(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 107977393 A
(43)申请公布日 2018.05.01

(21)申请号 201710363378.1

(22)申请日 2017.05.22

(71)申请人 海南大学
地址 570228 海南省海口市美兰区人民道路58号

(72)发明人 段玉聪 邵礼旭

(51)Int.Cl.
G06F 17/30(2006.01)
G06N 99/00(2010.01)

权利要求书1页 说明书5页 附图5页

(54)发明名称

一种面向5W问答的基于数据图谱、信息图谱、知识图谱和智慧图谱的推荐引擎设计方法

(57)摘要

本发明是一种面向5W问答的基于数据图谱，信息图谱，知识图谱和智慧图谱的推荐引擎设计方法，主要用于通过图谱回答用户使用自然语言提出的问题，属于分布式计算和软件工程学技术交叉领域。该方法允许人们通过提出自然语言形式的问题来表达他们的信息需求。知识图谱能以图形化的方式向用户展示经过分类整理的结构化知识，从而使用户从人工过滤网页寻找答案的模式中解脱出来。我们提出在数据、信息、知识和智慧层面上澄清知识图谱的整体表达，分别使用数据图谱、信息图谱、知识图谱和智慧图谱来解决用户提出的由5W(谁(Who)/何时(When)/何地(Where)，什么(What)，如何(How)，为什么(Why))引导的问题。

CN 107977393 A

1.一种面向5W问答的基于数据图谱、信息图谱、知识图谱和智慧图谱的推荐引擎设计方法，其步骤如下：

步骤1)构建问题模式库，将用户用自然语言提出的问题根据疑问词的不同划分为：a)由谁(who)或什么时候(when)或什么地点(where)引导的问题；b)由什么(what)引导的问题；c)由如何(how)引导的问题；d)由为什么(why)引导的问题；

步骤2)根据问题的类型，选择在哪种图谱上进行遍历查找答案；

步骤3)生成答案并将答案返回给用户。

2.基于数据图回答由谁(who)或什么时候(when)或什么地点(where)引导的问题：

利用对齐规则将用户提出的问题中的(疑问词，关系词，实体)映射到数据图谱中相关的(主体，关系词，客体)，形成精确的查询语句，利用该查询语句得到答案。

3.基于信息图谱回答由什么(what)引导的问题：

a)通过大量数据集训练，得出划分实体类型的规则；将用户问题中的实体属性与信息图谱中实体属性相匹配，计算相似度，将相似度最高的实体作为答案返回给用户；

b)在信息图谱上若无法直接找到答案，可以通过推理建立两个实体间的关系，增加图谱边密度，新建立的关系的正确度Cr是可计算的，P表示实体E1和实体E2之间的一条路径，Q表示所有路径，$\theta(\pi)$表示训练权重：

$$Cr(E_1,R,E_2)=\frac{\sum_{\pi\in Q}P(E_1\rightarrow E_2,\pi)\theta(\pi)}{|Q|};$$

c)在信息图谱上还能根据用户的问题进行相应的语义扩展，以推荐更多用户关心的信息，进一步提高查全率和查准率。

4.基于知识图谱回答由如何(how)引导的问题：

由如何(how)引导的问题，答案通常是以类似流程图的形式给出，因此我们在知识图谱上找到问题中的相关实体后，通过路径查询，将相邻的实体和关系词进行桥接。

5.基于智慧图谱回答由为什么(why)引导的问题：

a)本发明使用一种迭代的询问技术，来探索特定问题的因果关系，该技术的主要目标是通过重复“为什么”这个问题来确定缺陷或问题的根本原因，每个答案构成下一个问题的基础，通过设定询问次数来终止迭代询问；

b)对于两个实体间的因果关系，我们通过遍历两个实体间的所有路径找到所有可能的原因。

一种面向5W问答的基于数据图谱、信息图谱、知识图谱和智慧图谱的推荐引擎设计方法

技术领域

[0001] 本发明是一种面向5W问答的基于数据图谱、信息图谱、知识图谱和智慧图谱的推荐引擎设计方法。主要用于通过查询图谱回答用户使用自然语言提出的问题，属于分布式计算和软件工程学技术交叉领域。

[0002]

