(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112084312 A
(43)申请公布日 2020.12.15

(21)申请号 202010718229.4

(22)申请日 2020.07.23

(71)申请人 江苏海洋大学
地址 222000 江苏省连云港市新浦区苍梧路59号

(72)发明人 李慧 张舒 鲁尧 施珺 杨玉 樊宁 仲兆满 胡文彬 王国金

(74)专利代理机构 北京和联顺知识产权代理有限公司 11621
代理人 闫超良

(51)Int.Cl.
G06F 16/332(2019.01)
G06F 16/36(2019.01)
G06Q 30/00(2012.01)

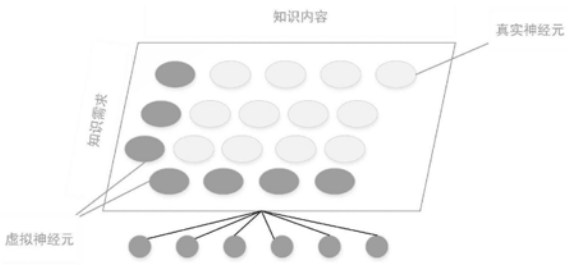
权利要求书4页 说明书11页 附图4页

(54)发明名称

一种基于知识图构建的智能客服系统

(57)摘要

本发明涉及网络数据搜索技术领域，具体提出了一种基于知识图构建的智能客服系统，通过利用问题-答案对的特征来确定问题-答案对的更精确位置，为问题和答案档案构建了一个知识图谱，该方法不仅构建了知识图谱，而且还提供了有效使用知识图谱的新方法，问答文档的特征由问题和答案组成，用于使地图的两个维度有意义，知识图谱在横向和垂直方向都被扩展，特别是在垂直膨胀期间，后续层的结构保持稳定，并提出了一种合并机制以避免稀疏性，LabelSOM选择每个神经元的特征词进行导航，并提取典型的Q&A文档以使用户快速了解全部内容。



CN 112084312 A

1.一种基于知识图构建的智能客服系统,其特征在于,包括以下步骤:

S1:对问答文档进行建模,即对问题和答案分别建模,并确定问题之间和答案之间的相似性;

S2:通过构建新的ClusterSOM模型对问答文档进行聚类,该新模型中神经元与周边神经元相关,且神经元与维度也相关;

S3:结合LabelSOM算法以及运用最小割理论分类后的特征单词创建知识图。

2.根据权利要求1所述的一种基于知识图构建的智能客服系统,其特征在于:S1中,首先对所有问题和答案都经过预处理,预处理包括分词和停止单词过滤,在预处理之后,再使用TF-IDF方法对问题和答案进行建模,使用此方法,可以将文本建模为由术语以及权重组成的向量。

3.根据权利要求2所述的一种基于知识图构建的智能客服系统,其中关于使用TF-IDF方法对问题和答案进行建模的部分,其特征在于:

每个问题都作为一个整体建模;t_i^q中的权重问答对q_j^a的问题可推导如下:

$$w_{ij}^{a} = f_{ij}^{q} \times \log \frac{N}{n_i^q} \quad (1)$$

$$f_{ij}^{q} = \frac{freq_{ij}^{q}}{\max_k f_{kj}^{q}} \quad (2)$$

其中N是整个数据集中的问答文档数量,n_i^q是其中问题包含项t_i的问答文件的数量,f_{ij}^q是问答对q_j^a中问题t_i^q的归一化频率,$freq_{ij}^a$是问答对q_j^a中问题t_i^q的频率,并且最大f_{kj}^q是问答对q_j^a问题中具有最大频率的项t_i^q的频率;

使用TF-IDF方法,答案中问答对t_i^a项的权重q_j^a的导出可如下所示

$$w_{ij}^{a} = f_{ij}^{q} \times \log \frac{N}{n_i^a} \quad (3)$$

$$f_{ij}^{a} = \frac{freq_{ij}^{a}}{\max_k f_{ki}^{a}} \quad (4)$$

可以得出问题q_m和q_n之间的相似性如下:

$$\mathrm{sim}(q_m, q_n) = \frac{\sum_{i=1}^{p} (w_{im}^q \times w_{in}^q)}{\sqrt{\sum_{i=1}^{q} (w_{im}^q)^2 \sum_{i=1}^{q} (w_{in}^q)^2}} \quad (5)$$

其中p是问题向量中的项数;

同样,答案a_m和a_n之间的相似性可以推导是如下:

$$\mathrm{sim}(a_m,a_n)=\frac{\sum_{i=1}^{p}(w_{im}^{a}\times w_{in}^{a})}{\sqrt{\sum_{i=1}^{a}(w_{im}^{a})^2\sum_{i=1}^{a}(w_{in}^{a})^2}} \quad (6)$$

其中q是答案向量中的项数。

4.根据权利要求1所述的一种基于知识图构建的智能客服系统，在S2中，关于构建新型ClusterSOM模型的部分，其特征在于：

首先，第0层包含三个神经元，其中一个虚拟神经元位于问题维度，答案维度中的一个虚拟神经元和一个真实神经元，这是两个虚拟神经元的交集，两个虚拟神经元存储问题和答案，以及真正的神经元存储问答对，问题维度中的虚拟神经元w_0^q，答案维度中的虚拟神经元w_0^a和真实神经元w_0^{qa}的突触权重被初始化为输入向量的平均值，如下所示：

$$w_0^q=\frac{1}{N}\sum_{1\le i\le N}X_i^q \quad (7)$$

$$w_0^a=\frac{1}{N}\sum_{1\le i\le N}X_i^a \quad (8)$$

$$w_0^{qa}=\frac{1}{N}\sum_{1\le i\le N}X_i^{qa} \quad (9)$$

其次，对知识图进行训练；首先，我们构造一个小的二维SOM，即ClusterSOM；它能在第0层以下的第1层中包含2×2真实神经元，在问题维度中包含2个虚拟神经元和在回答维度中包含2个虚拟神经元；再以λ步长训练第1层，参数λ确定训练层的迭代次数，在确定参数λ的值时，同时考虑匹配一致性和时间；

当问题维度的虚拟神经元的突触权重最接近问题向量时，问题的向量在问题维度被标记为虚拟神经元，对于问题的向量x_i^q，问题维度中的虚拟神经元可以如下得出：

$$c^q(t):\left\|x_i^q(t)-w_c^q(t)\right\|=\min i\left\{\left\|x_i^q(t)-w_c^q(t)\right\|\right\} \quad (10)$$

其中$c^q(t)$是当前学习迭代中答案维度中第i个虚拟神经元的突触权重向量；

使用相同的方法，将答案向量标记为答案维度中最接近的虚拟神经元，答案维度中获胜的神经元可以如下得出：

$$c^a(t):\left\|x_i^a(t)-w_c^a(t)\right\|=\min j\left\{\left\|x_i^a(t)-w_c^a(t)\right\|\right\} \quad (11)$$

