



(12) 发明专利申请

(10) 申请公布号 CN 113626574 A

(43) 申请公布日 2021. 11. 09

(21) 申请号 202110952984.3

G06N 3/08 (2006.01)

(22) 申请日 2021.08.19

(71) 申请人 成都数联云算科技有限公司
地址 610041 四川省成都市中国(四川)自由贸易试验区成都高新区吉泰五路88号3栋5层8号、9号

(72) 发明人 不公告发明人

(74) 专利代理机构 成都云纵知识产权代理事务所(普通合伙) 51316
代理人 熊曦 陈婉鹃

(51) Int. Cl.
G06F 16/332 (2019.01)
G06F 16/33 (2019.01)
G06F 16/338 (2019.01)
G06N 3/04 (2006.01)

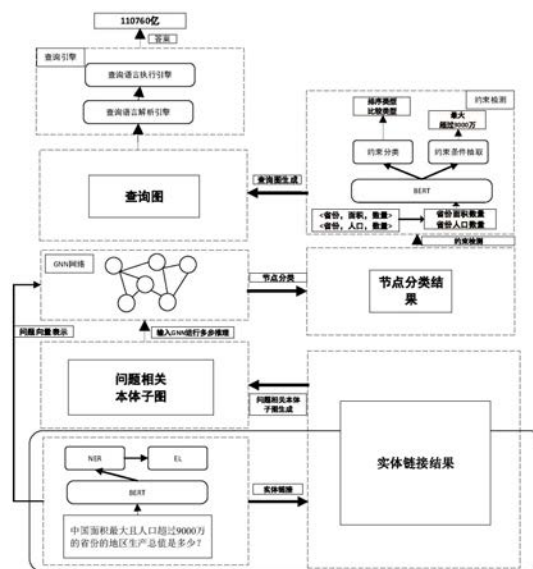
权利要求书3页 说明书12页 附图6页

(54) 发明名称

一种信息查询方法及系统及装置及介质

(57) 摘要

本发明公开了一种信息查询方法及系统及装置及介质,涉及自然语言处理领域,包括:对编码后的查询输入信息进行实体识别,生成本体子图;对所述本体子图中的节点进行分类,获得节点分类标签;生成所述本体子图中约束节点的类型和约束条件;基于所述分类标签,获得查询图;获得知识库中对应的实体集合;基于所述实体集合中的实体信息,获得所述信息查询路径的主题节点和连接节点;基于所述信息查询路径的主题节点、所述信息查询路径的连接节点、所述约束节点的类型、所述约束条件和所述查询图,生成所述信息查询路径;执行所述信息查询路径,获得与所述查询输入信息对应的结果,本方法提高了信息查询的效率。



1. 一种信息查询方法,其特征在于,所述方法包括:
 - 对编码后的查询输入信息进行实体识别,获得第一实体集合;
 - 基于所述第一实体集合中实体信息的类型,获得第一本体集合后生成本体子图;
 - 对所述本体子图中的节点进行分类,获得节点分类标签,节点类型包括:主题节点和约束节点,以及用于信息查询路径上各节点连接的连接节点;
 - 基于所述约束节点和所述信息查询路径上的其他节点,生成所述本体子图中约束节点的类型和约束条件;
 - 基于所述分类标签,获得查询图;
 - 获得知识库中与所述第一实体集合对应的第二实体集合;
 - 基于所述第二实体集合中的实体信息,获得所述信息查询路径的主题节点和连接节点;
 - 基于所述信息查询路径的主题节点、所述信息查询路径的连接节点、所述约束节点的类型、所述约束条件和所述查询图,生成所述信息查询路径;
 - 在所述知识库中执行所述信息查询路径,获得与所述查询输入信息对应的结果。
2. 根据权利要求1所述的信息查询方法,其特征在于,所述节点类型包括:主题节点、答案节点、中间节点、连接节点、约束节点和无关节点。
3. 根据权利要求2所述的信息查询方法,其特征在于,所述主题节点为所述信息查询路径的起始节点;
 - 所述答案节点为所述查询输入信息对应的结果在所述本体子图上所对应的节点;
 - 所述中间节点为所述主题节点到所述答案节点路径上的中间节点;
 - 所述约束节点为对所述主题节点到所述答案节点路径上节点进行约束的节点;
 - 所述无关节点为与所述查询输入信息对应结果查找无关的节点。
4. 根据权利要求1所述的信息查询方法,其特征在于,以所述第一本体集合中的本体类型为起始节点,在本体网络中遍历当前起始节点的N跳邻居节点,生成与所述查询输入信息对应的所述本体子图,N为推理的步数。
5. 根据权利要求1所述的信息查询方法,其特征在于,将所述本体子图中约束节点与所述信息查询路径上的节点进行组合,获得节点组合,将所述节点组合转为文本表示,获得节点组合文本,将所述节点组合文本与所述查询输入信息进行拼接,获得拼接结果;将所述拼接结果编码,获得第二编码信息;基于所述第二编码信息获得所述本体子图中约束节点的类型,从所述第二编码信息中抽取所述约束条件。
6. 根据权利要求1所述的信息查询方法,其特征在于,将所述第一实体集合中的实体通过实体链接对齐到知识库中对应的实体,获得所述第二实体集合。
7. 根据权利要求1所述的信息查询方法,其特征在于,所述约束节点的类型包括:排序类型、计数类型、比较类型和均值类型。
8. 根据权利要求1所述的信息查询方法,其特征在于,利用BERT模型对所述查询输入信息进行编码,获得所述第一编码信息,利用命名实体识别模型NER识别所述第一编码信息中的实体。
9. 根据权利要求1所述的信息查询方法,其特征在于,对于所述本体子图中的每个节点利用图神经网络得到节点的向量表示,将所述向量表示投影到节点的类别空间,利用分类

函数对节点进行分类。

10. 根据权利要求9所述的信息查询方法,其特征在于,利用图神经网络得到节点的向量表示,具体包括:

生成当前节点的邻居节点聚合信息;

基于所述邻居节点聚合信息,利用循环神经网络GRU更新得到当前节点的向量表示。

11. 根据权利要求1所述的信息查询方法,其特征在于,所述查询输入信息中的单词来源于词汇库,所述查询输入信息为 $q = \{w_1, w_2, \dots, w_{i_q}, \dots, w_{n_q}\}$, $w_{i_q} \in W^v$, $i_q = 1, 2, \dots, n_q$, n_q 表示所述查询输入信息中单词的个数, v 表示所述词汇库的大小, W 表示所述词汇库空间大小,所述第一实体集合 $E_q = \{e_1, e_2, \dots, e_{i_e}, \dots, e_{n_e}\}$,其中, $e_{i_e} = \{w_{i_e,1}, w_{i_e,2}, \dots, w_{i_e,m_e}\}$ 表示所述查询输入信息中的第 i_e 个实体, e_{i_e} 包含所述查询输入信息中的 m_e 个单词, n_e 代表所述查询输入信息中实体的个数,第一本体集合 $O_q = \{o_1, o_2, \dots, o_{i_o}, \dots, o_{n_o}\}$,其中, o_{i_o} 代表所述第一实体集合的第 i_o 个本体类型, n_o 代表所述第一实体集合的本体类型总数,所述第一本体集合中的本体类型 o_{i_o} 对应所述查询输入信息中一个或者多个实体 $\{e_{i_o,1}, e_{i_o,2}, \dots, e_{i_o,m_o}\}$ 的本体类型标签, m_o 代表本体类型 o_{i_o} 在所述查询输入信息中对应的实体个数。

12. 根据权利要求1所述的信息查询方法,其特征在于:

基于所述信息查询路径的主题节点和所述信息查询路径的连接节点在所述查询图中生成若干条第一中间查询路径;

基于所述约束节点的类型约束所述第一中间查询路径,获得若干条第二中间查询路径;

基于所述约束条件约束所述第二中间查询路径,获得所述信息查询路径。

13. 一种信息查询系统,其特征在于,所述系统包括:

实体识别单元,用于对编码后的查询输入信息进行实体识别,获得第一实体集合;

本体子图生成单元,用于基于所述第一实体集合中实体信息的类型,获得第一本体集合后生成本体子图;

节点分类单元,用于对所述本体子图中的节点进行分类,获得节点分类标签,节点类型包括:主题节点和约束节点,以及用于信息查询路径上各节点连接的连接节点;

约束节点的类型和约束条件生成单元,用于基于所述约束节点和所述信息查询路径上的其他节点,生成所述本体子图中约束节点的类型和约束条件;

查询图获得单元,用于基于所述分类标签,获得查询图;

第二实体集合获得单元,用于获得知识库中与所述第一实体集合对应的第二实体集合;

主题节点和连接节点获得单元,用于基于所述第二实体集合中的实体信息,获得所述信息查询路径的主题节点和连接节点;

信息查询路径生成单元,用于基于所述信息查询路径的主题节点、所述信息查询路径的连接节点、所述约束节点的类型、所述约束条件和所述查询图,生成所述信息查询路径;

结果获得单元,用于在所述知识库中执行所述信息查询路径,获得与所述查询输入信息对应的结果。

14. 一种信息查询装置,包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序,其特征在于,所述处理器执行所述计算机程序时实现如权利要求1-12中任意一个所述信息查询方法的步骤。

15. 一种计算机可读存储介质,所述计算机可读存储介质存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现如权利要求1-12中任意一个所述信息查询方法的步骤。

一种信息查询方法及系统及装置及介质

技术领域

[0001] 本发明涉及自然语言处理领域,具体地,涉及一种信息查询方法及系统及装置及介质。

背景技术

[0002] 智能问答是当前自然语言处理领域一个重要的研究方向,具有广阔的应用前景,随着知识库相关技术的研究,近年来不断取得新的突破,基于知识库的问答(KBQA: Knowledge Base Question Answer)技术也引起了越来越多研究者的关注。KBQA通常指通过对给定的自然语言问题进行语义理解和解析,再结合知识库中的结构化知识,推理得出答案。目前基于KBQA的智能问答通常使用基于信息检索的方法,即:直接利用问题所表达的语义信息从知识库中检索答案,例如将自然语言问题和知识库中的候选答案进行嵌入编码,再根据问题和答案的语义匹配程度查找和排序答案。基于语义解析的方法通常将自然语言问题转化成可以在知识库中进行查询的逻辑表达式,例如根据自然语言问题生成查询图,然后在知识图谱上进行查询。如对于如下的自然语言问题:

[0003] 广东省有多少人口?