背景技术

[0003] 知识图谱于2012年5月17日被Google正式提出，其初衷是为了提高搜索引擎的能力，增强用户的搜索质量以及搜索体验。目前，随着智能信息服务应用的不断发展，知识图谱已被广泛应用于智能搜索、智能问答、个性化推荐等领域。尤其是在智能搜索中，用户的搜索请求不再局限于简单的关键词匹配，用户的信息需求仅仅通过关键字是不能被完整表达的。自然语言问题是制定信息需求最直观的方式，人们可以通过提出问题来表达他们的信息需求。问题可用于表达不能表达为关键字的复杂信息需求，并且不会在结构和语义上产生重大损失。知识图谱具有丰富的自然语义，可以包含各种更完整的信息，其表达机制更接近于自然语言，能以图形化的方式向用户展示经过分类整理的结构化知识，从而使用户从人工过滤网页寻找答案的模式中解脱出来。我们提出在数据、信息、知识和智慧层面上澄清知识图谱的整体表达，分别使用数据图谱、信息图谱、知识图谱和智慧图谱来解决5W问题。

[0004] 在本发明做出之前，已有的智能语义搜索应用中，当用户发起查询时，搜索引擎会借助知识图谱的帮助对用户查询的关键字进行解析和推理，进而将其映射到知识图谱中的一个或一组概念之上，然后根据知识图谱中的概念层次结构，向用户返回知识卡片，其中包括指向资源页面的超链接信息。在深度问答应用中，系统同样会首先在知识图谱的帮助下对用户使用自然语言提出的问题进行语义分析和语法分析，进而将其转化成结构化形式的查询语句，然后在知识图谱中查询答案。我们将用户提出的问题根据疑问词的不同进行分类，由“谁(who)、什么时候(when)、什么地点(where)”等疑问词引导的问题将在数据图谱上进行遍历查找答案，由“什么(what)”引导的问题在信息图谱上查找答案，由“如何(how)”引导的问题在知识图谱上查找答案，由“为什么(why)”引导的问题在智慧图谱上查找答案。

发明内容

[0005] 技术问题：本发明的目的是提供一种面向5W问答的基于数据图谱、信息图谱、知识图谱和智慧图谱的推荐引擎设计方法，用于解决当前用户信息需求变得复杂，仅仅通过关键词查询效率低下的问题，对由5W引导的问题我们限定是最基本的问题类型，不涉及5W问题之间的转换。本发明可显著地提高用户查询的查全率和查准率。

[0006] 技术方案：一种面向5W问答的基于数据图谱、信息图谱、知识图谱和智慧图谱的推荐引擎设计方法，其步骤如下所示。

[0007] 1.构建问题模式库。将用户用自然语言提出的问题根据疑问词的不同划分为:a)由谁(who)或什么时候(when)或什么地点(where)引导的问题;b)由什么(what)引导的问题;c)由如何(how)引导的问题;d)由为什么(why)引导的问题。

[0008] 2.对用户提出的问题进行分词,从而确定问题类型。

[0009] 3.根据问题的类型,选择在哪种图谱上进行遍历查找答案。

[0010] (1)基于数据图谱回答由谁(who)或什么时候(when)或什么地点(where)引导的问题。利用对齐规则将用户提出的问题中的(疑问词,关系词,实体)映射到数据图谱中相关的(主体,关系词,客体),形成精确的查询语句,利用该查询语句得到答案。

[0011] (2)基于信息图谱回答由什么(what)引导的问题:

a)通过大量数据集训练,得出划分实体类型的规则;将用户问题中的实体属性与信息图谱中实体属性相匹配,计算相似度,将相似度最高的实体作为答案返回给用户;

b)在信息图谱上若无法直接找到答案,可以通过信息推理建立两个实体间的关系,增加图谱边密度,新建立的关系的正确度Cr是可计算的,P表示实体1和实体2之间的一条路径,Q表示所有路径,$\theta(\pi)$表示训练权重:

$$Cr(E_1,R,E_2)=\frac{\sum_{\pi\in Q}P(E_1\rightarrow E_2,\pi)\theta(\pi)}{|Q|};$$

c)信息图谱还能根据用户的问题进行相应的语义扩展,以返回更多用户关心的信息,进一步提高查全率和查准率。

[0012] (3)基于知识图谱回答由如何(how)引导的问题:

