系路径、实体存在的一些属性信息，如学习“中国银联”的向量表示时，并未考虑银联的建立日期、员工人数等属性信息，这样的向量表示模型会导致向量表示出现不精确，影响模型的预测结果。

**五、本提案所述技术方案的技术效果**

（按重要性顺序，列举并解释本专利提案的优点,技术效果需对应上述技术问题）

1．与传统的问答系统相比，本发明通过结合问题特征向量、用户信息特征向量，并结合知识图谱中的实体向量、实体属性向量、关系路径向量构建联合模型进行训练，预测知识图谱中的正确答案，有效的解决传统的问答中无法处理复杂推理的问题。

2．使用表示学习模型，学习实体的低维度向量表示；构建随机游走算法，学习关系路径的低维度向量表示；结合实体属性特征，学习实体属性的低维度向量表示。有效拓展了基于知识图谱推理的的问答系统的范围，提升答案预测的效果。

3．与传统的基于知识图谱推理的预测模型相比，本发明构建了基于Wide & Deep神经网络模型的推理系统，能更有效的利用不同类型的特征，推理的应用场景更为广泛。

**六、发明核心内容**

（简明扼要地描述整个技术方案/发明点，包含解决上述问题的必不可少的技术特征）

本发明提供了一种基于知识图谱推理的问答系统方法。给定一个提出的问题，结合提问的上下文，提问人的基本信息构建特征工程，基于向量空间模型将这些特征信息转化为向量。对于构建好的知识图谱，预先学习知识图谱中的实体向量，关系路径向量和实体属性向量。将问题和知识图谱结合训练神经网络排序模型，预测给定提问在知识图谱中的答案。

1．对于用户提问的**问题**，并结合此问题前提出的两个问题作为问题上下文，组成问题特征（query feature），使用中文分词，并去除停用词，使用查询重写（query rewriting）算法对部分错误问题重构。将这些字符串使用向量空间模型进行向量化表示。

2．对于**用户的基本信息**，如用户的职业、用户年龄、用户行为、用户兴趣爱好等信息构建成一个用户向量空间模型，和问题特征类似，将上述特征转化为用户信息向量。同时构建问题和用户信息交叉项特征，结合用户的基本信息和提问信息作为新的特征向量，构建排序模型进行训练。

3．对于知识图谱中的**实体**，利用实体和实体之间的关系，基于矩阵分解算法将“实体-关系-实体”组成的三维矩阵降维获取实体的低维度向量表示，如计算得到实体“中国银联”的五维向量为[0.1,0.8,0.3,0.4,0.6]。

4．对于知识图谱中实体和实体之间的**关系路径**，也进行低维度嵌入，计算关系路径的算法基于知识图谱中的随机游走算法。（1）给定两个实体，基于广度优先搜索或者深度优先搜索，进行随机游走，计算获取实体与实体之间存在的关系路径，如给定实体“中国银联”和“上海”，可以计算获取这两个实体之间的关系路径：“中国银联->股东->交通银行->位于->上海”、“中国银联->股东->交通银行->位于->浦东->位于->上海”。（2）获取上述的关系路径的关系路径类型: “-股东->位于-”和“-股东->位于->位于-”，并计算同一个关系特征下的所有相关关系路径类型，计算每个实体对之间的关系路径是否存在，存在即为1，否则为0。（3）构建实体和实体之间的关系路径特征矩阵，学习关系路径的低维度特征向量表示。

5．对于知识图谱中的**实体属性**，抽取实体的基本属性特征值，构建实体属性特征矩阵。如对于实体“中国银联”，抽取相关的实体属性信息如：成立日期、员工人数等实体属性值，并对于“实体-关系-实体”类型的三元组，计算头实体和尾实体之间的差值，将头实体属性、尾实体属性、头尾实体差值属性组合，构建实体属性特征矩阵。同时考虑实体属性值单位不同，将不同类型的实体属性特征值进行归一化，并学习实体属性矩阵的低维度向量表示。