[0004] 基于信息检索的方法主要包括以下几个步骤:

[0005] (1) 识别出问题中的主题词“广东省”,通过实体链接方式找到“广东省”在知识库中对应的实体,并将其邻居节点作为问题的候选答案,如图1中的“179725平方千米”,“10430万人”,“110760亿”。

[0006] (2) 使用LSTM或者Transformer等编码器对自然语言问题“广东省有多少人口?”进行编码。

[0007] (3) 在图一中,生成主题词“广东省”到候选答案的路径,如下:

[0008] <广东省,面积,179725平方千米>

[0009] <广东省,人口,10430万人>

[0010] <广东省,地区生成总值,110760亿>

[0011] 使用LSTM或者Transformer等编码器对上述路径进行编码。

[0012] (4) 计算问题编码和候选答案编码之间的语义匹配程度并进行排序,选择得分最高的候选答案作为问题的最终答案。

[0013] 基于信息检索的方法可以很好的处理简单问题的回答,但是对于添加了聚合等约束条件的复杂问题难以解决,无法根据问句的语义信息编码直接到知识库中匹配问题的答案,必须利用语义解析的方法识别出问题答案在知识库中的主题词,核心路径,约束条件等信息,对于这类问题通常利用语义解析的方法结合问句提供的信息生成查询图,再在知识库上执行查询图,查找到问题在知识库中的答案,

[0014] 目前业界主流的KBQA方法主要研究知识库中单跳简单问题的问答技术,对于多约束条件,多跳推理问题主要以问题中的实体作为种子节点在知识库中通过多步游走生成查询图,由于知识库中有的实体节点的邻居节点规模巨大,直接在实体知识库中游走效率低

下,急需一种高效的信息查询方法。

发明内容

[0015] 为了提高信息查询的效率,本发明提供了一种信息查询方法及系统及装置及介质。

[0016] 为实现上述发明目的,本发明提供了一种信息查询方法,所述方法包括:

[0017] 对编码后的查询输入信息进行实体识别,获得第一实体集合;

[0018] 基于所述第一实体集合中实体信息的类型,获得第一本体集合后生成本体子图;

[0019] 对所述本体子图中的节点进行分类,获得节点分类标签,节点类型包括:主题节点和约束节点,以及用于信息查询路径上各节点连接的连接节点;

[0020] 基于所述约束节点和所述信息查询路径上的其他节点,生成所述本体子图中约束节点的类型和约束条件;

[0021] 基于所述分类标签,获得查询图;

[0022] 获得知识库中与所述第一实体集合对应的第二实体集合;

[0023] 基于所述第二实体集合中的实体信息,获得所述信息查询路径的主题节点和连接节点;

[0024] 基于所述信息查询路径的主题节点、所述信息查询路径的连接节点、所述约束节点的类型、所述约束条件和所述查询图,生成所述信息查询路径;

[0025] 在所述知识库中执行所述信息查询路径,获得与所述查询输入信息对应的结果。

[0026] 其中,本方法找到查询输入信息中相关实体对应于知识库中的实体,然后根据所找到的实体选择其n跳的邻居节点构成本体子图,通过本体子图推理出答案候选路径得到相关答案,本方法在本体子图上推理,与传统方法直接在实体子图上进行推理更加高效快捷,避免了无目标的在知识库中庞大规模的邻居节点中游走,提供了信息查询的效率。

[0027] 优选的,本方法中所述节点类型包括:主题节点、答案节点、中间节点、连接节点、约束节点和无关节点。为了生成查询图,需要对本体子图的节点进行分类,然后再根据各个节点的类型构造查询图。对节点进行分类可以在生成查询图时根据节点的类型挑选需要的节点,然后过滤不需要的节点,可以使得减少不必要的查询推理,提高查询的效率。

[0028] 优选的,本方法中所述主题节点为所述信息查询路径的起始节点;

[0029] 所述答案节点为所述查询输入信息对应的结果在所述本体子图上所对应的节点;

[0030] 所述中间节点为所述主题节点到所述答案节点路径上的中间节点;

[0031] 所述约束节点为对所述主题节点到所述答案节点路径上节点进行约束的节点;

[0032] 所述无关节点为与所述查询输入信息对应结果查找无关的节点。

[0033] 其中,本方法定义了主题节点、答案节点、中间节点、约束节点和无关节点,通过上述的节点定义可以准确的对节点进行分类。

[0034] 优选的,本方法中以所述第一本体集合中的本体类型为起始节点,在本体网络中遍历当前起始节点的N跳邻居节点,生成与所述查询输入信息对应的所述本体子图,N为推理的步数。与现有方法的区别是,本方法以本体类型为起始节点,然后遍历相应的邻居节点,生成本体子图,本发明中的信息查询方法在本体子图上进行推理,相较于传统直接在实体子图上进行推理更加高效快捷。

[0035] 优选的,本方法中将所述本体子图中约束节点与所述信息查询路径上的节点进行组合,获得节点组合,将所述节点组合转为文本表示,获得节点组合文本,将所述节点组合文本与所述查询输入信息进行拼接,获得拼接结果;将所述拼接结果编码,获得第二编码信息;基于所述第二编码信息获得所述本体子图中约束节点的类型,从所述第二编码信息中抽取所述约束条件。通过将所述节点组合文本与查询输入信息进行拼接,然后将拼接结果编码,基于编码信息获得本体子图中约束节点的类型,从编码信息中抽取约束条件,通过上述方式,能够获得与查询输入信息对应的约束节点的类型和约束条件,进而能够通过约束节点的类型和约束条件约束查询图生成信息查询路径。

[0036] 其中,对于分类后的本体子图中的约束节点,需要进一步辨别其约束类型,抽取约束条件。

[0037] 优选的,本方法中将所述第一实体集中的实体通过实体链接对齐到知识库中对应的实体,获得所述第二实体集合。

[0038] 优选的,本方法中所述约束节点的类型包括:排序类型、计数类型、比较类型和均值类型。其中,通过对约束节点进行分类,便于生成更加准确有效的约束条件来约束查询图,进而能够更加准确高效的生成信息查询路径。

[0039] 优选的,本方法中利用BERT模型对所述查询输入信息进行编码,获得所述第一编码信息,利用命名实体识别模型NER识别所述第一编码信息中的实体。通过BERT模型能够方便对信息进行编码。

[0040] 优选的,本方法中对于所述本体子图中的每个节点利用图神经网络得到节点的向量表示,将所述向量表示投影到节点的类别空间,利用分类函数对节点进行分类。

[0041] 优选的,本方法中利用图神经网络得到节点的向量表示,具体包括:

[0042] 生成当前节点的邻居节点聚合信息;

[0043] 基于所述邻居节点聚合信息,利用循环神经网络GRU更新得到当前节点的向量表示。

[0044] 优选的,本方法中所述查询输入信息中的单词来源于词汇库,所述查询输入信息为 $q = \{w_1, w_2, \dots, w_{i_q}, \dots, w_{n_q}\}$, $w_{i_q} \in W^v$, $i_q = 1, 2, \dots, n_q$, n_q 表示所述查询输入信息中单词的个数, v 表示所述词汇库的大小, W 表示所述词汇库空间大小,所述第一实体集合 $E_q = \{e_1, e_2, \dots, e_{i_e}, \dots, e_{n_e}\}$,其中 $e_{i_e} = \{w_{i_e,1}, w_{i_e,2}, \dots, w_{i_e,m_e}\}$ 表示所述查询输入信息中的第 i_e 个实体, e_{i_e} 包含所述查询输入信息中的 m_e 个单词, n_e 代表所述查询输入信息中实体的个数,第一本体集合 $O_q = \{o_1, o_2, \dots, o_{i_o}, \dots, o_{n_o}\}$,其中 o_{i_o} 代表所述第一实体集合的第 i_o 个本体类型, n_o 代表所述第一实体集合的本体类型总数,所述第一本体集合中的本体类型 o_{i_o} 对应所述查询输入信息中一个或者多个实体 $\{e_{i_o,1}, e_{i_o,2}, \dots, e_{i_o,m_o}\}$ 的本体类型标签, m_o 代表本体类型 o_{i_o} 在所述查询输入信息中对应的实体个数。