由如何(how)引导的问题,答案通常是以类似流程图的形式给出,因此我们在知识图谱上找到问题中的相关实体后,通过路径查询,将相邻的实体和关系词进行桥接。

[0013] (4)基于智慧图谱回答由为什么(why)引导的问题:

a)本发明中使用迭代的询问技术,来探索特定问题的因果关系。该技术的主要目标是通过重复“为什么”这个问题来确定缺陷或问题的根本原因。每个答案构成下一个问题的基础,通过设定询问次数来终止迭代询问;

b)对于两个实体间的因果关系,我们通过遍历两个实体间的所有路径找到所有可能的原因。

[0014] 4.生成答案并将答案返回给用户。

[0015]

体系结构:

图1和图2分别给出了本发明的总体架构和流程示意图。首先由用户通过自然语言提出问题以表达自己的信息需求,本发明将用户提出的问题与问题模式库匹配,确定问题的类型,之后根据问题类型确定查询图谱的类型,通过遍历图谱最终将问题的答案返回给用户。

[0016] 下面给出数据图谱、信息图谱、知识图谱和智慧图谱的具体说明。

[0017] 数据图谱:数据是通过观察获得的数字或其他类型信息的基本个体项目,但是在没有上下文语境的情况下,它们本身没有意义。数据图谱可以通过数组、链表、队列、树、栈、图等数据结构来表达。

[0018] 信息图谱：信息是通过数据和数据组合之后的上下文传达的，经过概念映射和相关关系组合之后的适合分析和解释的信息。信息图谱可以通过关系数据库来表达。

[0019] 知识图谱：知识是从积累的信息中获得的总体理解和意识，将信息进行进一步的抽象和归类可以形成知识。知识图谱可以通过包含结点和结点之间关系的有向图来表达，知识图谱对需求语义的映射更完整，覆盖范围更宽。

[0020] 智慧图谱：智慧是一个外推过程，智慧使得人们可以明辨是非，从有限到无穷，从已知到未知进行推测。信息告诉人们做什么，知识告诉人们如何做，智慧告诉人们为什么要做。智慧图谱是在知识图谱的基础上体现出从已知到未知的推测过程，是一种混合型的难以剥离的结构。

[0021] 有益效果：本发明方法提出了一种面向5W问答的基于数据图谱、信息图谱、知识图谱和智慧图谱的推荐引擎设计方法，具有如下一些显著优点：

(1)将用户提出的问题有针对性地划分成5W问题，在数据、信息、知识和智慧层面上澄清知识图谱的整体表达，降低了查询的复杂性，提高搜索效率；

(2)具备语义推理功能，能根据用户的查询条件进行相应的语义扩展和语义推理，推荐更多用户关心的信息；

(3)支持自然语言的“问答式”查询，便于用户表达复杂中的信息需求。

[0022]

附图说明

[0023] 图1是本发明的总体架构示意图。

[0024] 图2是本发明的流程示意图。

[0025] 图3是数据图谱示例。

[0026] 图4是信息图谱示例。

[0027] 图5是知识图谱示例。

[0028] 图6和图7是智慧图谱示例。

[0029]

具体实施方式

[0030] 为了方便描述，我们通过例子来描述如何通过数据图谱回答由谁(who)或什么时候(when)或什么地点(where)引导的问题，通过信息图谱回答由什么(what)引导的问题，通过知识图谱回答由如何(how)引导的问题，通过智慧图谱回答由为什么(why)引导的问题。

[0031] 具体实施方案为：

(1)构建问题模式库。本发明通过对问题进行分词和词性标注处理后将问题归类为四种模式，分别是由谁(who)或什么时候(when)或什么地点(where)引导的问题，由什么(what)引导的问题，由如何(how)引导的问题和由为什么(why)引导的问题；

(2)根据问题模式选择在何种图谱上进行遍历。

[0032] a)基于数据图谱回答由谁(who)或什么时候(when)或什么地点(where)引导的问题。在图3中，我们假设用户提出的问题是“罗伯特的妻子是谁”，首先将该问题中的实体和

