网络开源多模态科技情报智能分析系统概要设计

撰写人：杨卉帆

2022-04

# 概述

## 研究内容

* 研究复杂网络条件下互联网开源数据稳定获取技术；
* 研究多模态开源数据治理与融合链接技术、人物属性关系推理补全技术、虚拟身份关联技术、跨数据源的情报知识迭代校验方法；
* 研究碎片化信息关联整合、多源异质科技情报信息网络构建技术，实现完整、高质量的科技情报信息网络构建和组织；
* 研究多语言环境下数据内容识别与信息认知技术；
* 研究情报智能推理与情报自动化加工技术。

## 考核指标

* 支持境内外社交媒体、新闻网络、信息站点、网络论坛等不少200种开源数据源的稳定获取；
* 支持文本、图像、视频、音频等数据的融合处理；
* 支持亿级实体规模的异质科技情报信息网络构建；
* 支持中、英、西、俄、藏等主要语种科技情报的自然语言处理、实体链接、属性抽取、关系抽取，准确率大于80％；
* 研制一套网络开源多模态大数据情报智能分析平台，在科技情报相关领域示范应用。

## 开发背景

Acalligence——网络开源多模态科技情报智能分析系统，面向学术界和工业界以及关注科技的广大学术爱好者。

（待描述功能）

## 名词解释

1. 工作：包括论文、专利、书籍、数据集等，会引用其他工作；
2. 作者：产出工作的人；
3. 出处：刊载工作的会议、期刊或资源库；
4. 机构：（通过作者）与工作相关的大学或其他组织；
5. 概念：即研究领域，以话题方式对工作进行标签；
6. 事件：社会上发生的具有一定影响力的事，一般应该包括时间、地点、人物、起因、经过、结果等属性，能够从若干篇新闻报道中提取出；

## 系统架构

系统由数据采集层、多模态知识获取层、知识库存储层、应用服务层四大层级组成。系统整体架构如图1-1所示。

图 1‑1系统架构图

### 数据采集层

数据采集层负责对原始多模态数据的获取。数据来源包括学术搜索系统、新闻站点、社交媒体、网络论坛四个部分。数据涉及不同的语言，包括中、英、西、俄、藏等主要语种。数据的不同模态有文本、图片、音频和视频。

### 多模态知识获取层

多模态知识获取层通过输入多模态数据，对不同模态的数据进行信息抽取，并进行跨模态数据融合，最终输出多模态知识库。

### 知识库存储层

知识库存储层负责数据的读取和保存，包括图数据库、关系型数据库、分布式文件系统及其配套的存取接口。

### 应用服务层（系统功能模块）

#### 信息抽取模块

##### 作者信息

该功能为展示作者的个人信息。本系统通过解析个人网站、机构官网、社交媒体、网络论坛、新闻站点上的作者信息，进行作者个人信息属性的挖掘，从而得到精准、多维的作者个人信息。作者个人信息包括以下属性：

* 姓名
* 照片
* 性别
* 所在机构
* 所在中心
* 职称
* 职务
* 个人视频
* 研究领域（包括所关联的概念）
* 教育经历
* 所承担的项目
* 所获奖项
* **按期刊、会议、专著统计成果**
* **按年份进行个人成果数**
* **一些学术方面的指数**
  + 1. Citation（引用量）：该作者所有出版物的被引用次数。
  + 2. Papers（论文数量）：该作者发表的所有论文的数量。
  + 3. H-index（H-指数）：如果一位专家的n篇论文中，每篇至少有 h 次引用，并且其他 (N-h) 篇论文每篇最多有 h 次引用，则该专家的索引为 h。
  + 4. Activity（活跃度）：作者的活跃度是根据过去几年发表的论文的数量界定的。
  + 5. Diversity（多样性）：一般来说，专家的研究可能包括几个不同的研究领域。多样性的定义是为了定量地反映作者的研究范围。
  + 6. Sociability（社交性）：社交性的分数基本上是根据一位专家有多少合著者来定义的。在下一步，我们将进一步考虑地点、组织、国籍信息和研究领域。
  + 7. G-index（G-指数）：G-指数是H-指数的衍生指数，主要是弥补h指数不能很好反应高被引论文的缺陷提出的，G-指数定义为: 论文按被引次数排序后相对排前的累积被引至少g2次的最大论文序次g，亦即第(g+1)序次论文对应的累积引文数将小于(g+1)2。
* ……

##### 机构信息

该功能为展示机构的信息。本系统通过解析机构官网、社交媒体、网络论坛、新闻站点上的机构信息，进行机构属性的挖掘，从而得到精准、多维的机构信息。机构信息包括以下属性：

* 名称
* 所属国家或地区
* 类型
* 官方网站
* 标志图片
* 别名
* 工作成果（按照年份统计数量等）
* 被引用量
* 地理位置
* 相关联的机构组织（上级机构等）
* 研究领域（包括所关联的概念）
* 创立时间
* 创立人
* 当前负责人
* 视频简介
* ……

##### 出处信息

该功能为展示出处（刊载工作的会议、期刊或资源库）的信息。本系统通过解析官网、社交媒体、网络论坛、新闻站点上的出处信息，进行出处属性的挖掘，从而得到精准、多维的出处信息。出处信息包括以下属性：

* 名称
* 出版商
* 工作成果（按照年份统计数量等）
* 被引用量
* 主页网址
* 研究领域
* 创立时间
* 更新时间
* ……

##### 作者/机构/出处多账号关联模块

随着社交网络快速发展，不断涌现的大规模社交网络（如Facebook，LinkedIn，新浪微博等）吸引了数以亿计用户。不同的社交网络在其功能，用户体验，目标用户群等各个方面都有不同的特点，例如Facebook是真实社交网络的线上版本，其内部的好友关系大多正是用户在线下的真实好友关系，且其好友关系是双向的；在Twitter或新浪微博上用户则更趋和自己的偶像或是意见领袖建立关系，这种关系是单向的；LinkedIn是职业化的社交网络，以便于用户更新自己展示自己的工作经历和能力水平，Google Scholar和AMiner等学术合作网络则反映了作者在发表学术论文时的合作关系。

正因为每种社交网络在用户的工作和生活中都各自扮演着不同的角色，用户常常在不同的社交网络上都拥有账户。每个账户都是用户完整形象的一个局部缩影，很显然，由于分散在各个不同的社交网络，这些局部是不相互连通的。因此，本系统通过机器学习手段，自动将作者、机构和出处的多个社交网络的账户进行自动关联。