[0045] 其中,将查询输入信息的文本信息引入到本体子图中节点邻居信息的更新,相较于传统节点的更新方法更能够关注到与查询输入信息相关的邻居节点,进而能够快速且准确的找到相关邻居节点,能够高效且准确的生成查询图,并且引入GRU更新本体子图的节点信息,更能够捕捉到本体子图中与查询输入信息相关路径的时序信息,进而能够高效快速

的生成相关路径。

[0046] 优选的,本方法基于所述信息查询路径的主题节点和所述信息查询路径的连接节点在所述查询图中生成若干条第一中间查询路径;

[0047] 基于所述约束节点的类型约束所述第一中间查询路径,获得若干条第二中间查询路径;

[0048] 基于所述约束条件约束所述第二中间查询路径,获得所述信息查询路径。

[0049] 本发明还提供了一种信息查询系统,所述系统包括:

[0050] 实体识别单元,用于对编码后的查询输入信息进行实体识别,获得第一实体集合;

[0051] 本体子图生成单元,用于基于所述第一实体集合中实体信息的类型,获得第一本体集合后生成本体子图;

[0052] 节点分类单元,用于对所述本体子图中的节点进行分类,获得节点分类标签,节点类型包括:主题节点和约束节点,以及用于信息查询路径上各节点连接的连接节点;

[0053] 约束节点的类型和约束条件生成单元,用于基于所述约束节点和所述信息查询路径上的其他节点,生成所述本体子图中约束节点的类型和约束条件;

[0054] 查询图获得单元,用于基于所述分类标签,获得查询图;

[0055] 第二实体集合获得单元,用于获得知识库中与所述第一实体集合对应的第二实体集合;

[0056] 主题节点和连接节点获得单元,用于基于所述第二实体集合中的实体信息,获得所述信息查询路径的主题节点和连接节点;

[0057] 信息查询路径生成单元,用于基于所述信息查询路径的主题节点、所述信息查询路径的连接节点、所述约束节点的类型、所述约束条件和所述查询图,生成所述信息查询路径;

[0058] 结果获得单元,用于在所述知识库中执行所述信息查询路径,获得与所述查询输入信息对应的结果。

[0059] 本发明还提供了一种信息查询装置,包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序,所述处理器执行所述计算机程序时实现所述信息查询方法的步骤。

[0060] 本发明还提供了一种计算机可读存储介质,所述计算机可读存储介质存储有计算机程序,所述计算机程序被处理器执行时实现所述信息查询方法的步骤。

[0061] 本发明提供的一个或多个技术方案,至少具有如下技术效果或优点:

[0062] 本发明中的信息查询方法在本体子图上进行推理,相较于传统直接在实体子图上进行推理更加高效快捷。

[0063] 本发明将查询输入信息的文本信息引入到本体子图中节点邻居信息的更新,相较于传统节点的更新方法更能够关注到与查询输入信息相关的邻居节点,进而能够快速且准确的找到相关邻居节点,能够高效且准确的生成查询图。

[0064] 本发明引入GRU更新本体子图的节点信息,更能够捕捉到本体子图中与查询输入信息相关路径的时序信息,进而能够高效快速的生成相关路径。

附图说明

[0065] 此处所说明的附图用来提供对本发明实施例的进一步理解,构成本发明的一部分,并不构成对本发明实施例的限定;

[0066] 图1是简单问题候选答案示意图;

[0067] 图2是复杂问题查询图示意图;

[0068] 图3是本方法架构示意图;

[0069] 图4是复杂问题知识库本体示意图;

[0070] 图5是问题相关本体的2阶邻居节点示意图;

[0071] 图6是问题相关本体子图的节点划分示意图;

[0072] 图7是本系统组成示意图。

具体实施方式

[0073] 为了能够更清楚地理解本发明的上述目的、特征和优点,下面结合附图和具体实施方式对本发明进行进一步的详细描述。需要说明的是,在相互不冲突的情况下,本发明的实施例及实施例中的特征可以相互组合。

[0074] 在下面的描述中阐述了很多具体细节以便于充分理解本发明,但是,本发明还可以采用其他不同于在此描述范围内的其他方式来实施,因此,本发明的保护范围并不受下面公开的具体实施例的限制。

[0075] 应当理解,本文使用的“系统”、“装置”、“单元”和/或“模组”是用于区分不同级别的不同组件、元件、部件、部分或装配的一种方法。然而,如果其他词语可实现相同的目的,则可通过其他表达来替换所述词语。

[0076] 如本说明书和权利要求书所示,除非上下文明确提示例外情形,“一”、“一个”、“一种”和/或“该”等词并非特指单数,也可包括复数。一般说来,术语“包括”与“包含”仅提示包括已明确标识的步骤和元素,而这些步骤和元素不构成一个排它性的罗列,方法或者设备也可能包含其它的步骤或元素。

[0077] 本说明书使用流程图说明根据本说明书的实施例的系统所执行的操作。应当理解的是,前面或后面操作不一定按照顺序来精确地执行。相反,可以按照倒序或同时处理各个步骤。同时,也可以将其他操作添加到这些过程中,或从这些过程移除某一步或数步操作。

[0078] 实施例一

[0079] 本实施例一提供了一种信息查询方法,本方法为基于本体推理的复杂问题问答方法,本方法中的查询输入信息通常为问句,也可以为其他形式的查询输入信息,本实施例不对查询输入信息的具体展现形式进行限定。

[0080] 本方法首先利用实体链接方法找到问句中相关实体对应于知识库中的实体。然后根据所找到的实体选择其n跳的邻居节点构成本体子图。通过本体子图推理出答案候选路径得到相关答案。具体步骤包括:

[0081] 问题中的实体识别与实体链接:

[0082] 对于输入的问题 $q = \{w_1, w_2, \dots, w_{i_q}, \dots, w_{n_q}\}$,其单词来源于词汇库, $w_{i_q} \in W^v$, $i_q = 1, 2, \dots, n_q$, n_q 表示问题中单词的个数, v 表示词汇库的大小, W 表示词汇库空间。首先利用BERT对输入的问题进行编码,得到问题的编码序列 x_q 和全局编码 x_{cls} ,然后将编码后的序列

x_q 经过命名实体识别模型NER检测问题中的实体,记实体集合 $E_q = \{e_1, e_2, \dots, e_{i_e}, \dots, e_{n_e}\}$,其中 $e_{i_e} = \{w_{i_e,1}, w_{i_e,2}, \dots, w_{i_e,m_e}\}$ 表示问句中的第 i_e 个实体,其包含问题中的 m_e 个单词, n_e 代表问题中实体的个数,公式表达如下:

$$[0083] \quad x_q, x_{cls} = \text{BERT}(q), x_q \in R^{n_q \times d}, x_{cls} \in R^{1 \times d}$$

$$[0084] \quad E_q = \text{NER}(x_q)$$

[0085] 其中, d 代表词向量的维度,对于实体集合 $E_q = \{e_1, e_2, \dots, e_{i_e}, \dots, e_{n_e}\}$,根据实体类型获取其本体集合 $O_q = \{o_1, o_2, \dots, o_{i_o}, \dots, o_{n_o}\}$,其中 o_{i_o} 代表问题中实体集合的第 i_o 个本体类型, n_o 代表问题中实体集合的本体类型总数,本体集合中的本体类型 o_{i_o} 对应问题中一个或者多个实体 $\{e_{i_o,1}, e_{i_o,2}, \dots, e_{i_o,m_o}\}$ 的本体类型标签, m_o 代表本体类型 o_{i_o} 在问题中对应的实体个数。

[0086] 问句相关实体的多跳本体子图生成:

[0087] 以 O_q 中的本体类型作为起始节点,在本体网络上遍历当前节点的 N 跳邻居节点,生成问题相关的本体子图 $O_s(V_s, E_s)$, N 代表推理的步数。 V_s 代表问句中提及的本体类型节点及其在本体网络中所有的 N 阶及以内的邻居节点, E_s 代表 V_s 中所有节点形成的边。

[0088] 本体子图的推理:

[0089] 本体子图的推理主要包括本体子图节点的分类以及查询图约束节点的类型划分和约束条件抽取。

[0090] (1) 本体子图的节点分类

[0091] 为了生成查询图,需要对本体子图的节点进行分类,然后再根据各个节点的类型构造查询图,具体地,将本体子图 $O_s(V_s, E_s)$ 上的节点划分为如下六类,如图2所示。

[0092] 1) 主题节点,代表答案查询核心路径的起始节点,为问句中主题词的本体类型。

[0093] 2) 答案节点,问题的答案在本体上所对应的节点。

[0094] 3) 中间节点,代表主题节点到答案节点路径上的中间节点。

[0095] 4) 连接节点,代表与答案路径上的主题节点、中间节点、答案节点相连接的本体节点,该节点需要在问句中出现,但是不必在答案查询核心路径上,主要作用是用于约束答案查询核心路径上的节点。