关系谓词提取出来，构造一个三元组(X,妻子,罗伯特)，将其转换为查询语句：

SELECT X WHERE (X,妻子,罗伯特) 然后遍历图谱，找到与实体罗伯特有“妻子”关系的另一端实体，作为答案返回给用户，即丽莎。

[0033] b)基于信息图谱回答由什么(what)引导的问题：

首先我们根据大量的数据集训练出对这些数据进行分类的规则，即找出每个类型的实体应满足哪些要求。我们假设对脊椎动物的分类有以下规则：

r1 :(是,飞翔的动物)∧(有,羽毛)∧(是,恒温动物)→鸟;

r2 :(是,水生动物)∧(有,鳞片)∧(呼吸,鳃)→鱼;

r3 :(是,变温动物)∧(有,鳞片)∧(呼吸,肺)→爬行动物;

r4:(是,胎生动物)∧(是,恒温动物)→哺乳动物;

r5 :(是,变温动物)∧(是,半水生动物)∧(呼吸,肺)→两栖动物。

[0034] 根据以上规则构建出脊椎动物分类的图谱如图4所示，当用户输入问题“燕子属于哪类脊椎动物”时，我们将燕子所拥有的属性与图谱中实体的属性相匹配，匹配度最高的实体类型将作为答案返回给用户。答案的正确率P可通过以下公式计算：

$$P=\frac{|\text{实体 1 的属性集合}|\cap|\text{实体 2 的属性集合}|}{|\text{实体 1 的属性集合}|};$$

在信息图谱中，可以通过信息推理建立更多实体之间的新关联，从而扩展实体之间的关系，增加信息图谱的边缘密度。推理需要有规则的支持，这些规则可以由人手动构建，但往往耗时费力。目前，它主要依靠关系的重现，利用协同挖掘技术自动找到推理规则。使用关系规则实现关系提取的经典方法是路径排序算法，它使用每个不同的关系路径作为一维特征。通过在信息图谱中构建大量关系路径来构建关系分类的特征向量和关系分类器来提取关系。新建立的关系的正确度Cr是可计算的，P表示实体E1和实体E2之间的一条路径，Q表示所有路径，$\theta(\pi)$表示训练权重：

$$Cr(E_1,R,E_2)=\frac{\sum_{\pi\in Q}P(E_1\rightarrow E_2,\pi)\theta(\pi)}{|Q|}。$$

[0035] c)基于知识图谱回答由如何(how)引导的问题：

由如何(how)引导的问题答案是一系列的流程，本发明使用路径查询来遍历图谱查找答案。路径查询由一个初始的实体s和要遍历的一系列关系，p =(r1,...,rk)组成。查询的答案或表示[q]是通过遍历p可以从s到达的所有实体的集合。在图5中，假设用户输入的问题是“如何展开一次招聘”，首先找到实体招聘，找到与它相关联的所有实体，要遍历的关系集合则是p=(下一步,下一步,…,下一步)。

[0036] d)基于智慧图谱回答由为什么(why)引导的问题。

[0037] 解决用户提问的由为什么(why)引导的问题分为两种情况：第一种是事务发生的原因来自自身，第二种是两个实体之间的因果关系。本发明使用迭代的询问技术，该技术的主要目标是通过重复“为什么”这个问题来确定缺陷或问题的根本原因。每个答案构成下一个问题的基础。在图6中，用户提出的问题是：“车子为什么无法启动”，根本原因来自与车辆本身没有按照推荐的服务计划进行维护。中间的原因是通过不断询问为什么得出的，通过设定有关询问次数的阈值，来终止迭代询问。

[0038] 对于两个实体间的因果关系，我们通过遍历两个实体间的所有路径找到所有可能的原因。在图7中，用户提出的问题是“吸烟是如何对肺造成损害的”，在图谱上分别找到烟和肺两个实体，将两个实体间的所有完整路径作为原因返回给用户。