6. 将上述实体向量、关系路径向量、实体属性向量、问题向量、用户信息向量作为输入向量，构建Wide & Deep神经网络模型进行打分，预测给定问题下最优的答案。通过将实体属性向量、关系路径特征向量、实体向量嵌入，并将用户信息向量和问题向量组合，作为输入层，训练神经网络模型，预测问题相关的答案。

**七、最佳技术方案（实施例）**

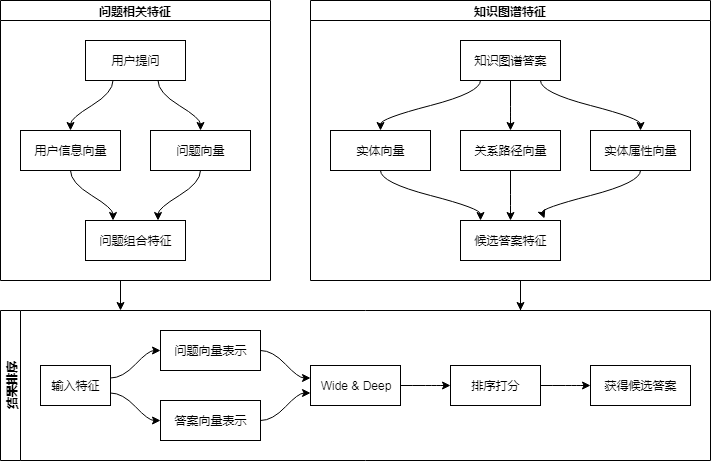


图 2问答系统的模型示意图

针对现有技术存在的不足和缺陷，本实施例公开一种基于知识图谱推理的问答系统方法，图2展示了基于知识图谱推理的问答系统中的模型示意图，图中将知识图谱中的实体向量、关系路径向量、实体属性向量、问题向量、用户信息向量作为输入向量，构建Wide & Deep神经网络模型进行排序打分，预测给定问题下最优的答案。具体实施方案的包括以下步骤。

**步骤一、问题（query）向量表示**

1.1对于用户提出的问题，基于中文分词算法对问题进行分词，并去除常见的停用词、标点符号。例如给定一个问题：“苹果的怎么样？”，使用中文分词技术将问题分词“苹果/怎么样/？”并剔除其中的停用词和标点符号。

1.2 获取问题的上下文提问，基于中文分词算法对问题进行分词，并去除常见的停用词、标点符号。对于提问“苹果的怎么样？”，获取这个提问之前的提问“苹果公司的新产品有哪些？”，同样对于上下文提问进行分词并去除停用词和标点符号。

1.3 将上述获得的问题和问题上下文使用向量空间模型转化为向量表示，获得问题相关的特征向量。

**步骤二、用户信息向量表示**

2.1 获取提问者的用户信息，如对于提问“苹果/怎么样/？”，获取提问者的基本信息如提问者的年龄，提问者的职业，提问者的兴趣爱好，提问者的用户标签等信息。

2.2 获取这些用户信息，对于获取的用户信息数据映射到向量空间中，并转化为用户信息特征向量。

2.3 对于提问信息和用户的基本信息进行交叉组合，获取新的组合特征作为训练预测模型的交叉组合特征向量。

**步骤三、实体的向量表示**

3.1 基于知识图谱中的“实体-关系-实体”类的关系三元组学习实体的低维度向量表示。对于知识图谱中所有的关系三元组如（中国银联，位于，上海），（交通银行，位于，上海），（上海，位于，中国）等转化为由“实体-关系-实体”组成的三维矩阵。

3.2 使用矩阵分解算法如RESCAL、TransE等表示学习模型，学习实体的低维度向量表示。图3展示了一个典型“实体-关系-实体”三维矩阵示意图，通过RESCAL、transE等表示学习算法可以学习实体的低维度向量表示，例如实体“中国银联”可以通过上述表示学习，获得一个5维度的特征向量[0.1, 0.5, 0.3, 0.6, 0.3]。