[0096] 5) 约束节点,通过聚合函数等对主题节点到答案节点的路径上的节点进行约束。

[0097] 6) 无关节点,代表与问题答案查找无关的节点。

[0098] 根据上述本体子图中各个节点的类型,对于简单问题,通过问题中的主题节点到答案节点的路径便可以在实体知识库中查找到具体的答案,对于存在多约束条件的问题,首先需要根据主题节点,中间节点和答案节点,生成答案的查询路径,然后根据连接节点和约束节点约束答案的范围。具体的步骤如下:

[0099] 通过在 x_q 上进行最大池化MAX_POOL,获取问句的向量表示 $c_q \in R^{1 \times d}$:

$$[0100] \quad c_q = \text{MAX_POOL}(x_q)$$

[0101] 本体子图 O_s 中每个节点的向量表示为 $h = \{h_1, h_2, \dots, h_{i_v}, \dots, h_{n_v}\}$, $h_{i_v} \in R^{1 \times d}$,其中

n_v 代表本体子图中节点的个数。初始化值可以是利用TransE等关系推理模型进行训练得到的节点向量值或者随机初始化的向量值： $\mathbf{h}^0 = \{\mathbf{h}_1^0, \mathbf{h}_2^0, \dots, \mathbf{h}_{i_v}^0, \dots, \mathbf{h}_{n_v}^0\}$ 。

[0102] 如图3,利用图神经网络GNN在本体子图中进行推理学习,更新本体子图的节点向量表示,对于图神经网络第1层的节点向量表示： $\mathbf{h}^l = \{\mathbf{h}_1^l, \mathbf{h}_2^l, \dots, \mathbf{h}_{i_v}^l, \dots, \mathbf{h}_{n_v}^l\}$,通过聚合更新得到第1+1层的节点向量表示 $\mathbf{h}^{l+1} = \{\mathbf{h}_1^{l+1}, \mathbf{h}_2^{l+1}, \dots, \mathbf{h}_{i_v}^{l+1}, \dots, \mathbf{h}_{n_v}^{l+1}\}$:

[0103] $\mathbf{h}^{l+1} = \text{GNN}(\mathbf{h}^l)$

[0104] 图神经网络GNN的具体算法主要包括对邻居节点信息的聚合和更新操作,本方法主要利用GRU神经网络更新当前节点及邻居节点的信息,每个节点具体的更新如下,首先生成当前节点的邻居节点聚合信息:

[0105] $\mathbf{g}_{i_v}^l = \sum_{j_v \in N_{i_v}} \alpha_{i_v j_v}^l \mathbf{h}_{j_v}^l, \alpha_{i_v j_v}^l \in R, \mathbf{g}_{i_v}^l \in R^{1 \times d}$

[0106] 其中: i_v 为当前节点, j_v 为 i_v 邻居节点, N_{i_v} 为当前节点 i_v 的所有邻居节点的集合, $\alpha_{i_v j_v}^l$ 为节点 i_v 和节点 j_v 的归一化权重, $\mathbf{h}_{j_v}^l$ 为 j_v 的节点向量;

[0107] $\alpha_{i_v j_v}^l = \text{softmax}(e_{i_v j_v}^l) = \frac{e_{i_v j_v}^l}{\sum_{k_v \in N(i_v)} e_{i_v k_v}^l}, e_{i_v j_v}^l \in R$

[0108] $e_{i_v j_v}^l = \text{ReLU}([\mathbf{c}_q || \mathbf{h}_{j_v}^l || \mathbf{h}_{i_v}^l] \mathbf{W}_h^T), \mathbf{W}_h \in R^{1 \times 3d}$

[0109] 其中, $e_{i_v j_v}^l$ 为节点 i_v 和节点 j_v 的权重, k_v 为 i_v 邻居节点, $N(i_v)$ 为当前节点 i_v 的所有邻居节点的集合, $||$ 代表将向量拼接,上标T代表转置矩阵。然后利用GRU网络更新当前节点的向量表示:

[0110] $\mathbf{r}^l = \sigma([\mathbf{h}_{i_v}^l || \mathbf{g}_{i_v}^l] \mathbf{W}_r^T), \mathbf{W}_r \in R^{1 \times 2d}, \mathbf{r}^l \in R$

[0111] $\mathbf{z}^l = \sigma([\mathbf{h}_{i_v}^l || \mathbf{g}_{i_v}^l] \mathbf{W}_z^T), \mathbf{W}_z \in R^{1 \times 2d}, \mathbf{z}^l \in R$

[0112] $\mathbf{g}_{i_v}^l = \mathbf{r}^l * \mathbf{g}_{i_v}^l, \mathbf{g}_{i_v}^l \in R^{1 \times d}$

[0113] $\mathbf{g}_{i_v}^{l+1} = \tanh([\mathbf{h}_{i_v}^l || \mathbf{g}_{i_v}^l] \mathbf{W}_o^T), \mathbf{W}_o \in R^{2d \times d}, \mathbf{g}_{i_v}^{l+1} \in R^{1 \times d}$

[0114] $\mathbf{h}_{i_v}^{l+1} = \mathbf{z}^l * \mathbf{g}_{i_v}^l + (1 - \mathbf{z}^l) * \mathbf{g}_{i_v}^{l+1}, \mathbf{h}_{i_v}^{l+1} \in R^{1 \times d}$

[0115] 其中, \mathbf{r}^l 、 \mathbf{z}^l 为当前节点的邻居节点聚合信息的权重, \mathbf{W}_r 、 \mathbf{W}_z 、 \mathbf{W}_o^T 、 \mathbf{W}_o 为神经网络层参数, $\mathbf{g}_{i_v}^{l+1}$ 为当前节点新增聚合信息, $\mathbf{h}_{i_v}^{l+1}$ 为第1+1层的节点向量;

[0116] 对于本体子图中的每个节点利用图神经网络推理得到的向量表示,将其投影到节点类别空间,在利用softmax分类函数对其进行分类:

[0117] $\mathbf{o}_{i_v} = \text{Linear}(\mathbf{h}_{i_v}^l), \mathbf{o}_{i_v} \in R^{1 \times m}$

[0118] $\mathbf{p}_{i_v} = \text{Softmax}(\mathbf{o}_{i_v}), \mathbf{p}_{i_v} \in R^{1 \times m}$

[0119] $c_{i_v} = \operatorname{argmax}(\mathbf{p}_{i_v}), c_{i_v} \in R$

[0120] 其中, \mathbf{o}_{i_v} 为线性变换后的输出, \mathbf{p}_{i_v} 为softmax分类得分, c_{i_v} 为分类结果, L代表图神经网络的层数, $\operatorname{Linear}()$ 代表线性变换函数, $\operatorname{Softmax}()$ 代表softmax分类器, $\operatorname{argmax}()$ 代表取最大值的下标, m代表节点的类别数目, 包括其它类别。

[0121] 约束节点的类别划分和约束条件抽取:

[0122] 对于分类后的本体子图中的约束节点, 需要进一步辨别其约束类型, 抽取约束条件。本发明将本体节点的约束类型划分为以下四类:

[0123] (1) 排序类型, 例如: 中国面积第二大的省份人口是多少?

[0124] (2) 计数类型, 例如: 中国有多少个省?

[0125] (3) 比较类型, 例如: 中国人口超过9000万的省份有哪些?

[0126] (4) 均值类型, 例如: 中国各个省份地区生产总值的平均值是多少?

[0127] 对于分类后的本体子图中的约束节点, 生成其与答案查询路径上节点所构成的三元组, 对于图6中的约束节点, 生成如表1所示的三元组及其文本表示:

[0128] 表1

三元组	文本表示
<行政区域, 面积, 数量>	行政区域面积数量
<行政区域, 人口, 数量>	行政区域人口数量

[0130] 将问题与三元组的文本表示拼接为t, 输入BERT编码模型, 对该节点的约束类型进行分类, 并抽取约束条件, 具体步骤如下:

[0131] $\mathbf{t}_s, \mathbf{t}_{cls} = \operatorname{BERT}(t), \mathbf{t}_s \in R^{n_t \times d}, \mathbf{t}_{cls} \in R^{1 \times d}$

[0132] 其中, n_t 为字符空间;

[0133] $\mathbf{t}_c = \operatorname{Linear}(\mathbf{t}_{cls}), \mathbf{t}_c \in R^{n_c \times d}$

[0134] $\mathbf{p}_c = \operatorname{Softmax}(\mathbf{t}_c), \mathbf{p}_c \in R^{n_c \times d}$

[0135] $c = \operatorname{argmax}(\mathbf{p}_c)$

[0136] 其中, c为约束类别;

[0137] $\mathbf{c}_{start} = \operatorname{softmax}(\mathbf{t}_s \mathbf{W}_s^T), \mathbf{W}_s \in R^d, \mathbf{c}_{start} \in R^{n_t}$

[0138] 其中, \mathbf{c}_{start} 为约束条件在拼接文本中的起始位置的概率分布;

[0139] $\operatorname{start} = \operatorname{argmax}(\mathbf{c}_{start}), \operatorname{start} \in R$

[0140] $\mathbf{t}_e = \operatorname{Transformer_Encoder}(\mathbf{t}_s), \mathbf{t}_e \in R^{n_t \times d}$