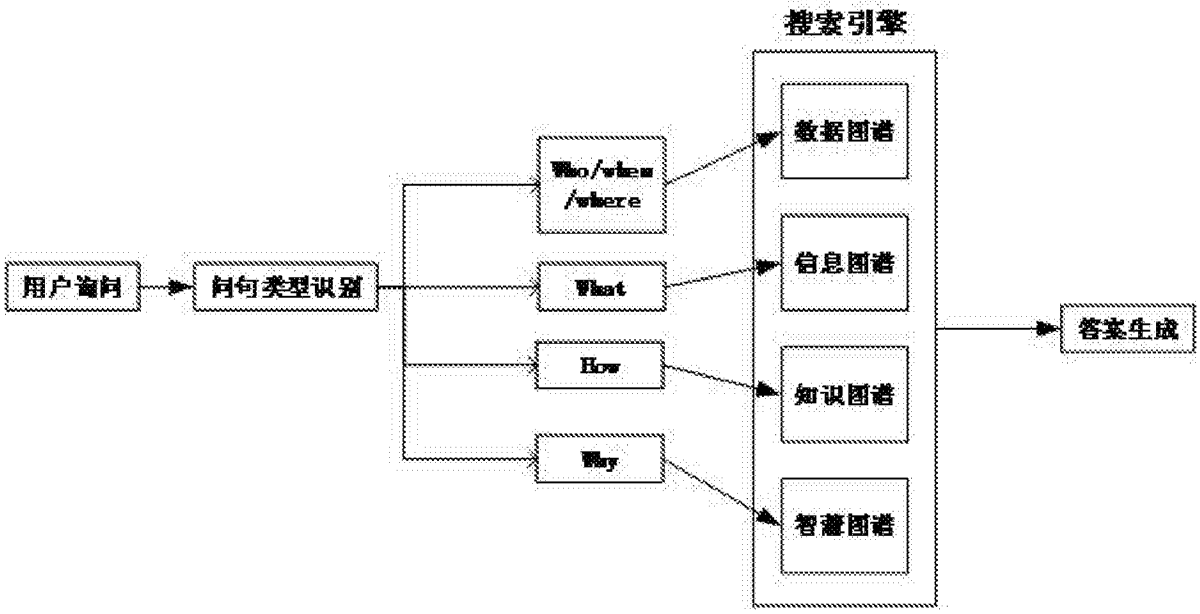
搜索引擎
用户询问
问句类型识别
Who/when
/where
What
How
Why
数据图谱
信息图谱
知识图谱
智慧图谱
答案生成

图1

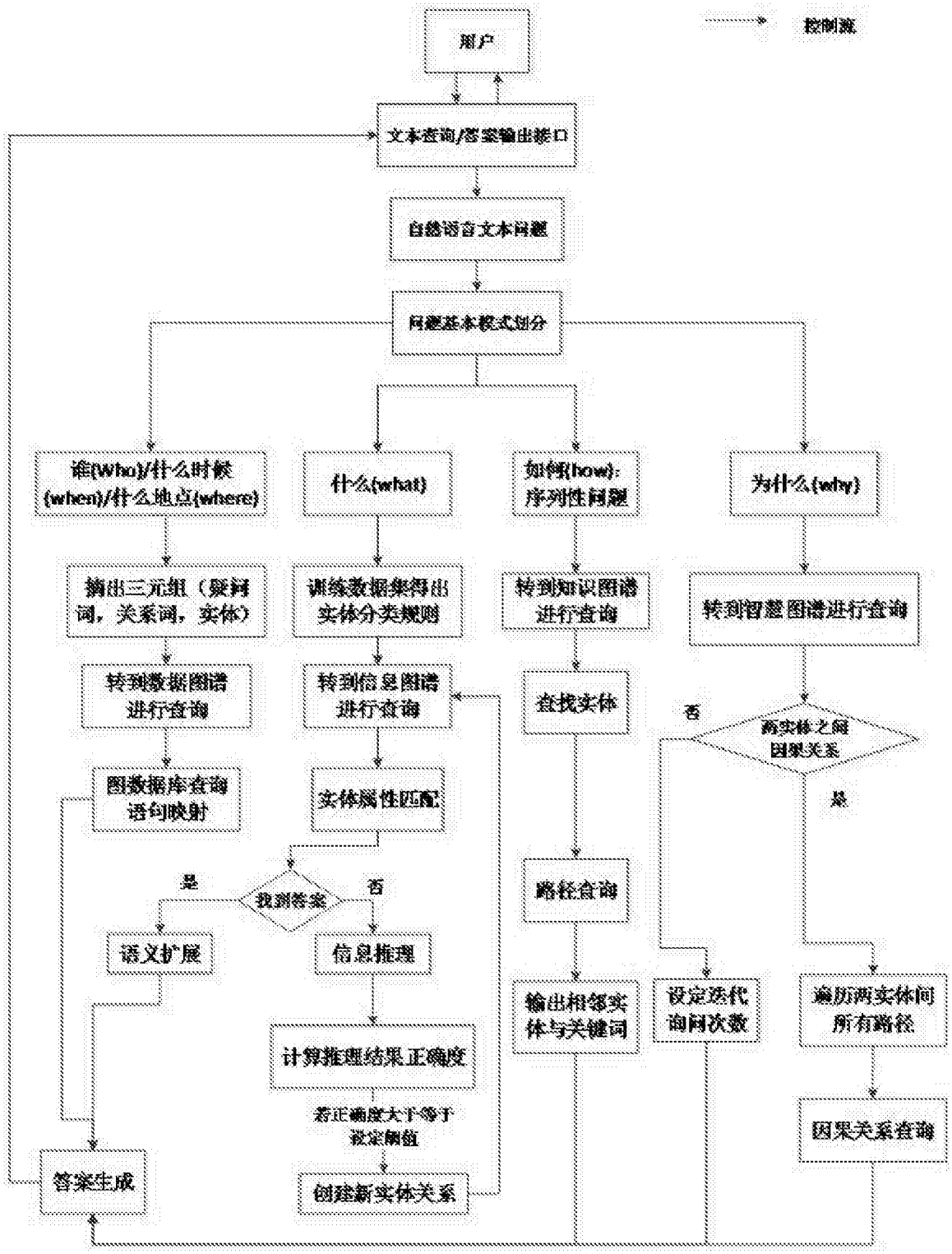
控制流
用户
文本查询/答案输出接口
自然语言文本问题
问题基本模式划分
谁(who)/什么时候(when)/什么地点(where)
什么(what)
如何(how)：序列性问题
为什么(why)
抽出三元组（疑问词，关系词，实体）
训练数据集得出实体分类规则
转到知识图谱进行查询
转到智慧图谱进行查询
转到数据图谱进行查询
转到信息图谱进行查询
查找实体
否
两实体之间因果关系
是
图数据库查询语句映射
实体属性匹配
是
找到答案
否
路径查询
语义扩展
信息推理
输出相邻实体与关键词
设定迭代询问次数
遍历两实体间所有路径
计算推理结果正确度
若正确度大于等于设定阈值
因果关系查询
答案生成
创建新实体关系