其中$c^a(t)$是当前学习迭代中答案维度中第j个虚拟神经元的突触权重向量；

训练知识图后，还需要对知识图拓展，并由平均量化误差mqe确定，问题维度中虚拟神经元i的平均量化误差mqe_q，可以计算如下：

$$mqe_i^q=\frac{1}{|U_a|}\sum_{i=1}Xmqe_i^q \quad (12)$$

其中Xq_i是训练向量的集合，标记了问题维度的第i个虚拟神经元，再通过平均图的问题

维度中每个神经元的平均量化误差,可以得出问题维度的mqe_m^q,如下所示:

$$mqe_m^q = \frac{1}{|U_q|}\sum_{i=1} mqe_i^q \tag{13}$$

其中U_q是问题维度中的神经元集合;

如果图中的mqe_m^q不小于上层问题尺寸中相应父神经元的量化误差的某个分数τ_1^q,则图的问题尺寸必须横向扩展;因此,在具有最高平均量化误差的神经元与其最不相似的邻居之间的图的问题维度中插入了新行;

新插入的神经元1_q的权重是通过平均其邻居B_1^q的权重得出的:

$$wql(t) = \frac{1}{|B_l^q|}\sum_{k\in B_l^q}^{n} w_k^q\ (t) \tag{14}$$

该过程一直持续到$mqe_m^q < \tau_1^q \times mqe_0^q$,$\tau_1^q$参数越小,地图的问题维度越大;

同样,可以按以下方式得出答案维的平均量化误差mqe_m^a:

$$mqe_m^a = \frac{1}{|U_a|}\sum_{j\in U_l^a}^{n} mqe_j^a = \frac{1}{|U_a|}\sum_{j\in U_l^a}(|Y_j^a|\sum_{i\in Y_j^a}\|x_i^a - w_j^a\|) \tag{15}$$

其中Y_j^a是标记答案维度的第j个虚拟神经元的训练向量的集合,而U_a是答案维度中的神经元的集合;

当$mqe_m^a \geq \tau_1^q \times mqe_0^a$时,在答案维度中具有最高平均量化误差的神经元和与其最相似的邻居之间会插入一个新列,新添加的神经元1^a的权重为其邻居B_1^a的权重,其计算公式如下:

$$w_l^a(t) = \frac{1}{|B_l^q|}\sum_{k\in B_l} w_k^q\ (t) \tag{16}$$

最后,重复第三步,不停地拓展,直到不需要任何层的神经元扩展方可结束。

5.根据权利要求1所述的一种基于知识图构建的智能客服系统,在S3中,

关于选取特征单词的部分,其特征在于:

LabelSOM算法用于查找每个虚拟神经元的特征词;首先,映射到虚拟问题神经元的向量中每个单词k_q的量化误差推导如下:

$$q_{ik}^q = \sqrt{\sum_{x_j\in x_l^q}(w_{ik}^q - x_{jk}^q)^2} \tag{17}$$

其中X_i^q是标记到问题维度的第i个虚拟神经元的训练向量集;相应地,

映射到虚拟答案神经元的向量中每个单词ka的量化误差推导如下:

$$q_{ik}^{a} = \sqrt{\sum_{x_j \in x_i^q} (w_{ik}^{a} - x_{jk}^{a})^2} \quad (18)$$

X_i^a 是标记到答案维度的第i个虚拟神经元的训练向量的集合，选择量化误差接近于0且大于权重阈值的单词作为特征单词。

6.根据权利要求5所述的一种基于知识图构建的智能客服系统，在S3中，关于分类特征单词以及构建知识图的部分，其特征在于：

假设一个文档中有n个句子x.x2，xi...n，将其分为两类即C_1，C_2；最小割计算公式如下：

$$CUT(S,T) = \sum_{x \in s \subseteq C_1} IND_2(x) + \sum_{x \in s \subseteq C_2} IND_1(x) + \sum_{x \in S, x_k \in T} assoc(x_i, x_k) \quad (19)$$

其中，$\sum_{x \in s \subseteq C_1} IND_2(x)$ 表示所有S中句子属于C2的概率；

$\sum_{x \in s \subseteq C_2} IND_1(x)$ 表示所有T中句子属于C1的概率；

$\sum_{x \in S, x_k \in T} assoc(x_i, x_k)$ 表示S，T的关联得分；

$S \cup T = C_1 \cup C_2$；

$S \cap T = NB$；

还有，

$IND_1(x) = \Pr_{sab}^{NB}(x)$

$IND_2(x) = 1 - IND_1(x)$ (20)

$$assoc(x_i, x_k) = \begin{cases} f(i,k) \cdot c & \text{当f(i,k)} \le \text{T时;} \\ 0, & \text{否则} \end{cases} \quad (21)$$

式(21)是由Navie Bayes分类器分类所得的句子x属于观点句集合的概率，其中，参数T为两个句子有邻近关系的距离阈值，大小可调，距离大于T表明两个句子间无邻近关系；函数f(i,k)＝e-li-i是有关句子物理距离的非递增函数；参数c为一个常量，c值越小，表明分类算法更容易将两个邻近句子分到两个类别中；根据上述得分，结合最小割算法，通过二次分类将特征单词分为观点单词和非观点单词两类；

知识图分类的标准是minCUT(S,T)，根据最小割计算公式，将问答文档中的句子抽象为一种特殊的网络图，该图为无向图且每条变得容量为第一次分类所得概率以及关联得分，即c(e)＝IND1(x)或c(e)＝IND1(x)或assoc(xi,xj)，流量f(e)＝1；最终，把结合了LabelSOM算法和运用最小割理论分类后的特征单词所创建的独特的网络称为知识图。

一种基于知识图构建的智能客服系统

技术领域

[0001] 本发明涉及网络数据搜索技术领域，具体涉及一种基于知识图构建的智能客服系统。

背景技术

[0002] 互联网的快速发展，尤其是在Web2.0时代，极大地增加了在线获取的知识量，例如Yahoo问答和知乎提供了共享知识的在线平台。用户在这些网站上发布问题(知识需求)并通过自由回答其他人的问题来分享他们的知识。传统的问答系统分为问句处理和答案检索两大部分。其中问句处理的基础是分词，然而用分词在处理一些专业领域的长文本名词进很容易造成名词的割裂。处理该问题的普遍思路是人工构建专业字典。该方式消耗大量的人力资料。迄今为止，大多数提议的方法都集中在搜索策略上，例如搜索相似的现有问题或重用现有的答案来解决未回答的问题。这种方法假定用户头脑中有一个主题或某些关键字，并且可以准确，真实地表达他们的知识需求。但是，大多数用户，尤其是新手，无法用准确的词清楚地表达他们的需求。

