



**气象知识图谱构建方案**

**目录**

[1. 基本方案 2](#_Toc482091395)

[2. 方案详细介绍 3](#_Toc482091396)

[2.1 原始数据源说明 3](#_Toc482091397)

[2.2.1. 内部数据 3](#_Toc482091398)

[2.2.2. 外部数据 3](#_Toc482091399)

[2.2 互联网数据采集与清洗说明 4](#_Toc482091400)

[2.3.1. 采集数据源说明 4](#_Toc482091401)

[2.3.2. 采集配置方案说明 4](#_Toc482091402)

[2.3 知识抽取说明 5](#_Toc482091403)

[2.4.1. D2R映射方案说明 5](#_Toc482091404)

[2.4.2. 行业类网站抽取说明 6](#_Toc482091405)

[2.4 知识融合与图谱构建说明 7](#_Toc482091406)

[2.5 气象知识图谱存储和应用 7](#_Toc482091407)

[2.5.1. 气象知识图谱存储 7](#_Toc482091408)

[2.5.2. 气象知识图谱应用 8](#_Toc482091409)

**气象知识图谱构建及其应用基本方案设计**

# 基本方案

气象知识图谱的构建及其应用方案的基本框架如下图1所示，依次可分为以下几层：

* 原始数据层：原始数据主要包括两方面的数据：一方面企业内部关系型数据库中的数据；另一方面是气象行业网站、百科、论文资源等互联网外部数据，包括气象科普、百度百科、百度学术等等；
* 互联网信息采集与清洗层：主要针对原始数据层中的开放互联网数据编写通用爬虫和特定的行业网站爬虫，将相应领域知识和数据采集存储到本地；**（提供通用采集器，需要开发特定行业爬虫）**
* 知识抽取层：知识抽取层包括D2R映射和行业网站抽取。D2R映射主要提供原始关系型数据库到气象图谱schema的文件映射（如表对应概念，记录对应实体，记录数据对应属性值等）和原始数据的多种更新方式；行业网站抽取主要利用wrapper组件结构化相关气象知识，同时提供周期性的数据更新；**（双方合作）**
* 知识融合层：知识融合层包括schema融合和数据融合。对抽取获得的气象数据进行实体与概念的识别、实体合并、实体对齐、上下位关系学习、关联关系学习，把记录型气象数据转化为知识图谱形式的气象知识；同时提供自动实体的冲突检测，并根据预定的冲突解决策略对融合过程中产生的冲突进行解决；**（双方合作）**
* 图谱存储层：基于MongoDB的大规模三元组知识存储为所构建的气象知识图谱提供百亿级别以上三元组知识的存储与更新，保证知识图谱的流式处理流程中的效率；一方面支持底层数据经过知识图谱学习过程不断对知识图谱进行补充和更新，另一方面为上层高速的数据访问提供支撑。同时提供基于Elastic Search构建气象图谱的分布式文件索引。**（提供图谱存储）**
* 图谱应用层：图谱应用层为上层应用提供数据访问的基础气象数据读取接口，并提供关联分析（包括实体关系发现和路径发现等）、基于图谱的标签传播语义标注和知识导航等；**（提供基础的分析功能、可视化功能，特定的业务需求需合作开发）**



图1 气象知识图谱基本框架图

# 方案详细介绍

## 原始数据源说明

### 内部数据

内部数据主要来源于企业内部关系型数据库中的关系型数据。**（暂时未知）**

### 外部数据

外部数据主要包括专家、院校、学术期刊等，信息主要来源于以下几类：

* 学术期刊类： 如知网、万方、百度学术等；
* 气象行业网站：中国气象科普网、中国天气、气象知识等；
* 开放知识库：百度百科、维基百科等。

## 互联网数据采集与清洗说明

### 采集数据源说明

数据采集主要包括采集气象类型、气象领域专家、气象领域科研机构等，采集来源包括垂直类网站、百科类网站和学术期刊类网站等，具体说明如下表1所示。

表1 互联网数据采集网站说明（部分）

|  |  |
| --- | --- |
| **数据来源** | **采集地址** |
| 气象科普园地 | http://www.cma.gov.cn/kppd/ |
| 气象知识 | http://www.qxzs.org.cn/ |
| 中国气象科普网 | http://www.qxkp.net/ |