[0141] 其中, \mathbf{t}_e 为 \mathbf{t}_s 通过Transformer层的编码;

[0142] $\mathbf{c}_{end} = \operatorname{softmax}(\mathbf{t}_e \mathbf{W}_e^T), \mathbf{W}_e \in R^d, \mathbf{c}_{end} \in R^{n_t}$

[0143] 其中, \mathbf{c}_{end} 为约束条件在拼接文本中的结束位置的概率分布;

[0144] $\operatorname{end} = \operatorname{argmax}(\mathbf{c}_{end}), \operatorname{end} \in R$

[0145] 其中 \mathbf{t}_s 表示问题和三元组文本表示拼接后通过BERT模型编码后输出的向量序列表示, \mathbf{t}_{cls} 表示问题和三元组文本表示拼接后通过BERT模型编码后输出的全局向量表示,

Transformer_Encoder代表一层的Transformer编码器, n_c 代表约束类别的数目, t_c 代表约束类别评分, p_c 代表约束类别的概率分布, c 代表约束类别的索引下标, W_s 和 W_e 是可训练的权重参数, $start$ 和 end 代表要抽取的约束条件在拼接文本中的起始和结束位置。

[0146] 问句答案生成:

[0147] 在得到本体子图上各个节点的标签后,通过如下步骤获取问题的最终答案:

[0148] (1) 根据本体子图上各个节点的标签,生成查询图 G_s 。

[0149] (2) 将 $E = \{e_1, e_2, \dots, e_j, \dots, e_m\}$ 中的实体通过实体链接模型EL对齐到知识库 $K(V, R)$ 中相应的 $E_k = \{e_{1k}, e_{2k}, \dots, e_{jk}, \dots, e_{mk}\}$ 节点上。

[0150] $E_k = EL(E)$

[0151] (3) 以 E_k 中的实体作为查询图中答案查询核心路径的主题节点和连接节点,结合约束节点的类型和约束条件,将查询图 G_s 转化为相应知识库的查询语言,通过知识库的查询引擎执行该查询语言获取问题的答案。

[0152] 其中,本实施例中,本方法基于所述信息查询路径的主题节点和所述信息查询路径的连接节点在所述查询图中生成若干条第一中间查询路径;

[0153] 基于所述约束节点的类型约束所述第一中间查询路径,获得若干条第二中间查询路径;

[0154] 基于所述约束条件约束所述第二中间查询路径,获得所述信息查询路径。

[0155] 即首先基于所述信息查询路径的主题节点和所述信息查询路径的连接节点在所述查询图中生成较多条第一中间查询路径,然后利用约束节点的类型和约束条件逐步约束出最终的信息查询路径。

[0156] 本方法将问句信息引入到节点的邻居节点注意力权重计算中,使得当前节点更关注和问题相关的邻居节点信息。

[0157] 本方法通过引入循环神经网络GRU对节点信息进行更新,使用序列生成的方式生成路径上的节点表示更能代表当前节点之前的路径信息。

[0158] 本方法中通过约束节点和查询路径中的节点构成的三元组生成文本表示,并与问题相拼接,使用BERT等编码器编码后用于约束类型的分类和约束条件的抽取。

[0159] 实施例二

[0160] 在实施例一的基础上,实施例二举例对本方法进行详细说明:

[0161] 对于问句“中国面积最大且人口超过9000万的省份的地区生产总值是多少?”的解决,本方法主要包括以下几步:

[0162] (1) 使用BERT模型对问句进行编码,得到问句的向量表示 $c_q = [0.12, 0.23, \dots, 0.44]$ 。

[0163] (2) 识别出问题中的实体“中国”的本体类型“国家”,在图4的本体图谱 O 上,以该本体类型为起点,遍历其 N 阶邻居节点, N 为推理的跳数,假设 $N=2$,可以得到如图5中的本体子图 S 。

[0164] (3) 初始化本体子图 S 中的各个节点向量表示,并结合问句的向量表示信息 c_q ,使用图神经网络模型GNN进行节点向量表示的聚合和更新。

[0165] (4) 利用softmax分类器,对子图 S 中的各个节点进行分类,将划分为主题节点,中间节点,连接节点,约束节点,无关节点,如图6。

[0166] (5) 找到问题的约束字段和约束所使用的聚合函数集合：

[0167] {<X?, 面积, max (?)>, <X?, 人口, gt (9000万)>}

[0168] 其中max (?) 表示取该类节点值得最大值, gt (9000万) 表示所有值大于9000万的该类节点。

[0169] (6) 生成如图2的查询图, 在知识库上执行查询图得到问题的答案。

[0170] 实施例三

[0171] 请参考图7, 本发明实施例三提供了一种信息查询系统, 所述系统包括：

[0172] 实体识别单元, 用于对编码后的查询输入信息进行实体识别, 获得第一实体集合；

[0173] 本体子图生成单元, 用于基于所述第一实体集合中实体信息的类型, 获得第一本体集合后生成本体子图；

[0174] 节点分类单元, 用于对所述本体子图中的节点进行分类, 获得节点分类标签, 节点类型包括：主题节点和约束节点, 以及用于信息查询路径上各节点连接的连接节点；

[0175] 约束节点的类型和约束条件生成单元, 用于基于所述约束节点和所述信息查询路径上的其他节点, 生成所述本体子图中约束节点的类型和约束条件；

[0176] 查询图获得单元, 用于基于所述分类标签, 获得查询图；

[0177] 第二实体集合获得单元, 用于获得知识库中与所述第一实体集合对应的第二实体集合；

[0178] 主题节点和连接节点获得单元, 用于基于所述第二实体集合中的实体信息, 获得所述信息查询路径的主题节点和连接节点；

[0179] 信息查询路径生成单元, 用于基于所述信息查询路径的主题节点、所述信息查询路径的连接节点、所述约束节点的类型、所述约束条件和所述查询图, 生成所述信息查询路径；

[0180] 结果获得单元, 用于在所述知识库中执行所述信息查询路径, 获得与所述查询输入信息对应的结果。

[0181] 实施例四

[0182] 本发明实施例四提供了一种信息查询装置, 包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序, 所述处理器执行所述计算机程序时实现所述信息查询方法的步骤。

[0183] 其中, 所述处理器可以是中央处理器 (CPU, Central Processing Unit), 还可以是其他通用处理器、数字信号处理器 (digital signal processor)、专用集成电路 (Application Specific Integrated Circuit)、现场可编程门阵列 (Fieldprogrammable gate array) 或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

[0184] 所述存储器可用于存储所述计算机程序和/或模块, 所述处理器通过运行或执行存储在所述存储器内的数据, 实现发明中信息查询装置的各种功能。所述存储器可主要包括存储程序区和存储数据区, 其中, 存储程序区可存储操作系统、至少一个功能所需的应用程序 (比如声音播放功能、图像播放功能等) 等。此外, 存储器可以包括高速随机存取存储器、还可以包括非易失性存储器, 例如硬盘、内存、插接式硬盘, 智能存储卡, 安全数字卡, 闪存卡、至少一个磁盘存储器件、闪存器件、或其他易失性固态存储器件。

[0185] 实施例五

[0186] 本发明实施例五提供了一种计算机可读存储介质,所述计算机可读存储介质存储有计算机程序,所述计算机程序被处理器执行时实现所述信息查询方法的步骤。

[0187] 所述信息查询装置如果以软件功能单元的形式实现并作为独立的产品销售或使用,可以存储在一个计算机可读存储介质中。基于这样的理解本发明实现上述实施例方法中的全部或部分流程,也可以通过计算机程序可存储于一计算机可读存介质中,该计算机程序在被处理器执行时,可实现上述各个方法实施例的步骤。其中,所述计算机程序包括计算机程序代码、对象代码形式、可执行文件或某些中间形式等。所述计算机可读介质可以包括:能够携带所述计算机程序代码的任何实体或装置、记录介质、U盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器、随机存储器、点载波信号、电信信号以及软件分发介质等。需要说明的是,所述计算机可读介质包含的内容可以根据司法管辖区内立法和专利实践的要求进行适当的增减。

[0188] 本发明已对基本概念做了描述,显然,对于本领域技术人员来说,上述详细披露仅作为示例,而并不构成对本说明书的限定。虽然此处并没有明确说明,本领域技术人员可能会对本说明书进行各种修改、改进和修正。该类修改、改进和修正在本说明书中被建议,所以该类修改、改进、修正仍属于本说明书示范实施例的精神和范围。

[0189] 同时,本说明书使用了特定词语来描述本说明书的实施例。如“一个实施例”、“一实施例”、和/或“一些实施例”意指与本说明书至少一个实施例相关的某一特征、结构或特点。因此,应强调并注意的是,本说明书中在不同位置两次或多次提及的“一实施例”或“一个实施例”或“一个替代性实施例”并不一定是指同一实施例。此外,本说明书的一个或多个实施例中的某些特征、结构或特点可以进行适当的组合。