[0003] 近年来，大规模的高质量知识图谱发展迅速，并在许多领域得到了广泛的应用，典型的包括如Freebase、DBpedia等英文知识图谱以及中文知识图谱。由于知识的结构化形式，知识图谱已经成为开放领域问答的重要资源，越来越多的研究工作也集中在知识图谱问答上。

[0004] 在维基百科的官方词条中:知识图谱是Google用于增强其搜索引擎功能的知识库。本质上，知识图谱是一种揭示实体之间关系的语义网络，可以对现实世界的事物及其相互关系进行形式化地描述。知识图谱是通过将应用数学、图形学、信息可视化技术、信息科学等学科的理论与方法与计量学引文分析、共现分析等方法结合，并利用可视化的图谱形象地展示学科的核心结构、发展历史、前沿领域以及整体知识架构达到多学科融合目的的现代理论。它能为学科研究提供切实的、有价值的参考。考虑到机构知识组织的个性化认知，基于刘等人的社会分类构建了知识图谱。Hao等人研究了领域知识图的构建和重要知识的识别他创建了一个地图，可以在其中从重要知识到不重要知识进行分层浏览。专注于虚拟实践社区中的知识管理，整合了信息检索和文本聚类方法以在Lin和Hsueh(2006年)中创建知识图谱。在虚拟的实践社区中，另一种类型的知识图谱是通过增长分层单元结构来构建的。每个簇的主题均通过LabelSOM算法选择。这些方法丰富了知识图谱的构建过程。但是，它们不适合为问答档案构建知识图谱。这些方法侧重于存储知识内容并将每个文档视为一个整体的传统文档。但是，问答文档由一个问题部分和一个答案部分组成，描述了知识的不同方面。因此，将一个问与答对视为一个单元，不仅使知识需求与知识内容混为一谈，而忽略了基于知识需求和知识内容维度可以更准确地定位问与答文档的事实。

[0005] 近几年人工智能技术逐渐应用到呼叫中心行业中，针对新渠道，在系统功能、技术、提供服务上均发生改变，进一步解放企业人力成本，这种添加人工智能技术的新渠道是目前新客服的典型代表。目前智能客服的应用方式有三种:在线智能客服、热线智能客服、

实体机器人客服。热线智能客服和实体机器人客服这两种方式比在线智能客服多了语音处理功能，虽然目前因识别技术发展相对成熟，但是各类方言和口音问题还是会给语音识别的准确率带来一定的影响。而在线智能客服多数为文字直接输入目前应用十分广泛，因此接下来主要深入探讨以文字直接输入的在线智能客服。虽然智能客服应用比较广泛，很多大型企业也已经搭建或者正在尝试搭建在线智能客服系统，但是根据一些企业用户的反馈，我们也发现了目前在线智能客服发展存在的一些问题。首先，用户将其信息需求以问题的方式提交给系统，并等待智能客服给出答案。智能客服根据问题，选取对应的解决方案来回答，以帮助用户解决问题。上面的方法着重于通过查询找到问题或答案并表现出良好的性能，并且它们在可以将知识需求清楚地表达为查询的条件下定位问题，从而找到了搜索策略。但是，当无法表达知识需求或必须确定问与答文档的分发时，则需要使用浏览策略并提出相应的方法。

[0006] 综上所述，当前的知识图构建方法处理传统文档并将每个文档视为一个整体，但是问答文档与仅包含知识内容的传统文档不同。每个问与答对不仅代表答案中的知识内容，还代表由问题表示的知识需求。因此，在知识图的构建过程中将问与答对视为单个单元无法区分知识需求和知识内容。

[0007] 为了解决这个问题，本文提出了一种构造问答档案知识图谱的方法。。

发明内容

[0008] 本发明的目的在于提供一种基于知识图构建的智能客服系统，以解决现有技术中存在的问题。

[0009] 问答档案的知识图如附图1所示，它包括知识需求维度和知识内容维度。在知识需求维度中，问题被聚类。在知识内容维度中，答案是聚集的。为了方便描述知识图的构建，神经元一词用于表示集群，因为知识图是通过扩展 SOM构建的。此外，两个边界处的簇(仅包含问与答文档的问题部分或答案部分)被称为虚拟神经元。内部簇存储整个问与答文档，它们被称为真实神经元。每个问与答对都映射到真实的神经元上。相应维度中的问题或答案将存储在虚拟神经元中。结合最小割理论再分类并利用所提出的知识图谱，可以从知识需求维度和知识内容维度中快速找到存在于问答中的知识。

[0010] 如图2所示，在知识图的构建中使用了三个主要步骤。首先，对问答文档进行建模，即对问题和答案分别建模，并确定问题之间和答案之间的相似性。其次，将问答文档扩展的问题和答案维度集中起来，以得出知识图的结构。为了保持结构的稳定性并减轻学习负担，对上层的结构进行了调整，并提出了一种合并神经元的机制。最后，构建知识图。在标签中，选择特征词来反映每个集群的主题，并对语句分类以快速识别每个群集的主要内容。有关步骤的详细说明如下：

[0011] S1:对问答文档进行建模，即对问题和答案分别建模，并确定问题之间和答案之间的相似性；

[0012] S2:通过构建新的ClusterSOM模型对问答文档进行聚类，该新模型中神经元与周边神经元相关，且神经元与维度也相关；

[0013] S3:结合LabelSOM算法以及运用最小割理论分类后的特征单词创建知识图。

[0014] 优选的，在S1中，首先对所有问题和答案都经过预处理，预处理包括分词和停止单

词过滤，在预处理之后，再使用TF-IDF方法对问题和答案进行建模，使用此方法，可以将文本建模为由术语以及权重组成的向量。

[0015] 其中，关于使用TF-IDF方法对问题和答案进行建模的部分：

[0016] 由于问答文档中的问题和答案都是文本性的，都必须由数值表示。首先，所有问题和答案都经过预处理，包括分词和停止单词过滤。分词是将句子分为有意义的词的过程。当使用不同的语言时，分词的方法是不同的。例如，单词之间的间隔可用于划分英语句子。但是，在处理中文句子时，由于它们以字符串形式表示，单词之间没有任何边界，因此无法直接进行划分，并且开发了许多工具来分割中文句子。在预处理之后，使用TF-IDF方法对问题和答案进行建模。由于该方法易于实现并且其含义易于理解，因此被广泛使用。使用此方法，可以将文本建模为由术语以及权重组成的向量每个问题都作为一个整体建模；t_i^q 中的权重问答对q_j^a 的问题可推导如下：