### 采集配置方案说明

针对不同的数据类型提供通用采集配置界面，配置界面如下图2所示（以豆瓣为例），采集数据结果如图3所示。

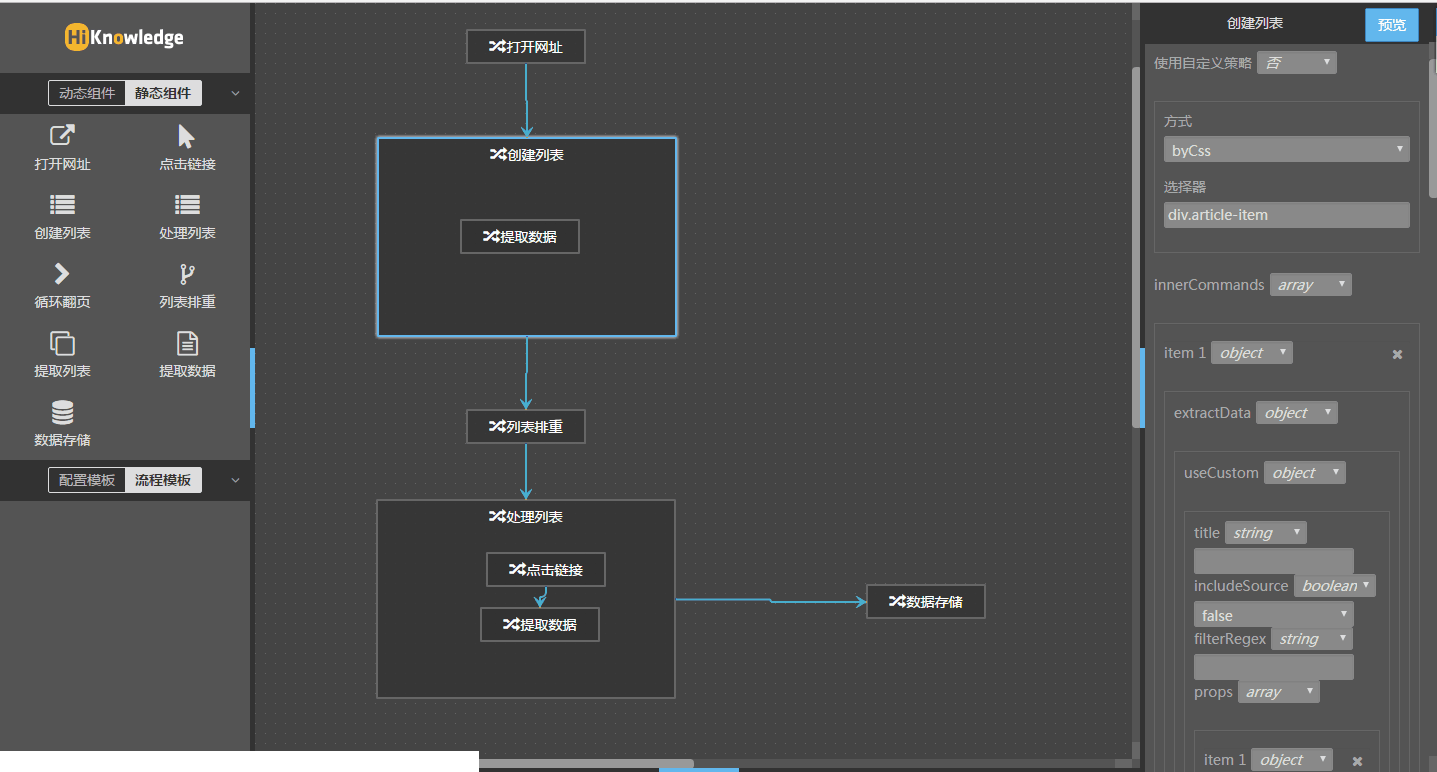


图2 可视化采集配置界面

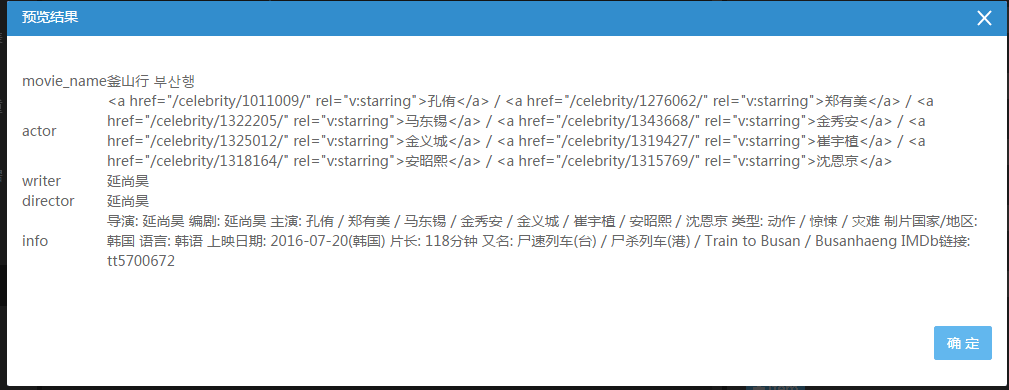


图3 可视化采集采集结果界面

## 知识抽取说明

知识抽取层包括D2R映射和行业网站抽取。D2R映射主要提供原始关系型数据库到气象图谱schema的文件映射（如表对应概念，记录对应实体，记录数据对应属性值等）和原始数据的多种更新方式；行业网站抽取主要利用wrapper组件结构化相关气象知识，同时提供周期性的数据更新。

### D2R映射方案说明

D2R即把关系型数据库中的数据转化为RDF三元组形式的语义数据。本次方案中制定了一组从关系型数据库映射到语义数据的映射规范，即D2RML，该规范使用XML语言描述。

D2RML中的主要关键词及相应的描述功能如下：

* dbtype：源数据库的类型，有mysql、oracle、sqlserver等，类型决定了连接时使用的驱动；
* dburl：数据库连接字符串，指定数据库的地址、端口和使用的数据库等信息；
* dbuser：数据库的用户名；
* dbpwd：数据库的密码；
* table：源数据表；
* concept：导入目标概念；
* name的colname属性：实体名称来源列；
* synonym的colname属性：同义实体来源列；
* parent的tablename属性：父概念的表名；
* attribute的colname指定属性来源列，attrname则指定属性名。

一个典型的映射文件如下所示，它描述了从气象数据库中映射气象类知识图谱的配置。

*<?xml version="1.0" encoding="utf-8"?>*

*<config>*

*<dbconfig>*

*<dbtype>mysql</dbtype>*

*<dburl>jdbc:mysql://8.8.8.8:8888/weather</dburl>*