##### 工作信息模块

该功能为展示工作（论文、专利、书籍、数据集等）信息。本系统通过解析学术搜索系统、作者的个人网站、机构官网、社交媒体、网络论坛、新闻站点上的工作信息，进行工作信息属性的挖掘，从而得到精准、多维的工作信息。工作信息包括以下属性：

* 标题
* 出版时间
* 出处
* 领域（包括所关联的概念）
* 公开性
* 作者列表
* 机构列表
* 被引用量
* 引用格式
* 所引用的工作
* 相关工作
* 创建时间
* 更新时间
* ……

##### 文献智能解析模块

智能PDF文献解析——参考文献解析、图表抽取、目录解析

##### 概念信息模块

该功能为展示概念信息。本系统通过解析学术搜索系统、作者的个人网站、机构官网、社交媒体、网络论坛、新闻站点上的概念信息，进行概念信息属性的挖掘，从而得到精准、多维的概念信息。概念信息包括以下属性：

* 名称
* 别名
* 所属概念层次
* 上级概念
* 相关概念
* 描述
* 相关工作成果
* 被引用量
* 图片标志
* 创建时间
* 更新时间
* ……

#### 智能推荐模块

学术推荐：权威审稿人推荐、优秀论文推荐、“伯乐”推荐等

领域专家发现、热点话题发现以及论文引用模式挖掘等

##### 相似作者

该模块功能是：在用户浏览一位作者的个人信息页面时，本系统将智能发现/推荐与该作者有类似属性（如研究领域、学术指数等）的作者。

##### 相似机构

该模块功能是：在用户浏览一个机构的信息页面时，本系统将智能发现/推荐与该机构有类似属性（如研究领域、相关联的机构组织等）的机构。

##### 相似出处

##### 相似工作

##### 相似概念

#### 关系网络模块

##### 作者关系网络

本系统通过对作者发表论文的协作者进行统计整理，将所有协作者分为合作关系、师生关系两类，并以图谱的形式进行展示。

1. 合作关系

以一个作者为中心，每一条线连接与该作者有过合作的专家学者，线条的长度体现合作紧密度，这些合作是指论文、专利等研究成果的合作。

1. 师生关系（学位论文数据集 -> 学术谱系）

传统的科技情报分析系统很少对作者的学术谱系进行描述。本系统以学位论文为数据集，使用实体识别与关系抽取等方法从中挖掘实体及其链接关系，使用关系推断等方法从网络中识别出潜在的师生关系，进一步归纳整理得到最终的学术谱系。如图8所示，这是北京邮电大学陈俊亮院士的多层学术谱系，从中可以看出，陈俊亮指导的学生王柏作为导师指导了另一批学生。

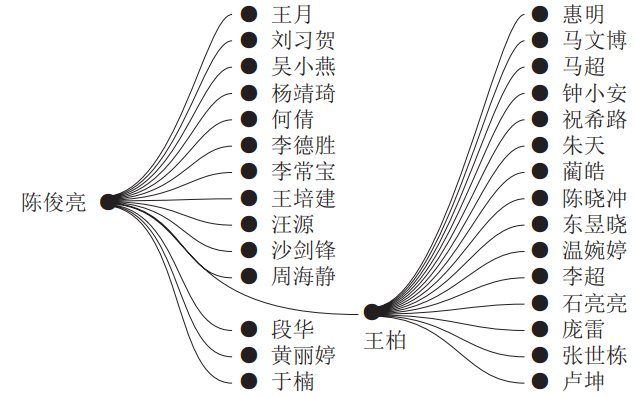


图8 学术谱系

本系统在作者关系网络方面的优势在于：以直观的关系图形式展示，关系的维度和关系图的数目较多。

##### 机构关系网络

本系统通过对机构发表论文的合作机构进行统计整理，以图谱的形式展示机构之间的合作关系。

以一个机构为中心，每一条线连接与该机构有过合作的其他机构，线条的长度体现合作紧密度，这些合作是指论文、专利等研究成果的合作。

##### 工作+概念关系网络

本系统通过对工作所属的研究领域/概念进行统计分析，以图谱的形式展示工作和概念之间的所属关系。

以一个概念为中心，每一条线连接该研究领域下的一个工作，线条的长度体现所属关系的相关程度，这些合作是指论文、专利等研究成果的合作。

##### 机构社团划分

传统的科技情报分析系统不包含机构社团划分功能，我们通过使用Fast Unfolding算法进行机构社团发现，将刻画两名作者关系的“一对一层次”拓展到刻画多名作者关系的“多对多层次”。这样，可以更深入、更直观地挖掘机构名下的科研团队信息。

本系统在机构社团划分功能方面的优势：有社团发现功能，且区分出的社团和真实分组情况较为接近。

#### 趋势预测模块

##### 作者研究趋势

作者的研究领域及学术关键词是作者画像中重要的组成部分，能够充分体现出作者的研究方向、学术偏好，甚至可以体现出作者对该学科领域热点的关注度。

传统的科技学术搜索系统，如谷歌学术、百度学术等，不涉及学术关键词功能；少数科技情报分析系统，如AMiner、中国知网等，只是对学术关键词进行简单的罗列。这种描述方式并不能描绘作者在不同时间段关注的学术关键词及作者的研究路线变迁趋势。