图2

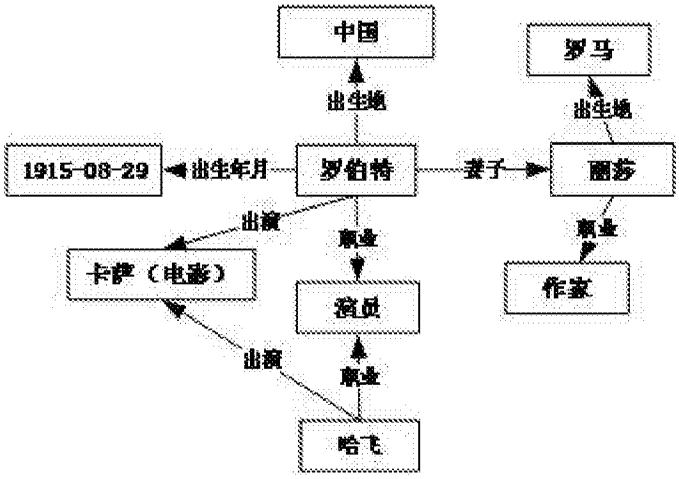
中国
罗马
出生地
出生地
1915-08-29
出生年月
罗伯特
妻子
丽莎
出演
职业
职业
卡萨（电影）
演员
作家
出演
职业
哈飞