*<dbuser>root</dbuser>*

*<dbpwd>root</dbpwd>*

*</dbconfig>*

*<mappings>*

*<mapping>*

*<table>weather</table> <!--源数据表-->*

*<concept>天气</concept> <!—导入目标概念 -->*

*<name colname="name"/> <!—实体名称来源列 -->*

*<synonym colname="en\_name"/> <!—同义实体来源列 -->*

*<attributes> <!—实体属性原来列-->*

*<attribute colname="definition" attrname="定义"/>*

*</attributes>*

*</mapping>*

*</mappings>*

*</config>*

在从结构化数据中进行知识映射，首先需要充分理解结构化数据中的基本结构，包括每个表格的含义以及表之间的关联，同时理解知识图谱的结构，然后使用D2RML把结构化数据中的表格与知识图谱中的概念或实体关联起来。知识抽取服务连接映射文件中的目标数据库，读取相应表格中的数据，把关系型数据库中的表和列数据分别映射成概念的实体以及实体的属性，然后把这些映射得到的知识存储到气象知识图谱中。

### 行业类网站抽取说明

针对不同的数据源，编写特定的wrapper包装器抽取如下图4所示气象相关的数据。



图4 待抽取数据示例

## 知识融合与图谱构建说明

经过知识抽取后的数据都是符合图谱的预先定义的三元组格式。利用生成好的三元组数据构建气象知识图谱的基本步骤如下：

* 构建气象知识分类体系：基于已有气象知识分类体系，利用外部气象分类（百科气象分类等）进行对齐和扩充；
* 定义专家、科研机构等实体的基本属性，包括专家的学术领域、就职单位、学术论文等；
* 添加实体节点；包括专家节点、科研机构节点等。在添加时需要确定图谱中是否已存在该节点，如已存在，则选取已有节点，如不存在，则新增节点。同时，对于一些重复名称的实体，还需要注意进行实体的合并。
* 实体属性值填充；把实体的属性填充完整。