因此，我们从时间维度出发，对作者的学术关键词进行刻画，将作者研究关键词的变化趋势直观地体现出来，从而发现作者的学术研究脉络。学术关键词变化趋势如图所示。

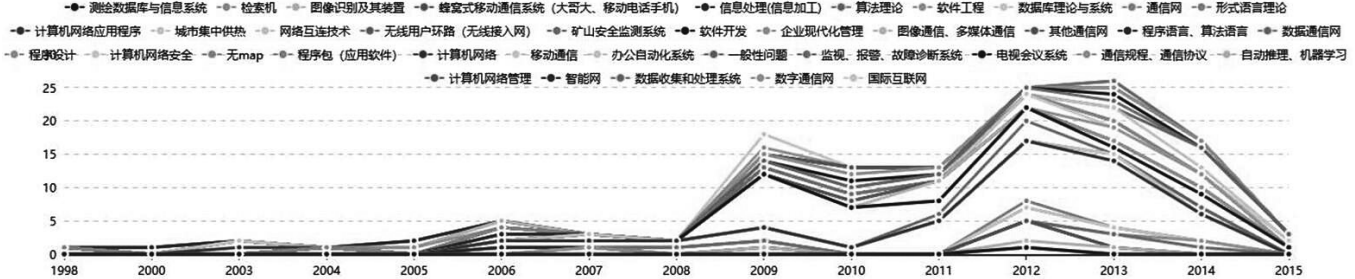


图7 学术关键词变化趋势

##### 机构研究趋势

机构研究趋势是某所机构研究重心、研究热点的直观体现。机构的研究趋势也会间接影响作者未来的研究方向。

传统的科技情报分析系统，如谷歌学术、百度学术、中国知网、AMiner等都不包含对机构学术研究关键词研究趋势的挖掘。

同样从时间维度出发，我们由机构名下作者的学术关键词趋势上卷得到该机构的学术 热词的演化趋势。

##### 出处研究趋势

出处研究趋势是某个出处研究重心、研究热点的直观体现。出处的研究趋势也会间接影响作者未来的研究方向。

传统的科技情报分析系统，如谷歌学术、百度学术、中国知网、AMiner等都不包含对出处学术研究关键词研究趋势的挖掘。

同样从时间维度出发，我们由出处名下作者的学术关键词趋势上卷得到该出处的学术 热词的演化趋势。

##### 概念研究趋势

概念研究趋势是某个研究领域的重心、热点的直观体现。概念的研究趋势也会间接影响该研究领域下作者未来的研究方向。

传统的科技情报分析系统，如谷歌学术、百度学术、中国知网、AMiner等都不包含对概念学术研究关键词研究趋势的挖掘。

同样从时间维度出发，我们由概念对应的研究领域下作者的学术关键词趋势上卷得到该概念的学术热词的演化趋势。

##### 团队核心人物演化分析

传统的科技情报分析系统缺少团队核心人物演化分析功能。团队核心人物的挖掘预测是对科研团队群体行为的分析，本系统能够预测学术机构内某位作者离职后，哪位作者会接替他的位置。

我们通过解决TNRP网络重塑问题求出网络中需要删除的作者集合，接着解决MTSP多作者网络继任者问题来预测出可能构建的新网络及相关的概率。当科研合作网络中移除一名团队核心人物时，该模块可以预测出网络的演化结果以及核心人物的更替。

##### 作者迁徙路线

#### 智能检索模块

研究者搜索、论文搜索、综述文献搜索、关联关系搜索以及基于话题的子图搜索

##### 搜索作者

1. 领域关键词/人名搜索

在搜索框内输入感兴趣的领域关键词或作者人名，通过h指数、性别、地区、语言等筛选条件和学术活跃度、领域新星、引用数以及论文数等排序方式查找。

1. 高级检索（附加“条件”）

在搜索框下的条件栏内输入您想要查找的作者所在机构，缩小检索范围。

##### 搜索机构

1. 领域关键词/机构名搜索

在搜索框内输入感兴趣的领域关键词或机构名称，通过地区、语言等筛选条件和学术活跃度、引用数以及论文数等排序方式查找。

1. 高级检索（附加“条件”）

在搜索框下的条件栏内输入您想要查找的机构所包括的作者，缩小检索范围。

##### 搜索出处

* 领域关键词/出处名搜索

在搜索框内输入感兴趣的领域关键词或出处名称，通过地区、语言等筛选条件和引用数、论文数等排序方式查找。

##### 搜索工作

1. 工作搜索+筛选

在搜索框内输入工作名称、关键词或相关领域，并限制作者、期刊、关键词等条件，查找您感兴趣的工作。

1. 通过查找该领域的学者找对应的工作

通过查找该领域的顶尖学者，可以进入学者主页查看该学者的工作。

1. 根据某篇工作查找相似工作

想要查找某篇工作的相似工作，可以直接在该工作的详情页面查看【相似工作】。

##### 搜索概念

1. 概念搜索+筛选

在搜索框内输入概念名称、关键词或相关领域，并限制作者、期刊、关键词等条件，查找您感兴趣的概念。

1. 通过查找该领域的学者找对应的概念

通过查找该领域的顶尖学者，可以进入学者主页查看该学者的研究领域所涉及的概念。

1. 根据某个概念查找相似概念

想要查找某个概念的相似概念，可以直接在该概念的详情页面查看【相似概念】。

##### 六度搜索

六度搜索的含义是指现实生活中的任意两个人之间建立起关联最多只需要通过六个人。在“六度分离” 理论中，社会中普遍存在人与人之间的弱纽带关系，这种关系能够拉近互不相识的两个人之间的距离，这在社会关系中发挥着巨大的作用。“六度分隔”产生的关系路径可以利用熟人之间的联系产生一个可信任的网络，这其中的潜能的确是无可估量的。

传统的科技情报分析系统不包含六度搜索功能。我们根据用户输入的两个人物实体对象，发掘实体间的关联路径及其路径中的人物。首先根据需要查询的人物关系从关系网络中进行实体搜索，再通过图算法获取极大连通子图作为网络关系的查询结果。本系统的六度搜索功能产生一个可信任的网络路径，通过这个网络路径，能够更清晰明确地观察到两名作者在关系网络中的信任路径，挖掘他们潜在的合作可能。六度搜索网络如图9所示。