[0190] 此外,本领域技术人员可以理解,本说明书的各方面可以通过若干具有可专利性的种类或情况进行说明和描述,包括任何新的和有用的工序、机器、产品或物质的组合,或对他们的任何新的和有用的改进。相应地,本说明书的各个方面可以完全由硬件执行、可以完全由软件(包括固件、常驻软件、微码等)执行、也可以由硬件和软件组合执行。以上硬件或软件均可被称为“数据块”、“模块”、“引擎”、“单元”、“组件”或“系统”。此外,本说明书的各方面可能表现为位于一个或多个计算机可读介质中的计算机产品,该产品包括计算机可读程序编码。

[0191] 计算机存储介质可能包含一个内含有计算机程序编码的传播数据信号,例如在基带上或作为载波的一部分。该传播信号可能有多种表现形式,包括电磁形式、光形式等,或合适的组合形式。计算机存储介质可以是除计算机可读存储介质之外的任何计算机可读介质,该介质可以通过连接至一个指令执行系统、装置或设备以实现通讯、传播或传输供使用的程序。位于计算机存储介质上的程序编码可以通过任何合适的介质进行传播,包括无线电、电缆、光纤电缆、RF、或类似介质,或任何上述介质的组合。

[0192] 本说明书各部分操作所需的计算机程序编码可以用任意一种或多种程序语言编写,包括面向对象编程语言如Java、Scala、Smalltalk、Eiffel、JADE、Emerald、C++、C#、VB.NET、Python等,常规程序化编程语言如C语言、Visual Basic、Fortran 2003、Perl、COBOL 2002、PHP、ABAP,动态编程语言如Python、Ruby和Groovy,或其他编程语言等。该程序编码可以完全在用户计算机上运行、或作为独立的软件包在用户计算机上运行、或部分在

用户计算机上运行部分在远程计算机运行、或完全在远程计算机或服务器上运行。在后种情况下,远程计算机可以通过任何网络形式与用户计算机连接,比如局域网(LAN)或广域网(WAN),或连接至外部计算机(例如通过因特网),或在云计算环境中,或作为服务使用如软件即服务(SaaS)。

[0193] 此外,除非权利要求中明确说明,本说明书所述处理元素和序列的顺序、数字字母的使用、或其他名称的使用,并非用于限定本说明书流程和方法的顺序。尽管上述披露中通过各种示例讨论了一些目前认为有用的发明实施例,但应当理解的是,该类细节仅起到说明的目的,附加的权利要求并不仅限于披露的实施例,相反,权利要求旨在覆盖所有符合本说明书实施例实质和范围的修正和等价组合。例如,虽然以上所描述的系统组件可以通过硬件设备实现,但是也可以只通过软件的解决方案得以实现,如在现有的服务器或移动设备上安装所描述的系统。

[0194] 同理,应当注意的是,为了简化本说明书披露的表述,从而帮助对一个或多个发明实施例的理解,前文对本说明书实施例的描述中,有时会将多种特征归并至一个实施例、附图或对其的描述中。但是,这种披露方法并不意味着本说明书对象所需要的特征比权利要求中提及的特征多。实际上,实施例的特征要少于上述披露的单个实施例的全部特征。

[0195] 针对本说明书引用的每个专利、专利申请、专利申请公开物和其他材料,如文章、书籍、说明书、出版物、文档等,特此将其全部内容并入本说明书作为参考。与本说明书内容不一致或产生冲突的申请历史文件除外,对本说明书权利要求最广范围有限制的文件(当前或之后附加于本说明书中的)也除外。需要说明的是,如果本说明书附属材料中的描述、定义、和/或术语的使用与本说明书所述内容有不一致或冲突的地方,以本说明书的描述、定义和/或术语的使用为准。

[0196] 最后,应当理解的是,本说明书中所述实施例仅用以说明本说明书实施例的原则。其他的变形也可能属于本说明书的范围。因此,作为示例而非限制,本说明书实施例的替代配置可视为与本说明书的教导一致。相应地,本说明书的实施例不仅限于本说明书明确介绍和描述的实施例。

[0197] 尽管已描述了本发明的优选实施例,但本领域内的技术人员一旦得知了基本创造性概念,则可对这些实施例作出另外的变更和修改。所以,所附权利要求意欲解释为包括优选实施例以及落入本发明范围的所有变更和修改。

[0198] 显然,本领域的技术人员可以对本发明进行各种改动和变型而不脱离本发明的精神和范围。这样,倘若本发明的这些修改和变型属于本发明权利要求及其等同技术的范围之内,则本发明也意图包含这些改动和变型在内。