[0017] $$w_{ij}^{a} = f_{ij}^{q} \times \log \frac{N}{n_i} \quad (1)$$

[0018] $$f_{ij}^{q} = \frac{freq_{ij}^{q}}{\max_k f_{kj}^{q}} \quad (2)$$

[0019] 其中N是整个数据集中的问答文档数量，n_i^q 是其中问题包含项t_i的问答文件的数量，f_{ij}^q 是问答对q_j^a 中问题t_i^q 的归一化频率，$freq_{ij}^a$ 是问答对q_j^a 中问题 t_i^q 的频率，并且最大f_{kj}^q 是问答对q_j^a 问题中具有最大频率的项t_i^q 的频率；

[0020] 使用TF-IDF方法，答案中问答对t_i^a 项的权重q_j^a 的导出可如下所示

[0021] $$w_{ij}^{a} = f_{ij}^{q} \times \log \frac{N}{n_i^a} \quad (3)$$

[0022] $$f_{ij}^{a} = \frac{freq_{ij}^{a}}{\max_k f_{ki}^{a}} \quad (4)$$

[0023] 可以得出问题q_m和q_n之间的相似性如下：

[0024] $$\mathrm{sim}(q_m, q_n) = \frac{\sum_{i=1}^{p} (w_{im}^q \times w_{in}^q)}{\sqrt{\sum_{i=1}^{q} (w_{im}^q)^2 \sum_{i=1}^{q} (w_{in}^q)^2}} \quad (5)$$

[0025] 其中p是问题向量中的项数；

[0026] 同样，答案a_m和a_n之间的相似性可以推导是如下：

[0027] $$\mathrm{sim}(a_m, a_n) = \frac{\sum_{i=1}^{p} (w_{im}^a \times w_{in}^a)}{\sqrt{\sum_{i=1}^{a} (w_{im}^a)^2 \sum_{i=1}^{a} (w_{in}^a)^2}} \quad (6)$$

[0028] 其中q是答案向量中的项数。

[0029] 在S2中，关于构建新型ClusterSOM模型的部分：

[0030] 知识图包括问题维度和答案维度，是通过对问题进行聚类得出的一种图谱。尽管SOM模型及其扩展模型(例如IESOM)可以将高维输入数据映射到低维映射上，但是使用这两个模型毫无意义。传统SOM模型的映射中的每个神经元仅与其邻居有关，而与维度无关。因此，在本节中，我们提出了一种新颖的，增长的，分层的二维SOM(ClusterSOM)模型。

[0031] 比较ClusterSOM模型的主要改进与SOM模型如下。在SOM中，尺寸即边界是没有意义的，神经元只与其边界有关邻居，但不是他们的尺寸。在提出的模型中，神经元不仅与它们的邻居有关，而且与维度有关。使用此模型，知识图谱的两个维度是有意义的，并且每个神经元不仅与其邻居而且还与维度有关。这两个维度由问题和答案的分布表示。只能通过使用SOM浏览每个神经元来定位。但是，在提出的模型中，可以从两个方面找到知识问题维度和答案维度，并为更详细的导航。在ClusterSOM中，下一层绝对是新建成的。

[0032] 下面简要总结了新型ClusterSOM模型的过程：

[0033] 首先，第0层包含三个神经元，其中一个虚拟神经元位于问题维度，答案维度中的一个虚拟神经元和一个真实神经元，这是两个虚拟神经元的交集，如图3所示。两个虚拟神经元存储问题和答案，以及真正的神经元存储问答对，问题维度中的虚拟神经元w_0^q，答案维度中的虚拟神经元w_0^a和真实神经元w_0^{qa}的突触权重被初始化为输入向量的平均值，如下所示：

[0034] $$w_0^q = \frac{1}{N} \sum_{1 \le i \le N} X_i^q \tag{7}$$

[0035] $$w_0^a = \frac{1}{N} \sum_{1 \le i \le N} X_i^a \tag{8}$$

[0036] $$w_0^{qa} = \frac{1}{N} \sum_{1 \le i \le N} X_i^{qa} \tag{9}$$

[0037] 其次，对知识图进行训练；首先，我们构造一个小的二维SOM，即 ClusterSOM；例如，它可能在第0层以下的第1层中包含2×2真实神经元，在问题维度中包含2个虚拟神经元和在回答维度中包含2个虚拟神经元，如图所示，在图4中。我们以λ步长训练第1层。参数λ确定训练层的迭代次数。大量的参数导致神经元和输入向量之间的更好匹配。但是，这将花费更多时间。因此，在确定参数λ的值时，需要同时考虑匹配一致性和时间；

[0038] 当问题维度的虚拟神经元的突触权重最接近问题向量时，问题的向量在问题维度被标记为虚拟神经元，对于问题的向量x_i^q，问题维度中的虚拟神经元可以如下得出：

[0039] $$c^q(t): \left\| x_i^q(t) - w_c^q(t) \right\| = \min i \left\{ \left\| x_i^q(t) - w_c^q(t) \right\| \right\} \tag{10}$$

[0040] 其中$c^q(t)$是当前学习迭代中答案维度中第i个虚拟神经元的突触权重向量；

[0041] 使用相同的方法，将答案向量标记为答案维度中最接近的虚拟神经元，答案维度中获胜的神经元可以如下得出：

[0042] $$c^a(t):\left\|x_i^a(t)-w_c^a(t)\right\|=\min j\{\left\|x_i^a(t)-w_c^a(t)\right\|\} \quad (11)$$

[0043] 其中$c^a(t)$是当前学习迭代中答案维度中第j个虚拟神经元的突触权重向量；

[0044] 训练知识图后，还需要对知识图拓展，并由平均量化误差mqe确定，问题维度中虚拟神经元i的平均量化误差mqe_q，可以计算如下：

[0045] $$mqe_i^q=\frac{1}{|U_a|}\sum_{i=1}Xmqe_i^q \quad (12)$$

[0046] 其中Xq_i是训练向量的集合，标记了问题维度的第i个虚拟神经元，再通过平均图的问题维度中每个神经元的平均量化误差，可以得出问题维度的mqe_m^q，如下所示：

[0047] $$mqe_m^q=\frac{1}{|U_q|}\sum_{i=1}mqe_i^q \quad (13)$$

[0048] 其中U_q是问题维度中的神经元集合；

[0049] 如果图中的mqe_m^q不小于上层问题尺寸中相应父神经元的量化误差的某个分数τ_1^q，则图的问题尺寸必须横向扩展；因此，在具有最高平均量化误差的神经元与其最不相似的邻居之间的图的问题维度中插入了新行；

[0050] 新插入的神经元1_q的权重是通过平均其邻居B_1^q的权重得出的：

[0051] $$wql(t)=\frac{1}{|B_l^q|}\sum_{k\in B_l^q}^{n}w_k^q(t) \quad (14)$$

[0052] 该过程一直持续到$mqe_m^q<\tau_1^q\times mqe_0^q$，$\tau_1^q$参数越小，地图的问题维度越大；

[0053] 同样，可以按以下方式得出答案维的平均量化误差mqe_m^a：

[0054] $$mqe_m^a=\frac{1}{|U_a|}\sum_{j\in U_i^a}^{n}mqe_j^a=\frac{1}{|U_a|}\sum_{j\in U_i^a}\left(\left|Y_j^a\right|\sum_{i\in Y_j^a}\left\|x_i^a-w_j^a\right\|_c\right) \quad (15)$$