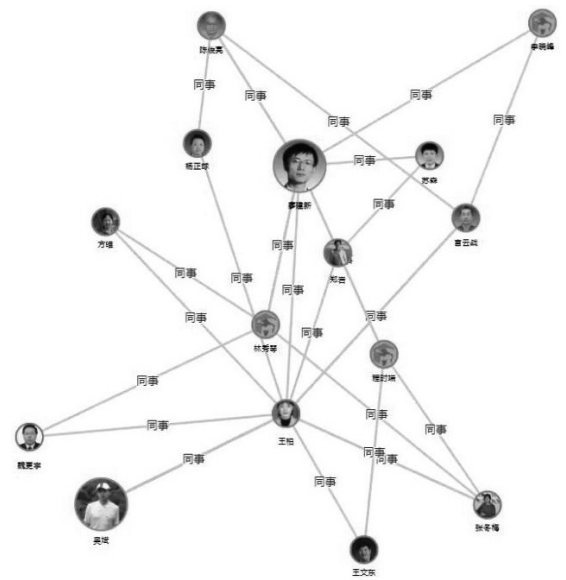


图9 六度搜索网络

#### 智能问答模块

##### 视频问答

##### 知识图谱问答

#### 评估排名模块

##### 作者评估

本系统评估作者的目的是确定来自互联网以及其他来源的大数据的研究人员的科学影响。我们的主要目标是使用各种指标根据文献计量数据对作者进行排名。

1. 数据源

我们使用近十年领域顶尖会议上发表的所有论文，对于每篇论文，我们提取作者信息，以及每篇论文的引文信息。

1. 指标

* Citation（引用量）：该作者所有出版物的被引用次数。
* Papers（论文数量）：该作者发表的所有论文的数量。
* H-index（H-指数）：如果一位专家的n篇论文中，每篇至少有 h 次引用，并且其他 (N-h) 篇论文每篇最多有 h 次引用，则该专家的索引为 h。
* Activity（活跃度）：作者的活跃度是根据过去几年发表的论文的数量界定的。
* Diversity（多样性）：一般来说，专家的研究可能包括几个不同的研究领域。多样性的定义是为了定量地反映作者的研究范围。
* Sociability（社交性）：社交性的分数基本上是根据一位专家有多少合著者来定义的。在下一步，我们将进一步考虑地点、组织、国籍信息和研究领域。
* G-index（G-指数）：G-指数是H-指数的衍生指数，主要是弥补h指数不能很好反应高被引论文的缺陷提出的，G-指数定义为: 论文按被引次数排序后相对排前的累积被引至少g2次的最大论文序次g，亦即第(g+1)序次论文对应的累积引文数将小于(g+1)2。

##### 机构评估

本系统评估机构的目的是确定来自互联网以及其他来源的大数据的组织的科学影响。我们的主要目标是使用各种指标根据文献计量数据对学术组织进行排名。

1. 数据源

我们使用近十年领域顶尖会议上发表的所有论文，对于每篇论文，我们从作者的隶属关系信息中提取组织，以及每篇论文的引文信息。

1. 指标

我们根据五个指标报告排名：论文数量、第一作者论文数量、科学信用、引文数量、AM-score。我们区分不同类型的论文，例如普通论文、短篇论文和长篇论文。基本上，在这个排名中，我们只考虑在领域顶尖会议上发表的长篇论文。

##### 出处评估

本系统主要评估会议和期刊两大类出处。用户可以根据不同指标的特性，自主选择他们看重的特征，查看会议和期刊的排名。评估出处的指标包括：

* H5指数

H5指数是过去五年内该会议的所有文章中，满足有h篇文章的引用量不小于h的，最大的满足条件的h就是这个会议的H5指数。例：一个会议近五年有10篇文章，它们的引用量分别为10,8,8,7,6,6,4,3,2,1，那么这个会议的H5指数就是6。

* H5中位数

H5中位数是基于H5指数的，它描述的是该会议前[H5指数]篇文章引用量的中位数。例：同上，因为这个会议的H5指数是6，所以H5中位数为10,8,8,7,6,6的中位数，即为7.5。

* TK5指数

TK5指数是为了反映一个会议引用数最高的那些文章的质量如何。TK5中位数的定义是：找出一个近5年会议引用量最高的K篇文章（这里我们取K=10）并得到这K篇文章的具体被引列表，也就是分别都有谁引用了这K篇文章。根据得到的K篇文章的K个全引用列表，分别用每一篇文章的引用列表中每篇文章的citation指数基于H指数计算模型进行计算，得到K个H指数。我们把这K个数的中位数就叫做TK5中位数。

* 上升指数

理论上来说，一个文章表现得更加活跃的会议比一个引用量年年降低的会议要优秀。我们选取了每个会议近五年来引用量前100名的文章，筛选出它们之中每年引用量都在上升的文章，将这些论文命名为rising paper（上升论文）。对于每个会议筛选出来的上升论文算出一个总的H5指数，作为上升指数。

* 基础研究创新指数

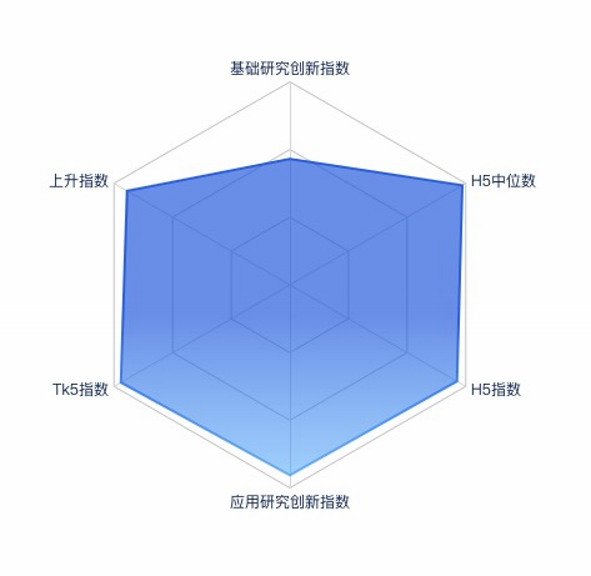
对于每个期刊/会议中近五年的全部论文，我们统计了其第一作者所在的机构。选取机构属于学术界的论文集合，将这一集合中的引用量平均值作为基础研究创新指数。

* 应用研究创新指数

对于每个期刊/会议中近五年的全部论文，我们统计了其第一作者所在的机构。选取机构属于产业界的论文集合，将这一集合中的引用量平均值作为基础研究创新指数。注：这两个创新指数分别体现学术界和产业界两个方向的研究创新情况，其比较结果对于学术界以及产业界的创新研究现状具有一定的参考价值。