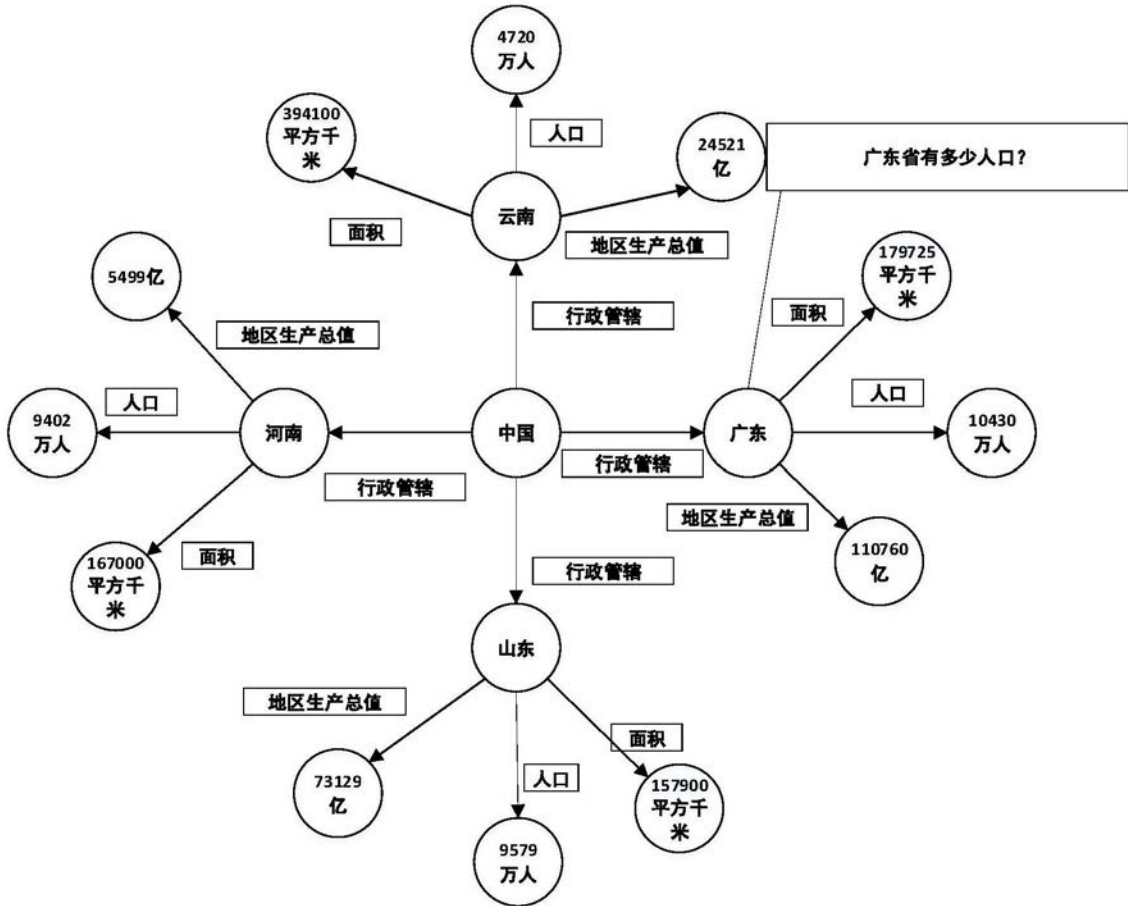


图1

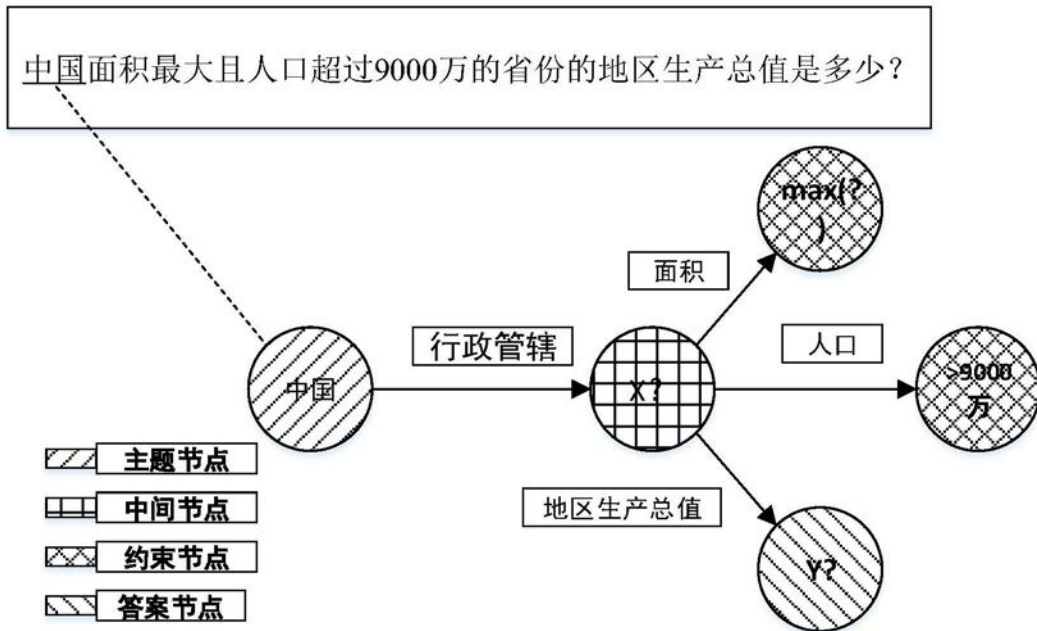


图2

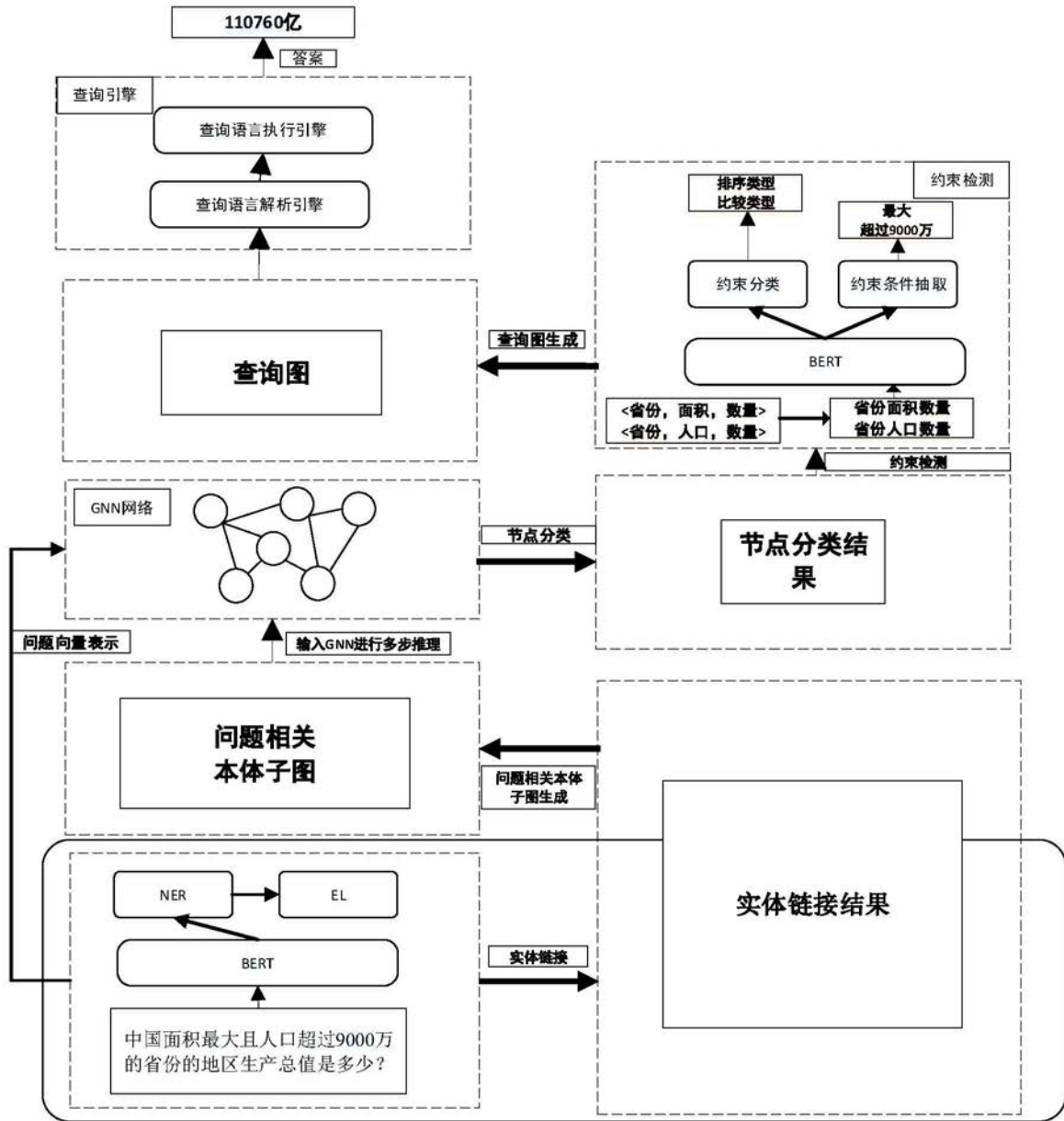


图3

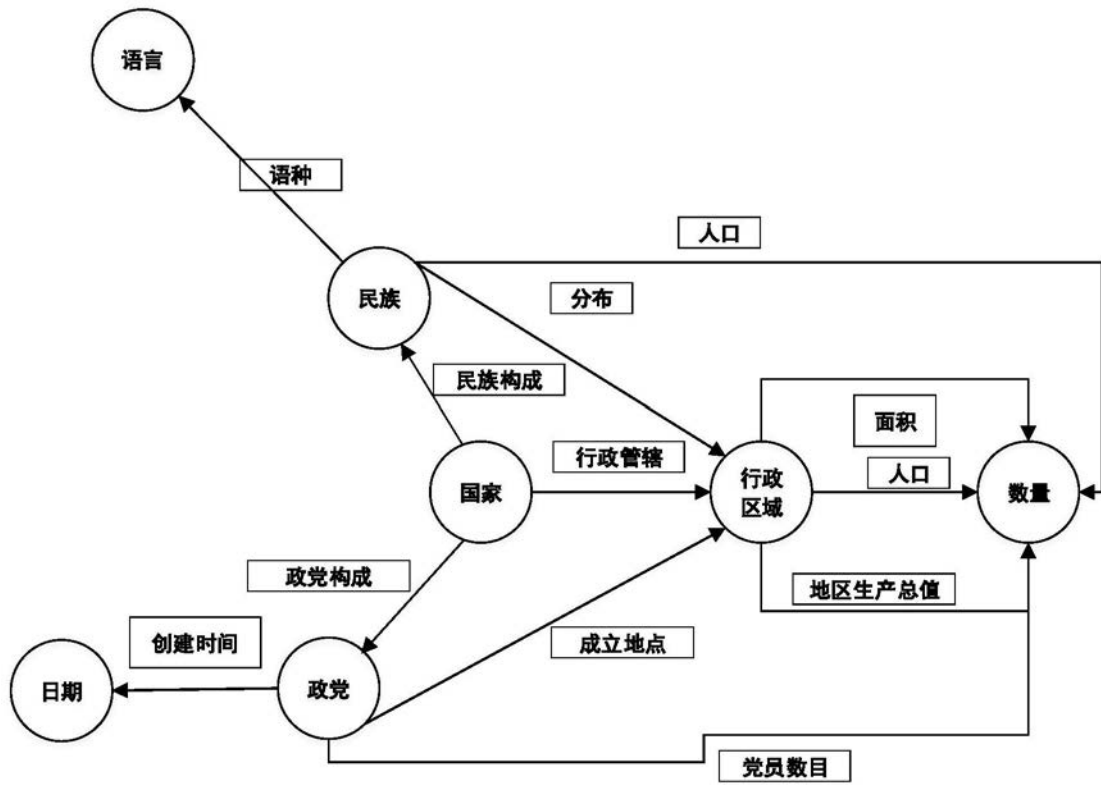