[0055] 其中Y_j^a是标记答案维度的第j个虚拟神经元的训练向量的集合，而U_a是答案维度中的神经元的集合；

[0056] 当$mqe_m^a\geq\tau_1^q\times mqe_0^a$时，如图6所示，在答案维度中具有最高平均量化误差的神经元和与其最相似的邻居之间会插入一个新列，新添加的神经元1^a的权重为其邻居B_1^a的权重，其计算公式如下：

[0057] $$w_l^a(t)=\frac{1}{|B_l^q|}\sum_{k\in B_l}w_k^q(t) \quad (16)$$

[0058] 最后，重复第三步，不停地拓展，直到不需要任何层的神经元扩展方可结束。

[0059] 尽管通过问答文件的组织可以方便地找到问答文件，但是很难解释每个集群，尤其是对于那些几乎没有有关问答档案的先行信息的新手。因此，构建知识图谱使其更易于理解。在提出的问答档案知识地图中，导航主要涉及两个维度。用户可以找到两个维度的虚拟神经元。那么，相交的真实神经元就是所需的问答档案集。同时，选择每个虚拟神经元的特征词来解释虚拟神经元，再运用最小割理论对特征单词进行分类。在真实的神经元中，可以通过分类后的特征单词快速识别文档和构建知识图。

[0060] 在S3中，关于选取特征单词的部分：

[0061] LabelSOM算法用于查找每个虚拟神经元的特征词；首先，映射到虚拟

[0062] 问题神经元的向量中每个单词k_q的量化误差推导如下：

[0063] $$q_{ik}^{q} = \sqrt{\sum_{x_j \in x_i^q} (w_{ik}^{q} - x_{jk}^{q})^2} \tag{17}$$

[0064] 其中X_i^q是标记到问题维度的第i个虚拟神经元的训练向量集；相应地，映射到虚拟答案神经元的向量中每个单词ka的量化误差推导如下：

[0065] $$q_{ik}^{a} = \sqrt{\sum_{x_j \in x_i^q} (w_{ik}^{a} - x_{jk}^{a})^2} \tag{18}$$

[0066] X_i^a是标记到答案维度的第i个虚拟神经元的训练向量的集合，选择量化误差接近于0且大于权重阈值的单词作为特征单词。

[0067] 关于最小割理论，如果对于一个(有)无向连通网络，去掉一个边集可以使．其变成两个连通分量，则这个边集就是割集。最小割集就是权和最小的割集。

[0068] 割(简称制集)是网络G＝<V,E>中顶点集合V的一个划分，它把网络中的顶点集合V划分成两个顶点集合S和T，记为[S,T]。T＝V-S，符号[S,T]＝{<s，t>∈E且s∈S,t∈T)。S中没有人度的顶点称为“源点”，T中没有出度的顶点称为“汇点”。图6(a)中[S,T]＝{<2,4>,<3,5>}．图6(c)中[S,T]＝{<1，2),<2,3),<3,5>}。图6中顶点1为源点，顶点5为汇点。图6(a)．(c)为割，(b)不是割，因为源点和汇点没有分开。对于图中每条边上的二元组，第1坐标为边的容量，记为c(e)，第2坐标为边的流量，记为f(e)，如c(<2,3>)＝2，f(<2,3>)＝1。从s中的顶点指向T中顶点的边称为正向割边，否则称为负向割边。图2的正向割边集＝(<2.4>.<3,5>}，负向割边集＝0。图6(c)的正向割边集{<1,2>,<3,5>}，负向割边集＝{<2,3>}。所有正向割边容量和称为割的容量(简称容量，记为c(S,T))，不同割图割的容量不同，如图6(a)的容量为7，图6(c)的容量为8。同理可以定义割的正向流量记为f(S,T)，割的负向流量，记为f(T,S)。割的网络流＝f(S,T)-f(T,S)。割的正向流量割的负向流量、割的网络流统称为网络的流。

[0069] 在S3中，关于分类特征单词以及构建知识图的部分，其特征在于：

[0070] 假设一个文档中有n个句子x.x2,xi...n，将其分为两类即C_1,C_2；最小割计算公式如下：

[0071] $$CUT(S,T) = \sum_{x \in s \subseteq C_1} IND_2(x) + \sum_{x \in s \subseteq C_2} IND_1(x) + \sum_{x \in S, x_k \in T} assoc(x_i, x_k) \tag{19}$$

[0072] 其中，$\sum_{x \in s \subseteq C_1} IND_2(x)$表示所有S中句子属于C2的概率；

[0073] $\sum_{x \in s \subseteq C_2} IND_1(x)$表示所有T中句子属于C1的概率；

[0074] $\sum_{x \in S, x_k \in T} assoc(x_i, x_k)$表示S,T的关联得分；

[0075] $S \cup T = C_1 \cup C_2$；

[0076] $S \cap T = NB$；

[0077] 还有，

[0078]
$$IND_1(x) = \Pr_{sab}^{NB}(x)$$
$$IND_2(x) = 1 - IND_1(x) \qquad (20)$$

[0079]
$$assoc(x_i, x_k) = \begin{cases} f(i,k) \cdot c & \text{当f(i,k)} \le \text{T时;} \\ 0, & \text{否则} \end{cases} \qquad (21)$$

[0080] 式(21)是由Navie Bayes分类器分类所得的句子x属于观点句集合的概率，其中，参数T为两个句子有邻近关系的距离阈值，大小可调，距离大于T表明两个句子间无邻近关系；函数f(i,k)＝e-li-i是有关句子物理距离的非递增函数；参数c为一个常量，c值越小，表明分类算法更容易将两个邻近句子分到两个类别中；根据上述得分，结合最小割算法，通过二次分类将特征单词分为观点单词和非观点单词两类；

[0081] 知识图分类的标准是minCUT(S,T)，根据最小割计算公式，将问答文档中的句子抽象为一种特殊的网络图，该图为无向图且每条变得容量为第一次分类所得概率以及关联得分，即c(e)＝IND1(x)或c(e)＝IND1(x)或assoc(xi,xj)，流量f(e)＝1；最终，把结合了LabelSOM算法和运用最小割理论分类后的特征单词所创建的独特的网络称为知识图。