* 指标图计算方法

在600个期刊会议中，我们找出每一个指数下的最高值，作为指标图该维度的最大值，并设定一个大于零的最小值，其余指标按比例显示在指标图上（如下图）。



##### 领域工作榜单

本系统结合学术指标筛选与行业学者复核的方式，提供某个领域内最具代表性、最前沿的工作清单。

##### 领域概念榜单

本系统结合学术指标筛选与行业学者复核的方式，提供某个学科/领域内最具代表性、最前沿的概念清单。

#### 事件/人格分析模块

##### 科技情报事件发现和自动摘要

##### 科技情报事件因果分析

##### 科技情报事件预测

##### 作者人格分析

##### 作者/事件社会评价

#### 用户管理模块

##### 用户订阅（作者/机构/出处/概念）

1. 订阅关注

关注您感兴趣的作者/机构/出处，可以获取该作者/机构/出处最新的科研动态。

关注您感兴趣的概念，可以在感兴趣领域的海量文献中发现最新的和高质量的工作。

1. 动态推荐

基于您的搜索浏览行为、关注的作者/机构/出处/概念以及收藏的工作等，可以推荐您感兴趣领域的最新工作和高质量工作，并基于此向您推送交叉领域和扩展领域的创新性工作。

##### 用户阅读管理

## 系统特色

### 与对标系统的比较

突出数据源，趋势分析和预测。

### 功能特色

本系统共设计大小功能共24项，通过对于市面上同类型系统的调研，本系统所独有的功能共7项，且均系首创，特色功能总结如下：

* 1. 作者/机构/出版商研究趋势

关键词丰富，折线图展示随年份变化趋势。

* 1. 机构社团划分
  2. 六度搜索
  3. 团队核心人物演化分析
  4. 科技情报事件因果分析
  5. 科技情报事件预测
  6. 作者人格分析
  7. 作者/事件社会评价
  8. 人物画像-人格分析：本系统首次将人格加入了对人物的刻画维度中；
  9. 人物画像-社会评价：本系统首次将人物的社会评价加入了对人物的刻画维度中；
  10. 人物画像-相关事件：本系统充分考虑了政治领域人物画像的特殊性，将人物相关事件（新闻舆情）加入到人物的刻画维度中，并首次设计了事件概述功能；
  11. 机构画像-机构内部关系网络：本系统首次将机构内部人物的关系网络加入了机构画像的刻画维度中；
  12. 事件分析-事件预测：本系统首次设计了基于新闻的政治事件预测功能；
  13. 群体画像-派系划分：本系统首次尝试按照党内派系来对政治人物进行划分；
  14. 群体画像-关键人物：本系统首次设计了群体关键人物发现功能。

### 算法特色

目前概要设计已经考虑实现的核心部分涉及后端算法12项，其中相比于现有算法已经有所创新的有：

1. 关系抽取

采用了一种基于BERT问答模型的开放式关系抽取算法QAORE，将开放式关系抽取任务建模为针对实体对的问题，预测句子中的文本段来作为关系指示词。QAORE通过使用预训练语言模型和问答机制，将实体对作为问题而关系指示词作为答案。相比于目前主流方法中仅使用实体的标签或位置指针来表示实体位置信息，QAORE能够更加充分地利用实体对的语义信息，通过计算语义的相关性实现在长句子或段落中抽取任意数量的关系。此外，本算法避免了使用额外的自然语言处理工具或依存关系解析，并且比非预训练的文本嵌入更为准确地刻画句子或段落的整体语义信息。在SAOKE和COER这两个中文标准数据集上的对比实验的结果表明QAORE在若干个评测指标上的结果均优于当前的主流模型。

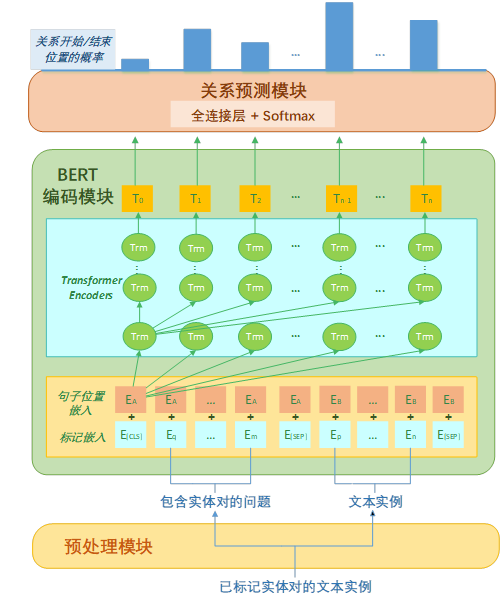


图1-2 开放式关系抽取算法QAORE的框架

1. 多文本摘要

使用了基于小规模模型的高效摘要方法。多文本摘要采用预训练语言模型从大量的语料库中学习到通用的知识。自然语言处理任务在预训练语言模型下都取得了更好的效果。目前的预训练语言模型的趋向于更大、更深、更多参数，但是这对硬件性能，尤其是显存大小有着更高的要求，同时也导致了算法存在鲁棒性差的问题，这使得这些学术成果难以被应用于生产实践。本算法采用预训练语言模型ALBERT。ALBERT采用参数共享等策略，模型的参数量仅为BERT的十分之一，却取得更好的效果。最近对于Transformer-block类型的预训练模型的研究表明：每个隐含层都蕴含丰富的语言，如词序、语法、语义知识。所以算法采用融合层根据所有句子表示计算出ALBERT每一层的权重，融合ALBERT每一层隐含的知识作为最后输出的补充。最终挑选出作为摘要的句子。

1. 人格分析

人格检测领域往往面临着新语种、低资源语言数据不足等多语言挑战，本系统从用户群体层面对多语言人格检测进行了探索。提出了一种基于多语言预训练和图神经网络的检测模型Per-GCN。该模型不从用户的多语言文本特征，多语言心理学特征，还结合了用户群体的结构信息来实现用户群体多人格检测。

多语言嵌入模型是一种功能强大的工具，可将不同语言的文本编码到共享的嵌入空间中，从而使其能够应用在一系列下游任务。针对人格文本数据的多语言性，本模型结合多语言预训练模型Multilingual BERT泛化能力，在中文英文语料上实现跨语言统一表达。

LIWC是一个透明的文本分析程序，用来统计不同心理意义类别单词的个数。由于LIWC能够检测各种情绪、思维方式、社会关系等，很多人格分析方法都使用LIWC来提取特征。我们的模型使用LIWC2015中文英文词典提取103个维度的语言特征作为每个用户语言特征的一部分，包括标准计数、心理过程、相关性、个人兴趣等维度。然后，结合用户文本语义特征和LIWC语言统计特征结合作为用户基本属性特征。

再从用户群体视角出发，利用用户文本相似度构造一个完整的用户群体结构图。具体的，对于每个用户通过设置语义相似度阕值选取top-k个相似度用户连边。需要说明的是，我们是根据用户群体之间的相似度分布来选择阈值的。我们的目的是要筛选出相似度低的用户，同时保证网络的不稀疏，让模型通过用户本身和相似度较高的邻居用户来决定用户的个性特征。实验表明，保留10%-20%的边缘可以获得最好的结果，这正是我们为每个数据集选择阈值的方法。

在构建户群体结构图之后，我们的模型通过图卷积神经网络聚合来自其相邻节点的信息来学习更好的实体表示。因此，我们的模型考虑了整个用户群体的结构信息。利用GCN对不规则网络结构建模优势我们将文本检测任务转换成一个用户群体图节点回归问题。具体的，在通过L层GCN之后，我们得到一组新的用户节点嵌入，它们在用户群体结构图中聚合来自其相似度高的邻居的属性。我们采用平均绝对误差(MAE)作为训练的目标函数。通过最小化预测值和真实人格值的MAE来更新训练权重。最后在对测试集中用户节点进行回归预测后得的用户的大五人格得分。利用该方法，本模型可以更好地利用用户群体之间的多语言语义关系，并增强其全局一致性，更加精准的建模用户群体多语言人格检测。