图3

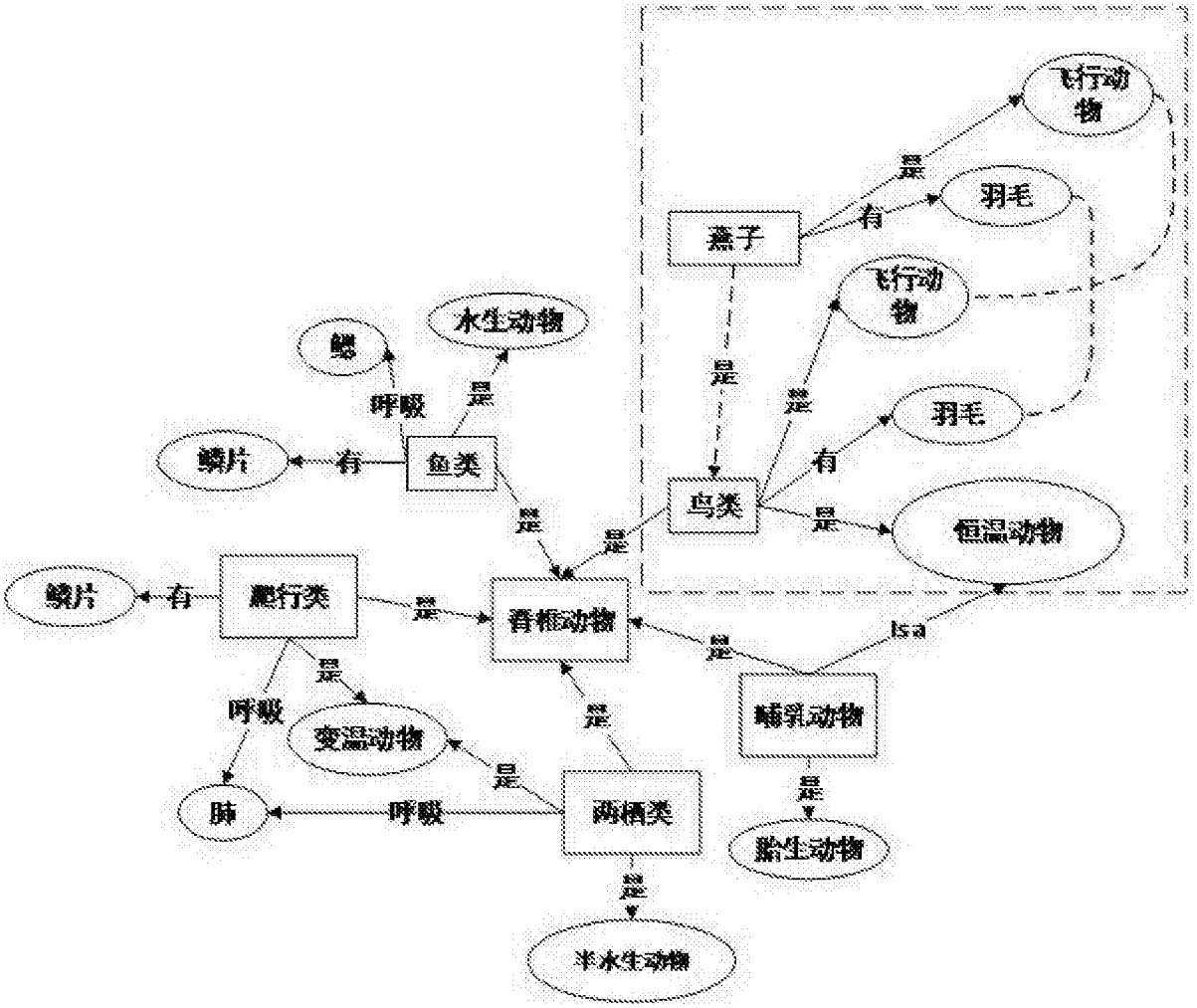
飞行动物
是
羽毛
有
燕子
飞行动物
是
是
羽毛
有
鸟类
是
恒温动物
水生动物
鳃
是
呼吸
鳞片
有
鱼类
是
是
鳞片
有
爬行类
是
脊椎动物
is a
是
呼吸
是
哺乳动物
变温动物
是
是
肺
呼吸
两栖类
是
胎生动物
是
半水生动物

图4

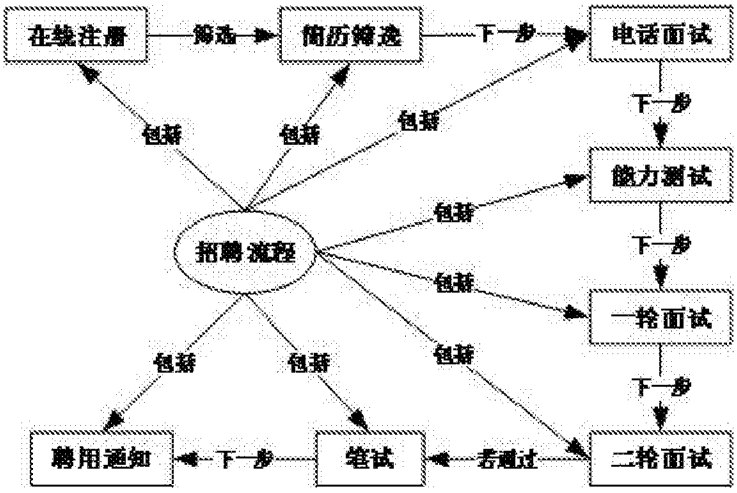
在线注册
筛选
简历筛选
下一步
电话面试
包括
包括
包括
下一步
能力测试
包括
下一步
招聘流程
包括
一轮面试
包括
包括
包括
下一步
聘用通知
下一步
笔试
未通过
二轮面试