[0082] 与现有技术相比，本发明的有益效果是：本方案不仅构建了知识图谱，而且还提供了有效使用知识图谱的新方法，问答文档的特征由问题和答案组成，用于使地图的两个维度有意义，知识图谱在横向和垂直方向都被扩展，特别是在垂直膨胀期间，后续层的结构保持稳定，并提出了一种合并机制以避免稀疏性，LabelSOM选择每个神经元的特征词进行导航，并提取典型的Q &A文档以使用户快速了解全部内容。

附图说明

[0083] 为了更清楚地说明本发明实施例的技术方案，下面将对实施例描述所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0084] 图1为本发明问答档案的知识图；

[0085] 图2为本发明知识图构建流程图；

[0086] 图3为本发明知识图第0层图；

[0087] 图4为本发明知识图第一层图；

[0088] 图5为本发明扩展答案维度示意图；

[0089] 图6为本发明最小割理论示意图；

[0090] 图7为知识库样例图。

具体实施方式

[0091] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其它实施例，都属于本发明保护的范围。

[0092] 实施例

[0093] 一、实验数据

[0094] 本文实验使用的数据集为NLPCC-ICCPOL 2016KBQA (tcci.ccf.cn/conference/2016/pages/page05_evadata.html)数据集。该数据集是目前最大的公开中文知识图谱问答数据集，其包含14,609个问答对的训练集和包含9870个问答对的测试集。并提供一个知识库，包含6,502,738个实体、587,875个属性以及43,063,796个三元组。知识库文件中每行存储一个事实(fact)，即三元组(实体、属性、属性值)。知识库样例附图7所示；

[0095] 这些问答数据集将首先进行预处理。在分词中，使用了在处理中文单词时很流行的解霸分词软件包。在停用词的过滤中，使用了许多流行的中文停用词列表，例如四川大学机器智能实验室的停用词库，百度的停用词列表和哈尔滨工业大学的停用词列表。然后，如第3.2节所述训练知识图的结构。在训练过程中，初始2DSOM的大小为2×2，学习率最初设置为0.3，并且随着时间的推移而减小。学习速率确定了获胜虚拟神经元的权重向量的更新幅度。较低的值会导致更准确和稳定，但更新过程较慢，而较低的值会导致更新过程加快，但由于更新幅度较大，因此可能无法生成准确且稳定的网络。因此，该值是通过多次重复的实验和分析来平衡速度和精度来确定的。在研究中，首先使用较大的学习率值，然后得出获胜虚拟神经元矢量的相应权重。然后，该值减小，并且减小的值再次用于计算获胜虚拟神经元的向量的权重。重复执行该减小和计算的过程，直到获胜的虚拟神经元矢量的导出权重不变为止。然后，将该值选择为学习率的最终值。

[0096] 二、实验设置

[0097] 对于上述表示模型，本文采用排序模型进行训练，该方法驱动模型输出包含在训练集中的问题实体和问题谓词对的高分，同时为不合理配对产生较低分数。在训练期间最小化的损失函数由下式给出:

[0098] $-\sum(q,p)\in C\ \max(0,S(q,p-)-S(q,p+)+\gamma)$

[0099] 因此在训练中，模型主要关注负例和正例得分之差小于边界γ的数据对，以使得正例和负例得分相差越大越好。

[0100] 三、评价标准

[0101] 精度和召回率是信息检索领域中常用的指标。在这项研究中，问题维度中聚类的精度定义为聚类中相关问题与总问题之比：

[0102] $$精度^{q}=\frac{集群中相关问题的数量}{集群中的问题数量} \qquad (22)$$

[0103] 相应地，召回率定义为问题维度中集群中相关问题与集群相关的总问题之比：

[0104] $$召回率^{q}=\frac{集群中相关问题的数量}{所有相关问题数量} \quad (23)$$

[0105] 提取典型问答文件的目的是掌握数量最少的问答文件的神经元的主要内容。因此，第一个标准涉及覆盖范围。主题的覆盖率肯定不会低于问答文件的提取比例。实际上，我们更关心提高主题的覆盖范围。因此，我们提供了以下新颖的标准：

[0106] $$提高覆盖率=\frac{(\Sigma qa_j \in \max(sim(qaj \bullet qai))) - |Nc|}{|Nc|} \quad (24)$$

[0107] 其中Ne是映射到神经元的问答文档的集合，N_c是提取的典型问答文档的集合。

[0108] 满意程度用于衡量与相应声明有关的绩效，并且还要求检查员评估该指标。较高的值表示检查员对语句的结果更满意。语句F1的满意度可得出如下：

[0109] $$F1=\frac{\sum_{i=1}^{N} r_{ij}}{N} \quad (25)$$

[0110] 其中r_{ij}是第i个检查员对语句S_jj给出的评级，并且N是检查员人数。

[0111] 四、实验结果

[0112] 对于知识图模型，本文采用的是100维的字符级别向量。ClusterSOM隐层维度为100，dropout为0.5，学习率为0.001，本文采用TD-INF方法来更新训练中的参数。实验中，随机选取10％训练数据作为验证集，结果如表 1所示，可以看出在测试集上，F1值为97.36％，取得了较好的满意度，证明该模型的有效性，也为属性选择实验提供了有效的实验结果。

[0113]

	准确率/％	召回率/％	F1/％
验证集	97.56	97.48	97.51
测试集	97.41	97.32	97.36

[0114] 表1实体抽取实验结果

[0115] 分类实验中，本文对比了特征词分类和未分类的实验结果，如表2所示。可以看出特征词未分类采用50维时并不能较好地表示问题，其实验结果甚至低于对比实验中单独采用100维的词向量的结果。对于本实验，分类特征词和未分类特征词分别取100维时得到最优实验结果，而随着维度的增加，知识图模型的F1并没有明显的提升。因此，最终本文选择分类特征词和未分类特征词维度都为100。其他参数如GRU编码器隐层维度为200，dropout设置为0.3。

[0116]

Word-Level	Char-Level/%		
	50 dims	100 dims	200 dims
50 dims	71.45	72.49	72.28
100 dims	72.34	73.96	73.90
200 dims	72.11	73.78	73.57

[0117] 表2不同维度字向量与词向量实验结果

[0118] 文同时与NLPCC官方提供的基线模型以及只采用词级别嵌入表示并通过GRU进行编码的模型进行对比。实验结果如表3所示，可以看出本文模型在最终结果上比基线模型有了很大的提高，且与只采用词级别信息表示模型相比，结合字符级别、词级别以及独热编码

信息的组合模型，更能充分对数据进行表示。相比于词级别模型，知识图模型包含更加丰富的表示信息，其包含的特征词可以更好地处理问题的语义信息，例如对问句“列克星敦号航空母舰能载多少人?”，特征词的引入能够使得属性“人员编制”的置信度更加准确，同时对于未登录词，特征词也能较好地进行处理。因此知识图谱模型可以达到比传统方法更好的实验结果。

[0119]