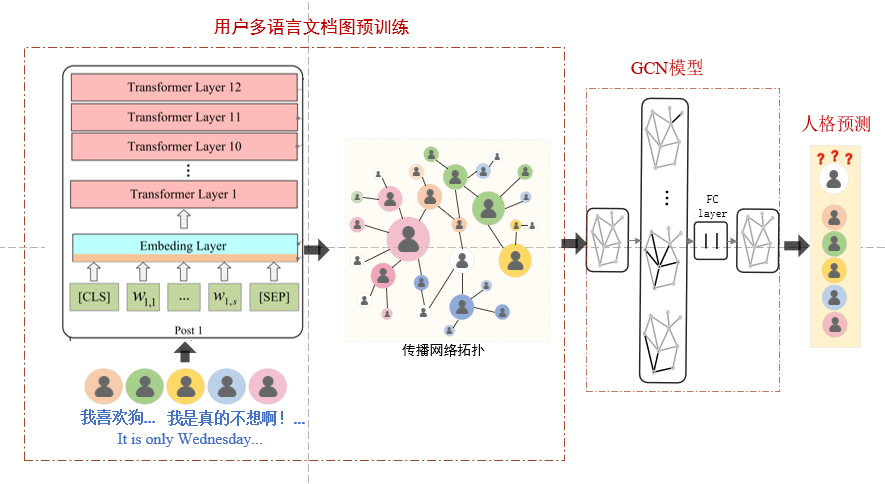


图 1-3 多语言与GCN结合的人格分析模型

1. 实体消歧：基于HITL和知识嵌入的实体消歧方法

使用实体抽取所得到的实体名，在wiki的http接口中进行查询，获取若干个wiki中候选实体的部分属性和简介，通过基于神经网络的实体链接算法，将这些候选实体的名称、属性、简介的语言学特征与待消歧实体的名称、所在句子、所在文章的语言学特征进行匹配，从候选实体中找出待消歧实体所指代的实体。

为适应实际应用的需要，实体消歧模型借用了HITL（Human-in-the-loop annotation）的思路，将数据标注和模型训练两个步骤结合了起来。当一个待消歧的实体名输入模型时，模型首先从wiki的查询接口中获取候选实体，经过与待消歧实体的相似度排序后输出候选实体列表。之后由人工从中选择出正确的实体，此时模型将从人类的选择中进一步进行学习，优化排序策略，如此迭代反复，直到训练完成。预测时则直接将排序后得到的相似度最高的候选实体作为结果输出。训练和预测的流程如下图所示：

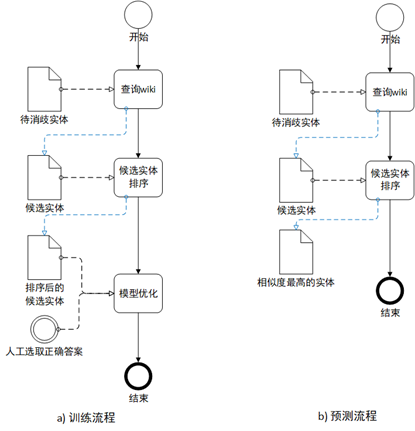


图1-4 实体消歧训练和预测流程

模型本身也可以提前使用一定量的标准数据进行初步训练，不过在实际的生产实践中，基于机器学习的算法模型往往需要部分实际数据的训练才能够更加适应使用场景，为此需要人工标注一定量的数据。相比于传统的实体消歧模型，这种HITL的方法借用了少样本学习和强化学习的思路，一方面，通过人类手把手的训练，模型可以快速地适应应用场景；另一方面，模型在迭代过程中准确率逐步提高，也让人工标注的工作量随着模型的优化而减少。这种训练方式对于样本不足的领域具有很高的适配性，能够大大减少生产过程中的时间和人力成本。

之后将所指代的实体的属性和简介以及wikipedia的实体编号组合为基础实体对象，将实体对象保存在neo4j数据库中，该实体与所在的文章之间也建立关联，此关联须保存至数据库中。同时，在实体抽取阶段所得到的实体原句中标记实体的wiki编号，以备关系抽取算法的使用，文章中的所有实体对象组合形成文章实体列表，以备其它算法的使用。

1. 关系归一：基于HITL和表示学习的归一化方法

关系的归一化同样借鉴了HITL的思想，算法的核心为关系词的相似度计算模型，我们将本系统的关系归一化操作分为两个阶段，其流程如下图所示：

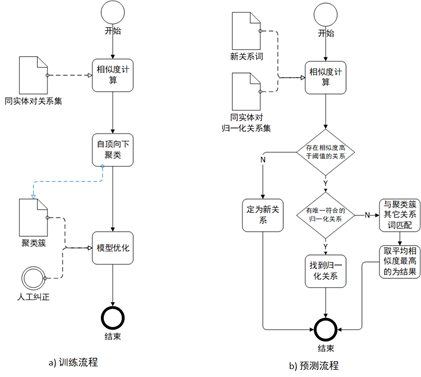


图1-5 关系归一化训练预测流程图

第一个阶段为训练阶段，在本阶段，系统会将该领域一定规模的历史数据进行开放式关系抽取的处理，所得到的<实体, 关系, 实体>三元组存入文件中备用。当所有的历史数据处理完成后，对于同一个实体对之间的关系，模型会根据每个关系间的相似度进行自上而下的层次聚类，并由人工对聚类错误的点进行重新分配，模型则从人工修改后的结果中进一步优化关系的相似度算法，如此迭代进行，直到训练结束，其余的数据均以聚类的方式进行划分，将每个聚类簇的中心点作为归一化的表述，此时整个聚类簇的归一化关系即为该中心点。

第二个阶段则为预测阶段，与实体链接问题类型，当基础数据的处理完成，系统投入使用并更新数据时，需要将新挖掘出的关系词链接到原有的归一化关系或者新的关系。具体来说，从文章中提取的新关系词将首先与该实体对间已有的归一化关系（聚类簇中心点关系）进行相似度计算：若存在唯一的一个归一化关系与新的关系词相似度大于某个阈值，将新的关系词归类于该归一化关系；若出现超过一个归一化关系与新的关系词相似度大于阈值，则分别计算每个归一化关系所属的聚类簇中其它关系与新关系词的相似度的平均值，取平均值最大者为新关系词的归一化关系；若相似度与所有归一化关系均低于某个阈值，则定为新型关系，新关系词本身就是的归一化关系。