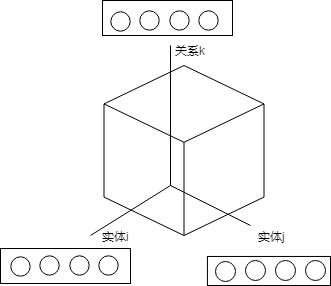


图 3一个“实体-关系-实体”三维矩阵示意图

**步骤四、关系路径向量表示**

4.1 基于知识图谱中的“实体-关系-实体”类的关系三元组学习关系路径的低维度向量表示。对知识图谱中的所有的关系，对每一个给定一个关系，我们获取知识图谱中关系下所有的实体对集合。

4.2对于该集合下所有实体对计算能连接头尾实体的关系路径，获取关系路径集合，其中表示能连接头尾实体的关系路径,l表示关系路径的长度不超过l。其中图中的路基计算算法可以基于广度优先搜索，也可以基于深度优先搜索，通常采用随机游走获取图中的关系路径。

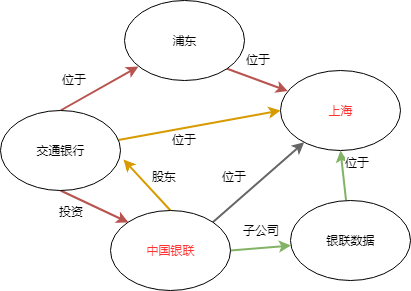


图 4 抽取知识图谱中的关系路径示意图

4.3 举例来说，对于给定的关系如“位于”，获取知识图谱中所有和“位于”相关的头尾实体对，如：（上海，中国），（中国银联，上海），（交通银行，上海）基于随机游走算法获取这些头尾实体对在知识图谱中相互连接的关系路径，如对于（中国银联，上海）这个实体对，可以基于广度优先搜索算法或深度优先搜索算法，计算抽取相关路径如：“中国银联->股东->交通银行->位于->上海”、“中国银联->股东->交通银行->位于->浦东->位于->上海”，“中国银联<-投资<-交通银行->位于->浦东->位于->上海”。对于上述“位于”关系下所有实体对的关系路径获取关系路径类型: “-股东->位于-”、“-股东->位于->位于-”、“-投资<-位于->位于-”等，图中的“->”和“<-”表示实体和实体之间的关系路径的方向。

4.4 对于给定的关系下的实体对，构建“实体对-关系路径”组成的稀疏特征矩阵，对于由“实体对-关系路径”构建的特征矩阵进行降维，获取关系路径的低维度特征向量表示。

**步骤五、实体属性的向量表示**

5.1 基于知识图谱中的“实体-属性-属性值”类的实体属性三元组学习实体属性的低维度向量表示。

5.2 对于知识图谱中的实体，枚举所有的实体属性类型，获取实体属性的类型，如对于实体“中国银联”，枚举所有的实体属性类型，如创建年月日、员工人数等。

5.3 计算获取这些实体属性类型的特征值，构建“实体-属性”组成的特征矩阵。对特征矩阵中的特征进行归一化，解决不同的属性如年龄、创建年月日、员工人数等单位不统一带来的问题。

5.4 对于构建获得的“实体-属性”特征矩阵，考虑到矩阵数据稀疏性的特点，进行降维分解，获取实体属性的低维度向量表示。

**步骤六、神经网络模型构建**

6.1 对于上述构建的不同类别的特征进行结合，构建神经网络模型，预测用户提出问题后知识图谱中候选的答案。对于上述的问题特征向量（query vector）、用户特征向量（user vector）、实体特征向量（entity vector）、关系路径特征向量（path vector）、实体属性特征向量（literal vector）进行拼接，并构建Wide & Deep神经网络训练模型。图5展示了基于上述特征构建的Wide & Deep神经网络排序模型示意图。