图4

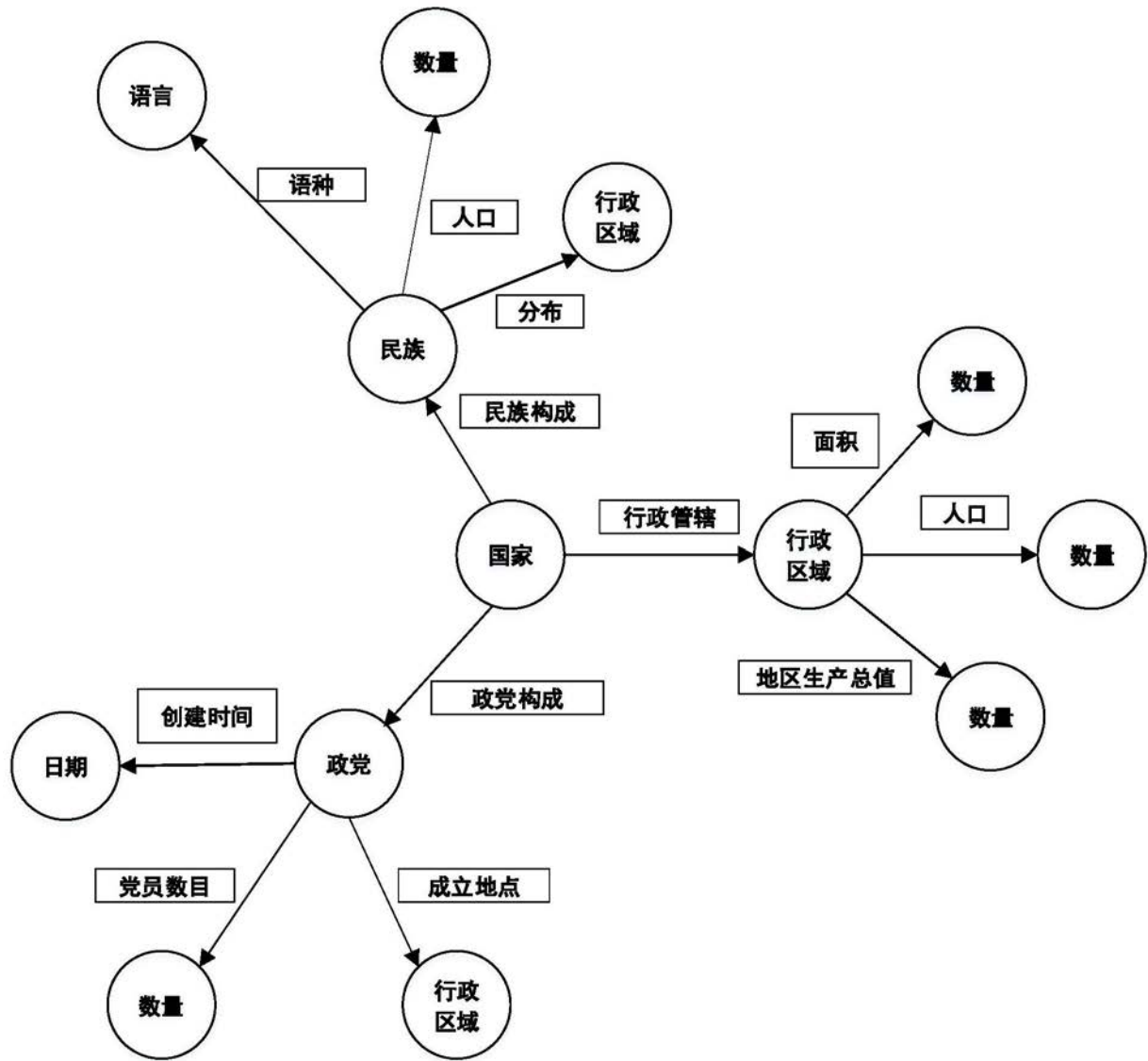


图5

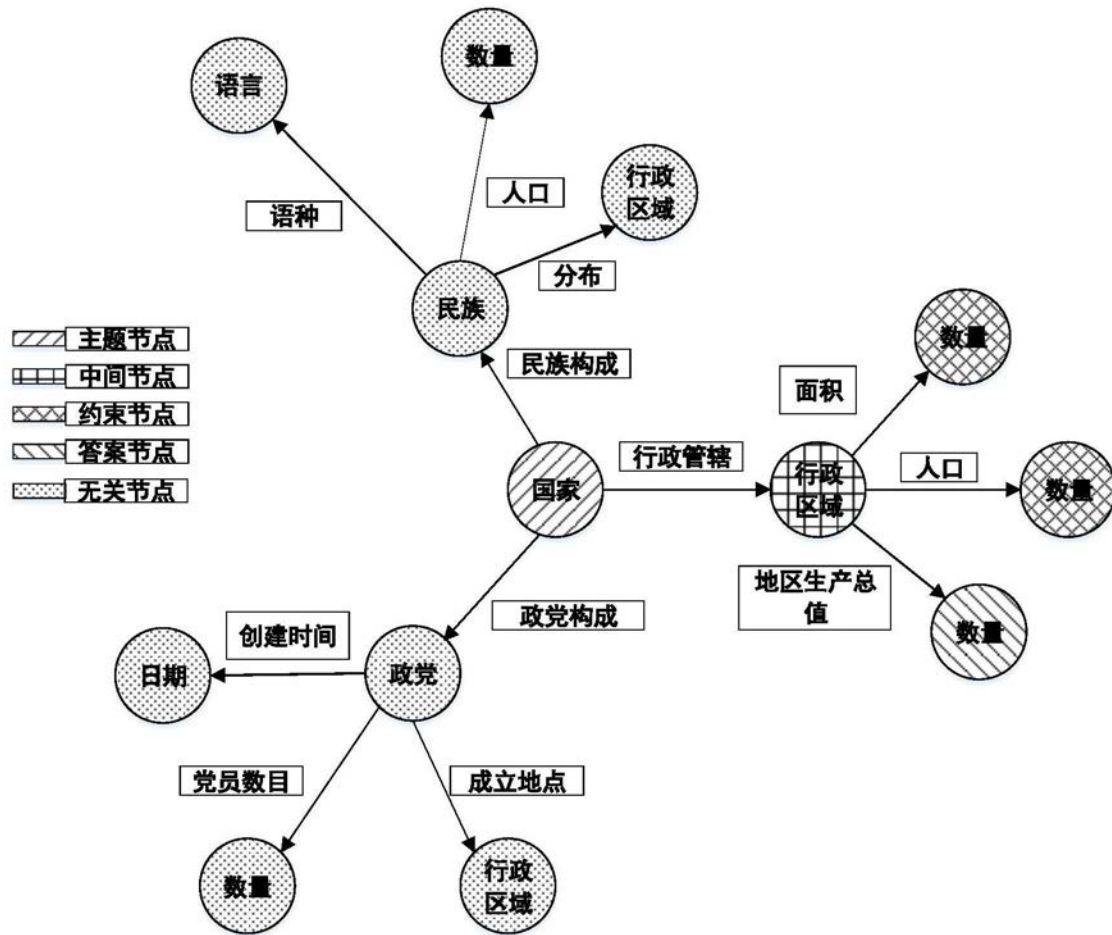


图6

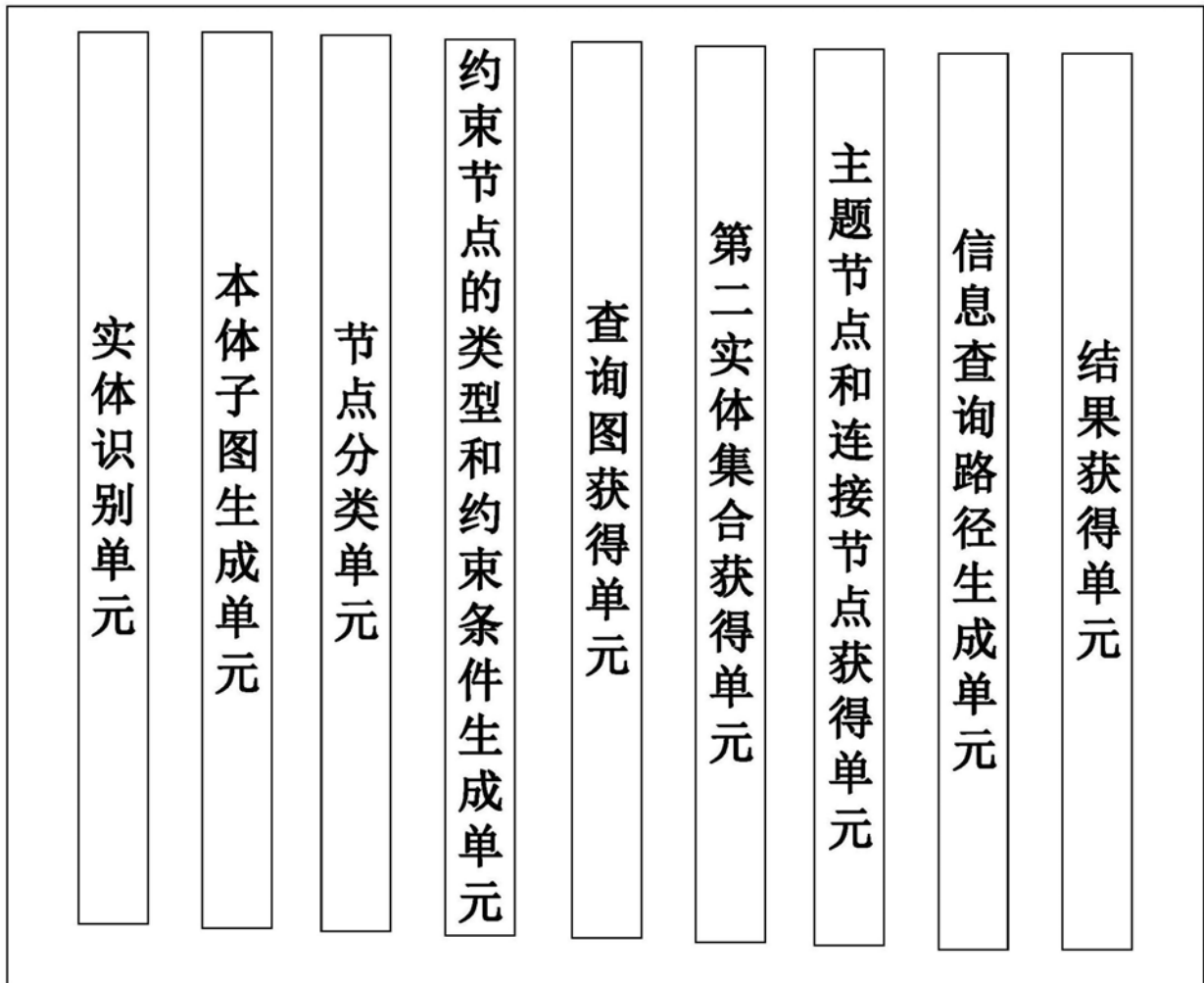


图7