在上述流程中得到的归一化关系将与实体对组成三元组，存储在neo4j数据库中，其它关系与归一化关系将被存储在Hbase数据库中。

1. 多媒体目标检测

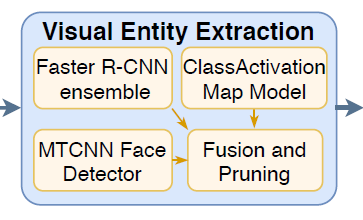


图 1-6 视觉目标检测模块图

数据源由文本，图片以及视频中抽取的关键帧组成

视觉上的实体抽取包括四个部分：

1. 通过Faster R-CNN来进行目标检测与识别。
2. 通过Class Activation Map模型进行特定情境下的实体与事件的检测与识别。
3. 使用启发式方法来给识别出来的所有实体进行去重操作，从而保证实体的质量。
4. 单独使用MTCNN来进行人脸检测，并作为person实体。

文本上的实体抽取可以依赖2.3.1模块进行。

1. 多媒体实体链接和共指

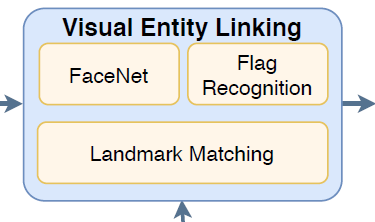


图 1-7 视觉实体链接模块图

实体链接的目标是将识别出来的实体与现实世界中的实体链接对应

对于人脸。由Google图片搜索根据公众人物的名称搜索公众人物的人脸，将可识别的公众人物存储在数据库中，作为一个可识别列表。对于从上一步提取出来的人脸实体，通过FaceNet从可识别列表中选择出对应的公众人物

对于地点。与人脸相似，先产生一个可识别的列表存储在数据库中，再将识别到的地点实体通过DELF模型对应到可识别的实体上。

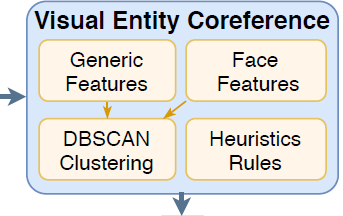


图 2‑6视觉实体共指模块图

实体共指的目标是识别出多个图片或者视频关键帧中出现的相同实体。

特征提取包括两个方面，一般性特征与人脸特征：

一般性特征，训练一个CNN模型，其任务是判别出不同图片中的相同实体，之后通过该模型来提取各个实体的特征向量。

人脸特征，通过FaceNet得出特征向量

共指方法包括两类：

第一类是通过DBSCAN来进行聚类，将同一个类簇中的实体视作共指实体。

第二类是利用启发式规则，比如，如果两张人脸图片对应的公众人物相同，则也识别为共指。

1. 多媒体数据融合

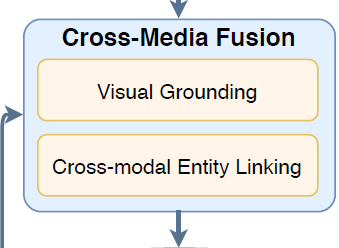


图 2‑7 跨模态融合模块图

此部分包括Visual Grounding系统，具体做法：

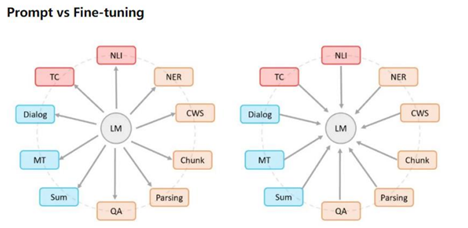
将Text实体及其句子feed到ELMo模型中，并其周围的图片通过CNN提取出的feature map进行比较。

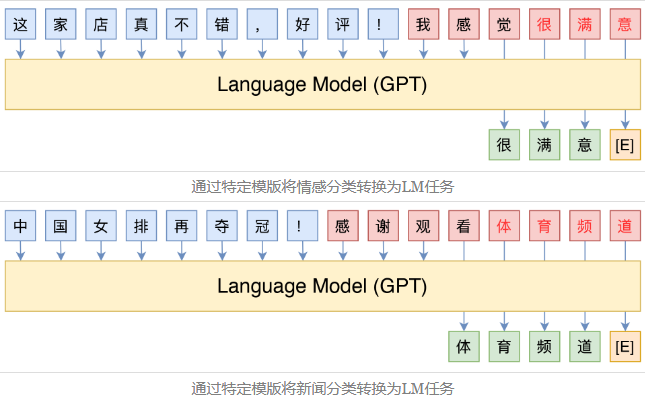
可以得到粗粒度每个图片的相关分数，以及图片之中的相关分数。如果图片相关分数超过阈值，则通过其图片内的相关分数得出bounding box，如果与已知实体出现了重叠，则赋值为已知实体，如果没有实体则新建实体。之后通过NIL聚类链接两个模态的实体。

1. 因果事理图谱构建

近年来深度学习技术、预训练模型的发展促进了NLP相关任务普遍采用Pre-train + Fine-tuning的范式。在该范式中，大规模预训练语言模型通过去微调去适配下游任务，但该范式严重依赖有标签的数据，未能充分挖掘预训练语言模型训练过程中学习到的语义知识。

一种新的通过对下游任务进行改造来适配语言模型的新NLP范式可以更加充分的挖掘预训练语言模型中的知识，并且极大降低模型对于有标签数据的依赖，在few-shot甚至zero-shot场景下也能取得非常好的效果。





在该事件因果关系判别算法中，将需要判别是否存在因果关系的两个事件利用一个模板来进行适当的改造，然后利用预训练模型的MLM任务来判断两个事件是否存在因果关系。具体来说，若需要判别的两个事件分别为a:”今天天气很冷”和b:“我需要多穿衣服“,通过一个模板p：”event1和event2之间[MASK]因果关系”，将事件a和事件b通过模板p分别填入event1和event2的位置，得到p(a,b)=”今天天气很冷和我需要多穿衣服之间[MASK]因果关系”，将该\_p(a,b)输入到经过预训练的BERT等预训练模型中，让这些模型来预测[MASK]位置应该填写的词，并且我们规定，[MASK]位置能被预测的词只有[存在，不存在]\_两种选项。

在实际模型中，设置的模板和[MASK]位置能被预测到的词有各种形式和选项，然后通过相关的实验来挑选出更符合自然语言形式的模板和[MASK]的词典，提高模型的性能。这样将原始的分类任务进行改造去贴近预训练任务，能更充分的挖掘预训练模型中的知识，在few-shot和zero-shot场景中取得非常好的效果。