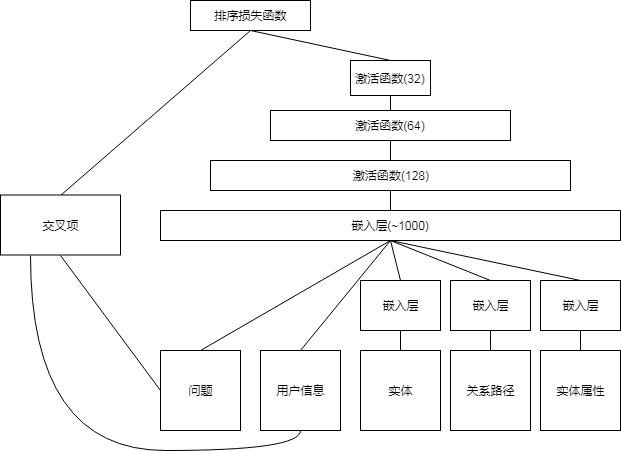


图 5 问答系统的神经网络排序模型示意图

6.2 对于上述的步骤，构建训练集，并对于标注正负例问答结果对构建预测模型，对知识图谱实体等相关特征和问题等特征进行计算。对于每个提问，计算获取相关的特征向量，构建Wide & Deep神经网络特征模型。

6.3 对于Deep层神经网络，输入层包括上述的问题特征向量、用户特征向量、实体特征向量、关系路径特征向量、实体属性特征向量等，隐藏层，其中采用ReLU激活函数，为第l层的学习权重，为偏置，为激活函数学习得到第l+1层的值，此外对不同的数据，为了能获得更好的预测结果，激活函数也可以采用softmax、sigmoid等。

6.4 对于Wide层神经网络，基于问题特征向量、用户特征向量以及两者的交叉项训练，构建逻辑回归模型。对于已构建的Wide & Deep神经网络模型，统一使用对数损失函数进行联合训练。

6.5 对于上述构建的神经网络模型，采用随机梯度优化算法学习模型的参数，模型的损失函数设为对数损失函数，模型的评价指标为AUC评价方式。其中模型的评价指标也可以选择其他分类、排序算法使用的损失函数如MAP、f1等指标，模型的优化算法也可以基于其他基于动量的梯度下降算法、基于最小批处理的梯度下降算法如Adam、RMSprop，模型的损失函数也可以选择均方误差（mean squared error）、最大边际损失（hinge loss）等。

**八、可替换技术方案、补充实施例**

（描述可替换上述技术方案的其他技术方案。例如：（1）修改该发明，也能得到实质性同样或接近的结果；（2）方法流程顺序有变化，或者有新增步骤加入；（2）装置方案中有组件替换，或者结构变形等。）

1. 上述自然语言处理方法使用jieba工具实现，中文分词的方法可替换为其他自然语言处理的工具；
2. 上述知识图谱中实体的向量表示使用TransE，TransH、TransR等表示学习算法都可用来学习实体的向量表示；
3. 上述知识图谱中关系路径的向量表示采用路径排序算法获取实体和实体之间的关系路径类型，
4. 上述的神经网络采用随机梯度优化算法学习模型的参数，可替换为其他的不同种类梯度学习算法；
5. 上述的损失函数采用对数损失函数，可替换为均方误差、边际损失等其他类型的损失函数；
6. 上述的激活函数采用ReLU激活函数，根据预测模型训练效果的不同，可以采用softmax、sigmoid等激活函数；
7. 上述的模型评价指标采用AUC进行评价，可替换为其他基于分类或排序算法使用的损失函数如MAP、f1等指标；

**九、术语说明表**

（请将本提案涉及的技术术语按照下表形式填写，尤其是发明点相关的技术术语和本发明自定义的术语）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 英文缩写 | 英文全称 | 中文全称 | 中文解释 | 其他说明 |
| TransE | Translating Embedding |  | 知识图谱中实体和关系的低维度向量表示 |  |
| Wide & Deep | Wide & Deep  Neural Network |  | 一种包含Wide和Deep两个组件的神经网络模型 | 使用Wide和Deep组件，使得训练得到的模型能够同时获得记忆和泛化能力 |
| Random Walk | Random Walk | 随机游走算法 | 发现图中节点和节点之间关系的算法 | 给定图和出发点，随机选择邻居结点，移动到邻居结点，并把新结点作为出发点重复游走，获取关系路径 |