模型	F1	Pre@1	Pre@2	Pre@5
NLPCC	52.48	52.48	60.46	67.33
Word-Level	71.60	71.60	76.38	79.55
本文	73.96	73.96	79.45	82.51

[0120] 表3问答实验结果对比

[0121] 同时，本文也将实验结果与其他在该中文知识图谱问答数据集上进行实验的论文结果进行对比，如表4所示。前3名的满意度结果分别为 82.47％、81.59％、79.57％，且作者在实验中基本都采用了一些预定义的规则以及集成方法对模型进行优化。本文在仅使用单一神经网络模型、结构尽量简单的情况下，也取得了较好的实验结果，验证了模型的有效性。

[0122]

模型	F1/％
PKU[20]	82.47
NUDT[21]	81.59
CCNU[22]	79.57
NEU	72.72
本文	73.96

[0123] 表4不同实验结果比较

[0124] 本文提出了一种新颖的知识图谱，用于浏览问答档案。我们不仅构建了知识图谱，而且还提供了有效使用知识图谱的新方法。问答文档的特征由问题和答案组成，用于使地图的两个维度有意义。知识图谱在横向和垂直方向都被扩展。特别是在垂直膨胀期间，后续层的结构保持稳定，并提出了一种合并机制以避免稀疏性。LabelSOM选择每个神经元的特征词进行导航，并提取典型的Q&A文档以使用户快速了解全部内容。我们使用真实数据集进行了实验，结果表明该方法既可行又实用。

[0125] 由于CQA网站的成员不断发布问题和答案，因此知识图需要相应地更新。尽管知识图重建是可行的，但是由于发布和更新之间存在时间差，因此无法保证知识图与当前问答文档之间的一致性。因此，在未来的研究中，需要研究一种实时处理新的问答文档以维护最新知识图的方法。而且，对连续词空间的处理也将在未来的研究中得到更多的关注。

[0126] 在本说明书的描述中，参考术语“一个实施例”、“示例”、“具体示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本发明的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不一定指的是相同的实施例或示例。而且，描述的具体特征、结构、材料或者特点可以在任何的一个或多个实施例或示例中以合适的方式结合。

[0127] 以上公开的本发明优选实施例只是用于帮助阐述本发明。优选实施例并没有详尽叙述所有的细节，也不限制该发明仅为所述的具体实施方式。显然，根据本说明书的内容，

可作很多的修改和变化。本说明书选取并具体描述这些实施例，是为了更好地解释本发明的原理和实际应用，从而使所属技术领域技术人员能很好地理解和利用本发明。本发明仅受权利要求书及其全部范围和等效物的限制。

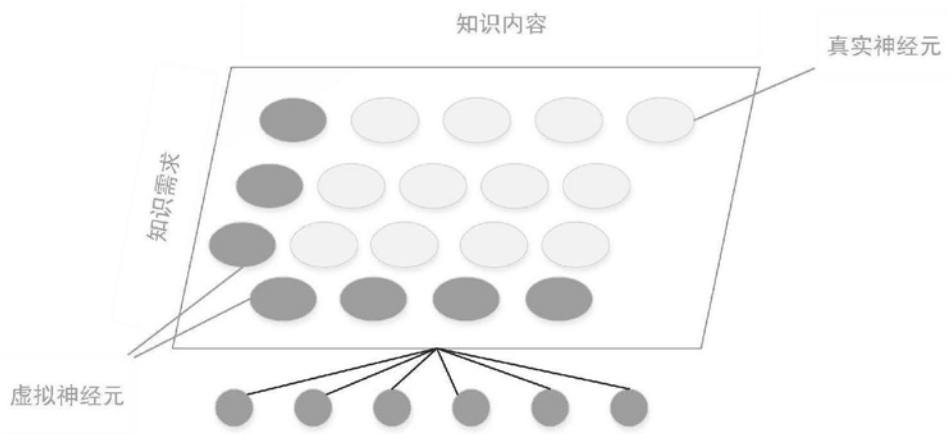

知识内容
真实神经元
知识需求
虚拟神经元

图1

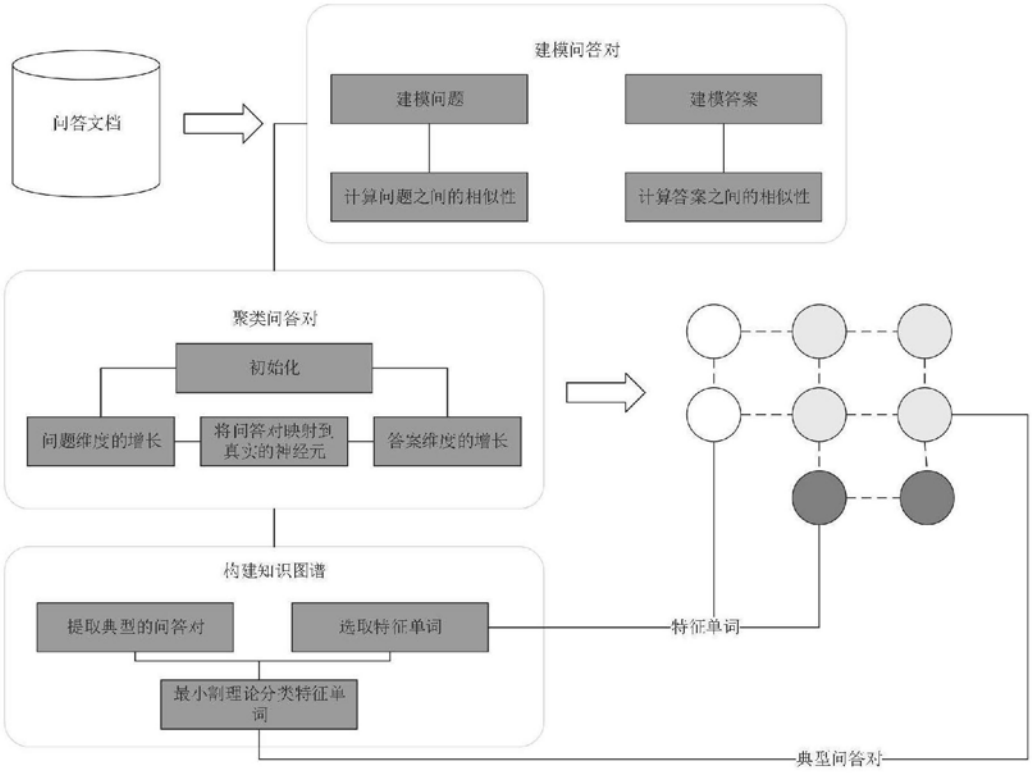

问答文档
建模问答对
建模问题
建模答案
计算问题之间的相似性
计算答案之间的相似性
聚类问答对
初始化
问题维度的增长
将问答对映射到真实的神经元
答案维度的增长
构建知识图谱
提取典型的问答对
选取特征单词
最小割理论分类特征单词
特征单词
典型问答对