在构建过程中，主要难点在于气象分类体系的扩充、同名专家节点的合并以及专家与技术点的关联。

## 气象知识图谱存储和应用

### 气象知识图谱存储

海翼知大规模图存储软件由四部分组成，依次是底层存储层、倒排索引层、图挖掘层、应用接口层。

* 底层存储层：最终存储数据的介质，通过使用不同类型的数据库来存储不同类型的数据；
* 缓存与索引：为数据中需要频繁访问的数据（如图的Schema）提供缓存功能，通常使用内存数据库实现；
* 图挖掘：对图形数据的挖掘算法，实现图形数据的高效使用；典型的算法包括图遍历、最短路径、图匹配和网络流等。
* 应用接口层：以RESTful、JavaAPI等方式提供对基础数据、图形数据、图检索和图挖掘的算法访问。

海翼知提供了百亿级别三元组存储、读取、查询的完整的解决方案，在现有的存储解决方案能够有效的满足气象知识图谱的存储、访问要求。

### 气象知识图谱应用

图谱的应用包括：

* 路径分析：两个实体之间的路径发现，发现实体间二度、三度关系；

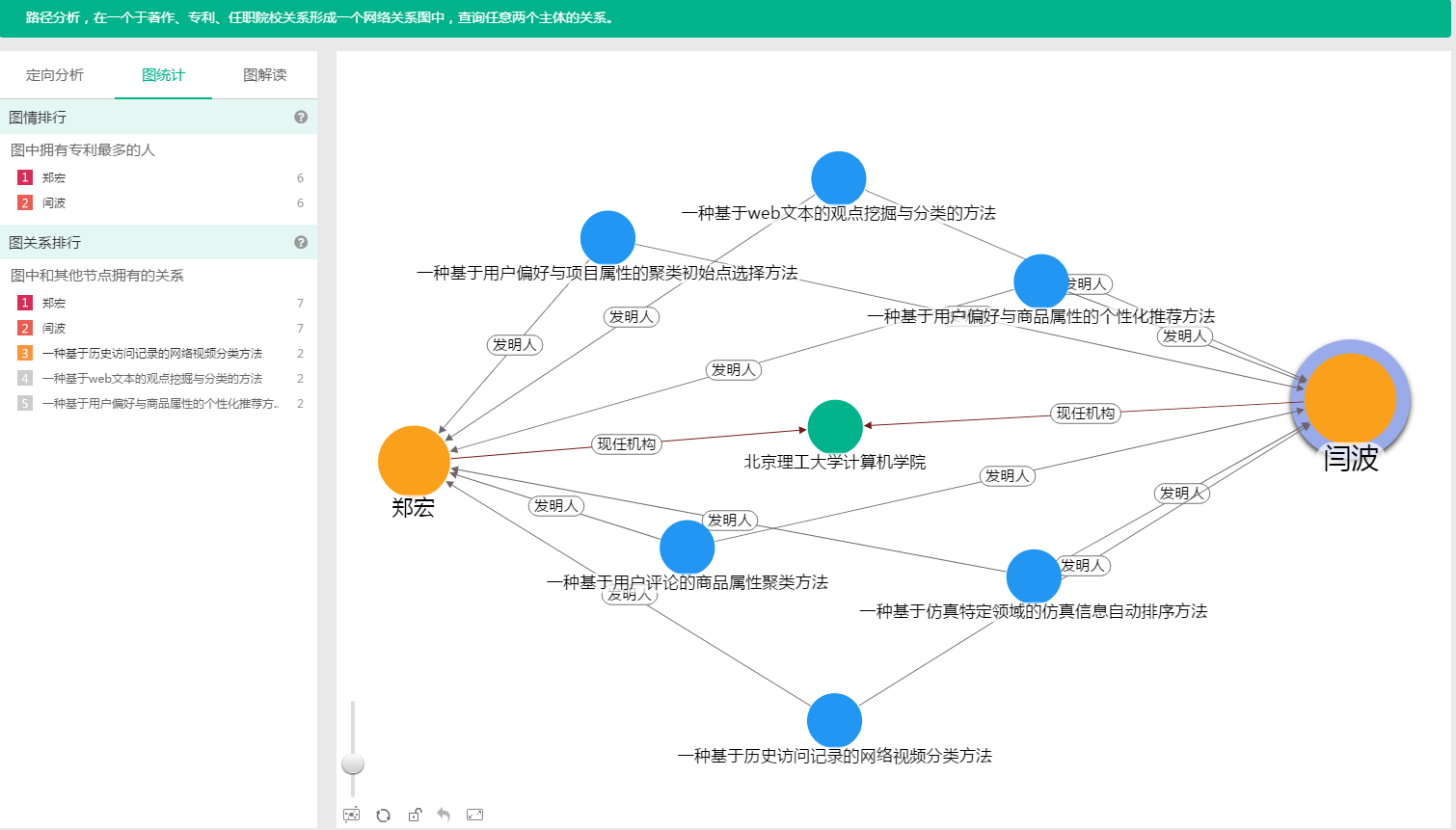


图1 路径分析示例

* 关联分析：发现多个实体与实体之间的关联关系；

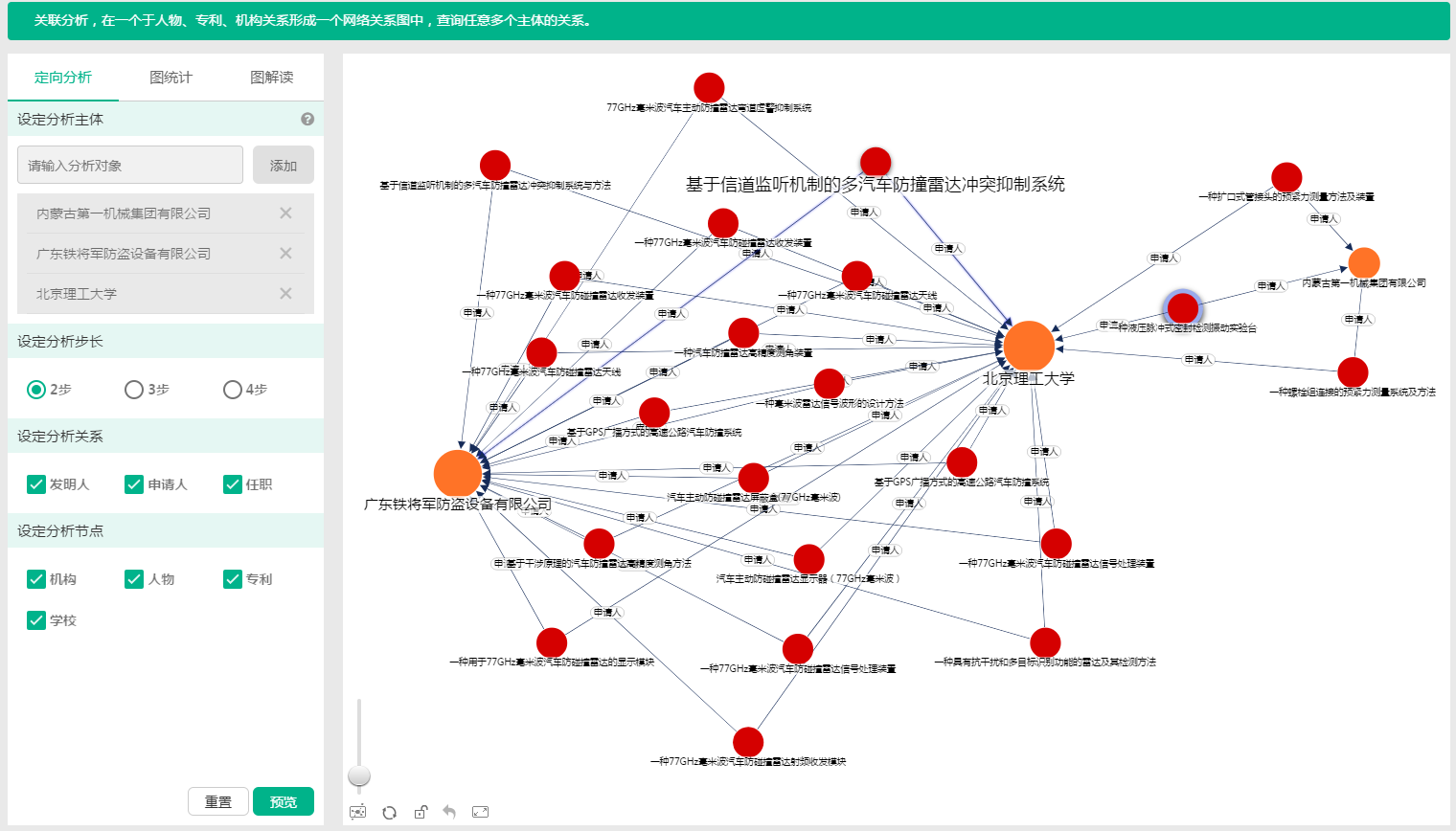


图2 关联分析示例

* 模板检索：基于语义的检索模板，如\*\*气象领域的专家有哪些等等；