图5

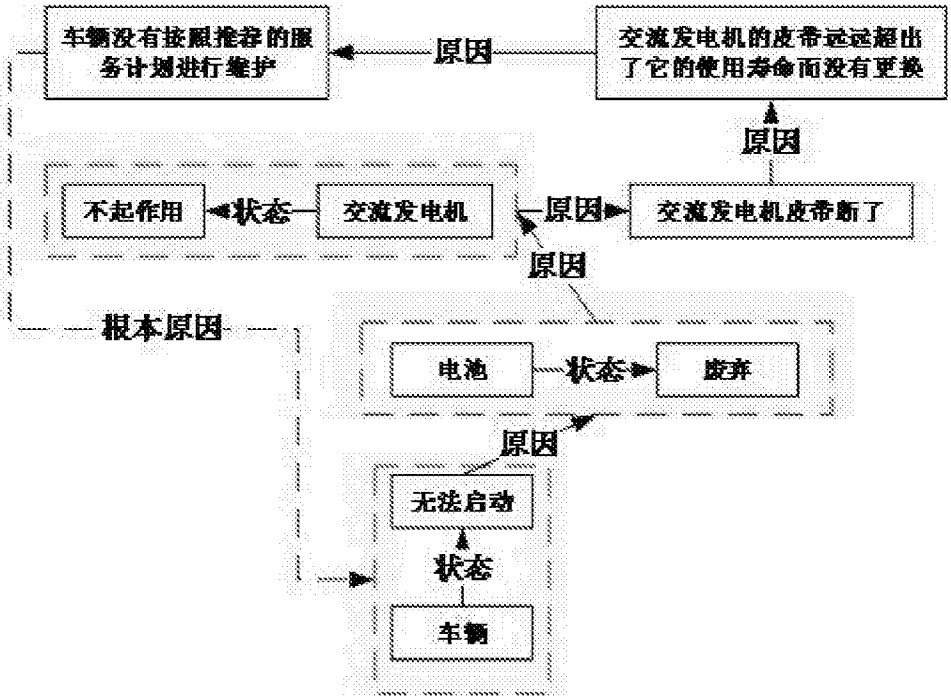
车辆没有按照推荐的服务计划进行维护
原因
交流发电机的皮带运转超出了它的使用寿命而没有更换
原因
不起作用
状态
交流发电机
原因
交流发电机皮带断了
原因
根本原因
电池
状态
废弃
原因
无法启动
状态
车辆

图6

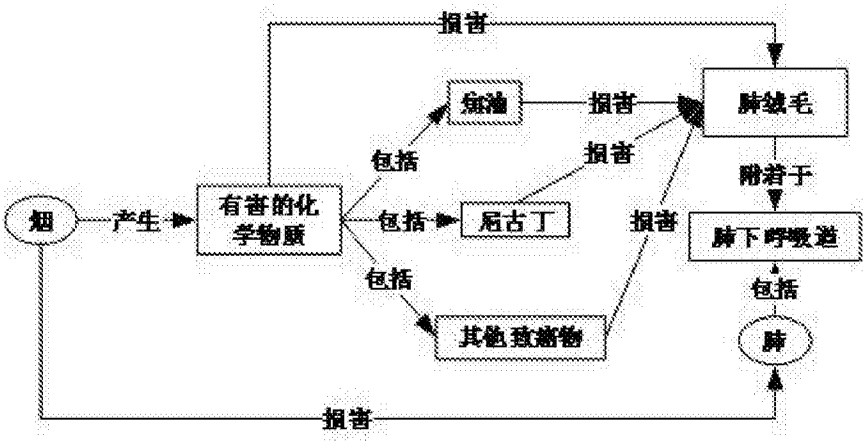
损害
鱼油
损害
肺泡毛
包括
损害
附着于
烟
产生
有害的化学物质
包括
尼古丁
损害
肺下呼吸道
包括
包括
其他致癌物
肺
损害

图7