图2

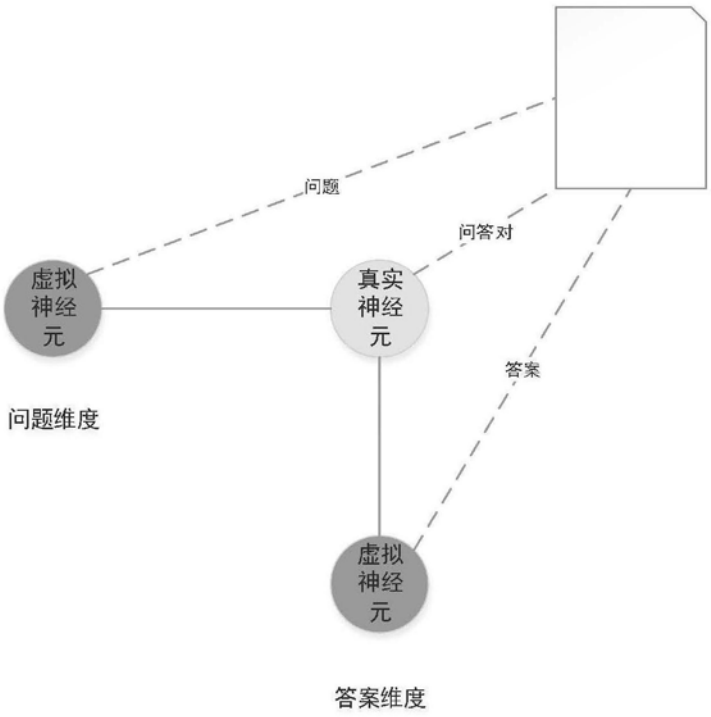


图3

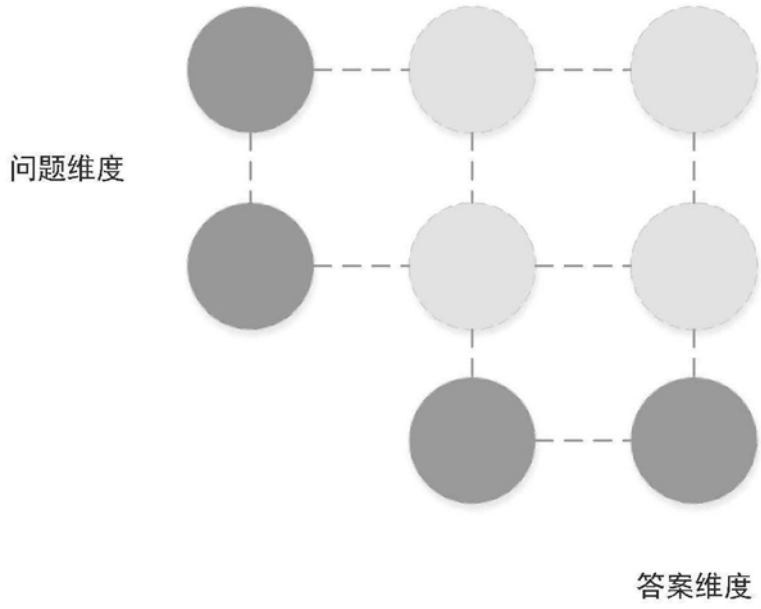


图4

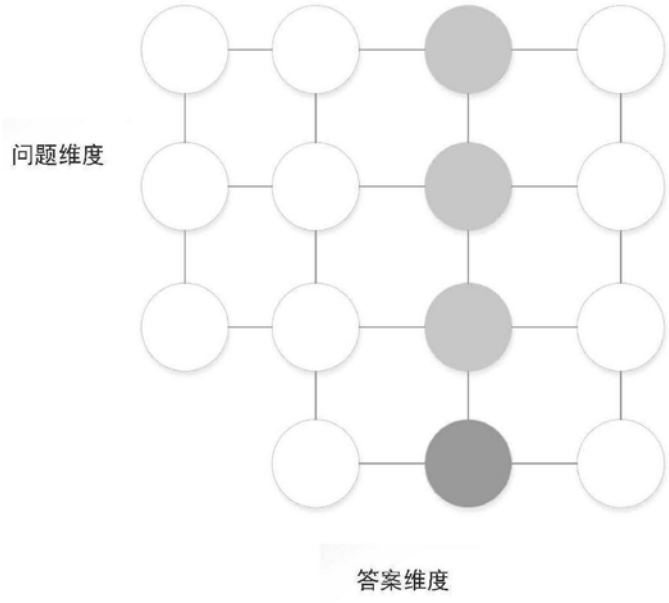


图5

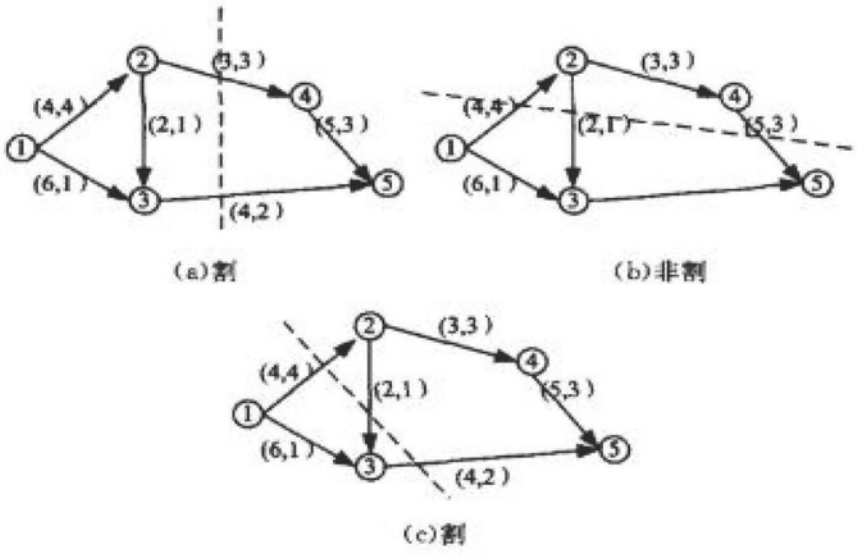


图6

```
"希望之星"英语风采大赛|||中文名|||"希望之星"英语风采大赛
"希望之星"英语风采大赛|||主办方|||中央电视台科教节目中心
"希望之星"英语风采大赛|||别名|||"希望之星"英语风采大赛
"希望之星"英语风采大赛|||外文名|||Star of Outlook English Talent Competition
"希望之星"英语风采大赛|||开始时间|||1998
"希望之星"英语风采大赛|||比赛形式|||全国选拔
"希望之星"英语风采大赛|||节目类型|||英语比赛
```

图7