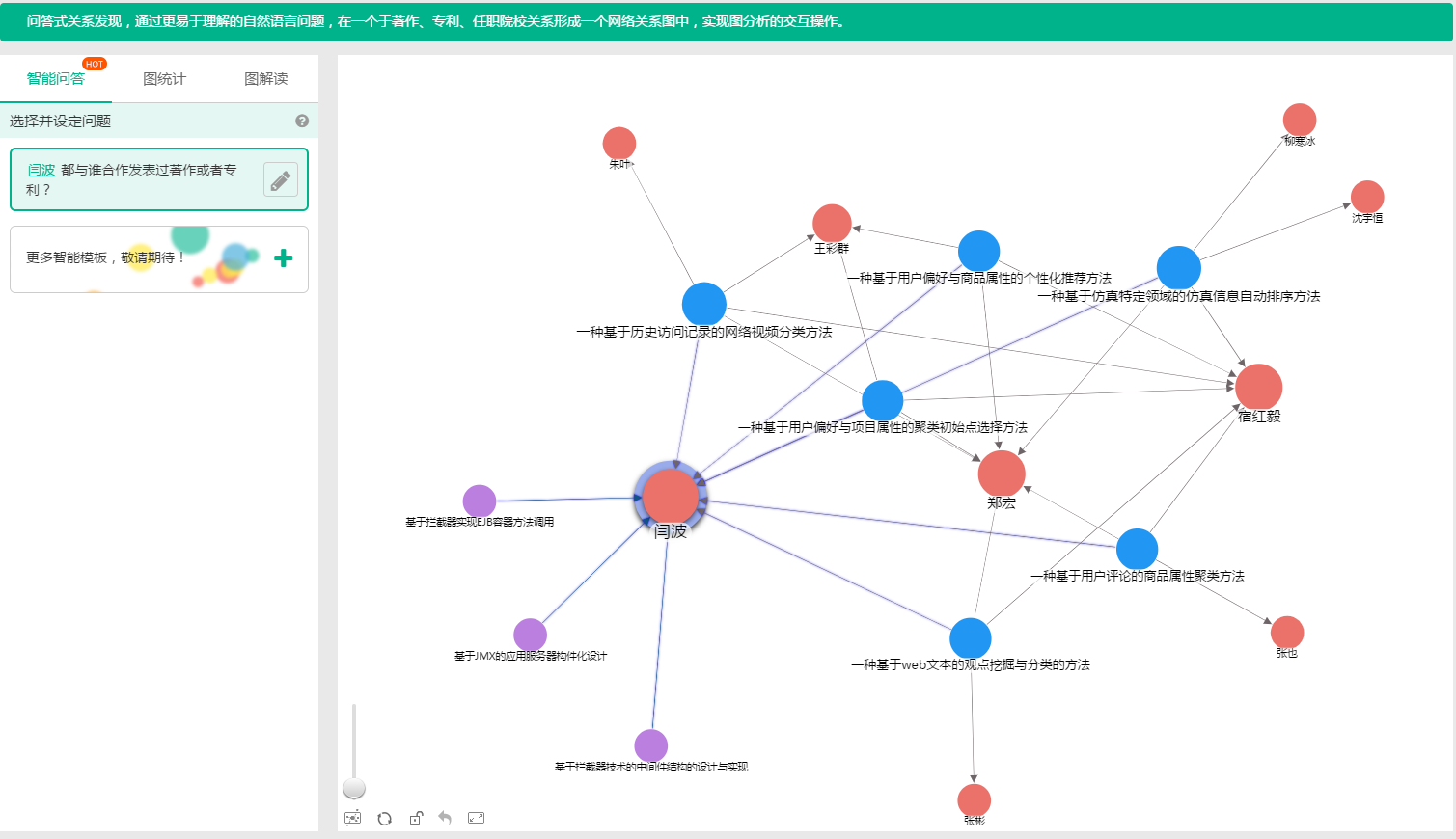


图3 智能问答示例

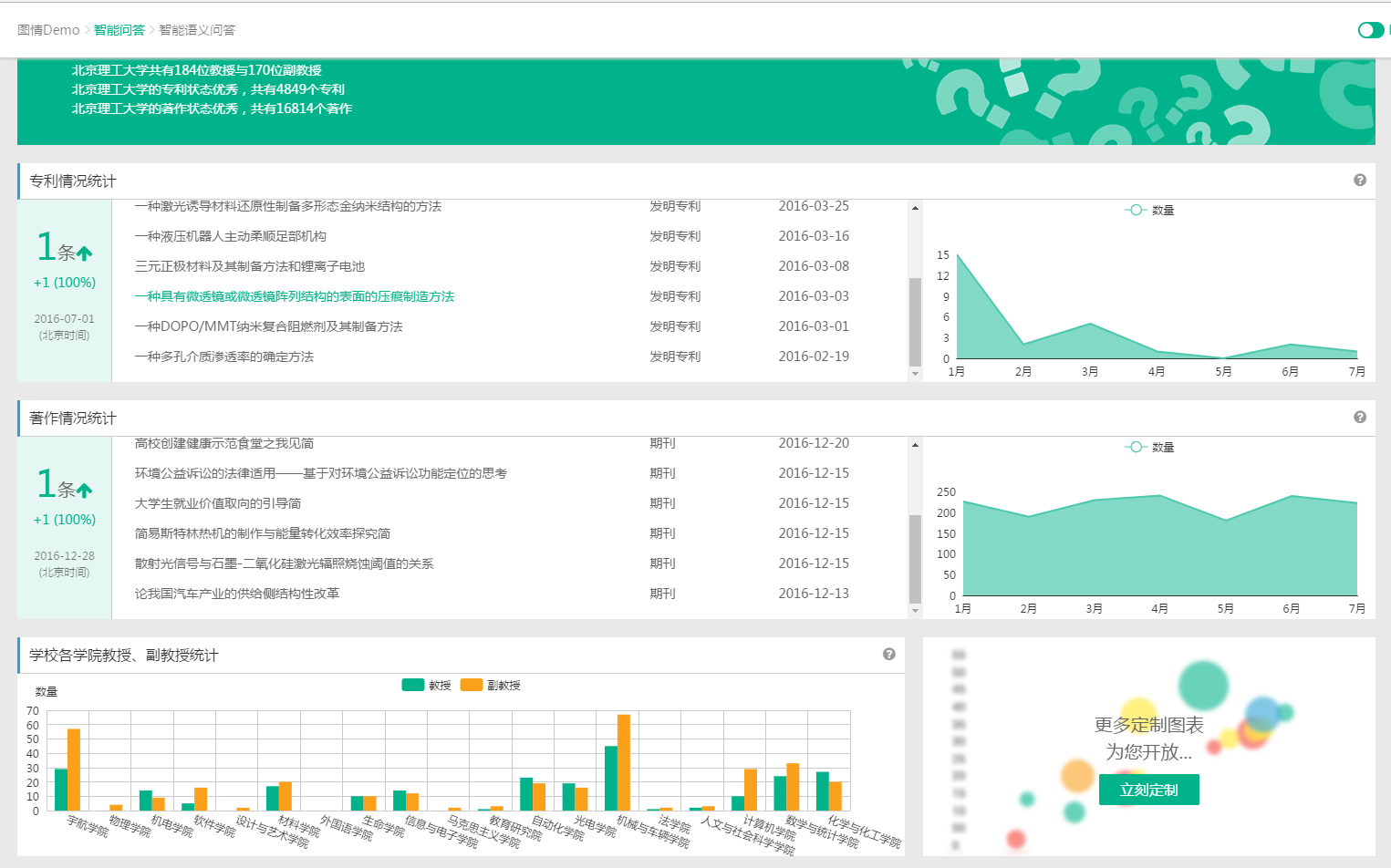


图4语义问答示例

* 资源统计分析：针对知识点下的资源进行可视化的统计分析。



图5 资源统计分析示例

（**备注：**路径分析和关联分析目前只支持实体类型过滤和边类型的过滤计算，暂未支持边上权重的计算。）