然后通过预测出来的事件间的因果关系为边，将因果事件进行相连构建大规模的事理图谱，然后该图谱可以应用到后续的事件预测、可解释性分析、故事结尾生成等NLP任务中。

1. 群体画像

从整体上看，计划爬取台湾政治人物的推特账号关注信息、简介以及推特历史；构建出图后，使用图聚类算法进行群体划分，使用节点中心性和GNN估计节点重要性；通过构建时序图反应群体面对事件时的演化。

数据获取：属性图的构造

目前计划分析twitter上的关注及互动关系，同时将账号简介、推文等作为节点的特征进行统一分析。

先手工罗列一些知名的台湾政治人物（如马英九，蔡英文等）作为“种子”，从这些“种子”出发，沿着他们的关注列表对图进行广度优先搜索，即可获得台湾各界政治人物的关注关系图；而且，由于知名人物关注的人往往也不是无名小卒，通过这种方式爬取到的人物也更符合“有影响力的关键人物”这一标准。

群体划分：图聚类

在以文本信息为特征、以关注关系为边的图上运行各种聚类算法，就可以无监督地得到节点（政治人物）的类别信息；目前图上的聚类算法大致可以分为以下两种：

1. 谱聚类：对图的拉普拉斯矩阵进行特征分解，将较小的k个特征值对应的特征向量拼接堆叠成的矩阵作为图上节点的低维表示，之后在低维空间用常规的聚类算法（如K-means）进行聚类；
2. 基于GNN的图聚类：很多这个方向的工作都是以GNN为模型的基础结构，以重建图结构为学习目标，将习得的embedding用于聚类；此类方法在用到的特征、正则化技术和学习目标上有较多变化；

根据模块的具体场景，计划采用的算法可以大致划分入第二类中，此处的创新点在于，在进行聚类的同时，还试图得到关于文本特征中关键词的归因（Attribution）信息，这可能需要模型结构和损失函数上的特殊设计；这种思路的优势在于，把群体划分和词云生成放在了同一个阶段进行，提高了处理数据的效率，又通过二者的关系提供了一定的解释性。

群体关键人物发现：节点重要性估计

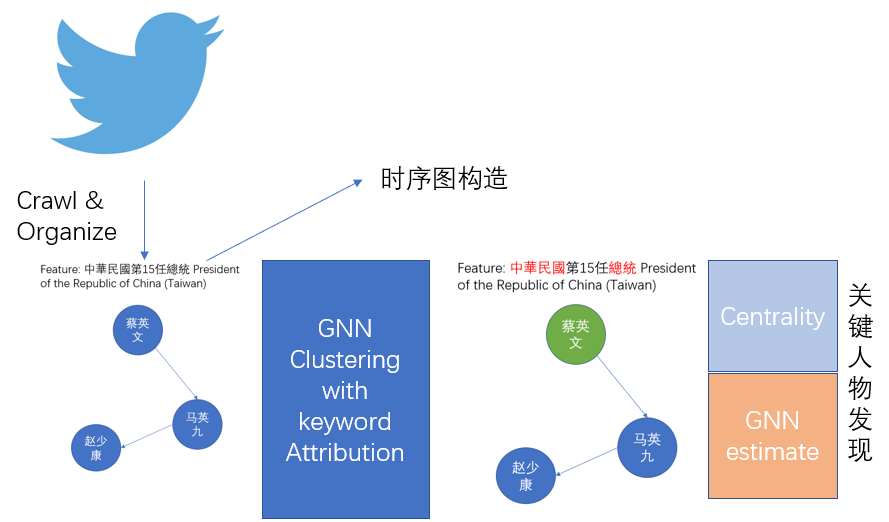
在构造好的关系图中，可以进行关键节点的分析，此处也有多种做法：

1. 直接根据图结构计算节点中心性；通过网络科学中常用的度中心性、特征向量中心性，PageRank中心性等，这些中心性能在一定程度上体现图中节点（政治人物）的重要性（影响力）；
2. 基于GNN的节点重要性估计：此类方法将GNN输出的节点与他的一阶邻居的embedding输入打分网络，经过若干轮聚合后输出这个节点的重要程度；其缺点在于训练中需要真实的重要性标签进行监督；

在当前的群体画像场景中，方法2的局限性主要体现在对于真实重要性的依赖；所以此处的难点和创新点就在于如何设计无监督的GNN节点重要性估计算法。

群体演变：动态图的构造与分析

根据数据形式，可以把关键人物的twitter活动（发了新twitter，参与到了某个话题的讨论）看作是一个“事件”，而把“事件”之间图中节点间的交互情况看作是边，这样就可以构造出一系列的时序动态图。而通过对某个“事件”发生前后，图的变化的分析，则可以窥探某些事件对政治群体演化的影响。作为更进一步的工作，还可以探索直接用GNN处理时序图，作为群体划分和关键人物发现的补充



# 系统架构的层次设计

## 数据采集层

### 学术搜索系统爬虫

### 新闻站点爬虫

利用爬虫程序从新闻网站获取新闻数据，摘取新闻的标题、网址、发布时间、主题、关键词、正文、图片、音频、视频，并以文件的形式保存。

新闻文章的正文和关键词部分是后续实体画像和事件抽取的主要数据。

目前来自《中时电子报》的数据共有68271篇文章，占用存储空间270MB。如按照相同的规模计算新加入的蓝绿阵营媒体《联合报》和《自由时报》的数据，则应共有20万篇左右的文章，占用存储空间810MB。

### 社交媒体爬虫

从社交网站（twitter，facebook）中查找重要人物的账号主页，并爬取人物发布的官方推文和社会大众评论。

这些数据将为人物画像中的人格分析和社会评价服务。

### 网络论坛爬虫

## 多模态知识获取层

### 文本实体抽取

通过基于神经网络的实体抽取算法，以序列标注的形式标记出文本中出现的人物、组织、地点、工作、概念实体，及其所处的句子，以备进一步的处理。

### 文本实体链接和消歧

使用文本实体抽取所得到的实体名，在wiki的http接口中进行查询，获取若干个wiki中候选实体的部分属性和简介，通过基于神经网络的实体链接算法，将这些候选实体的名称、属性、简介的语言学特征与待消歧实体的名称、所在句子、所在文章的语言学特征进行匹配，从候选实体中找出待消歧实体所指代的实体。

为适应实际应用的需要，实体消歧模型借用了HITL（Human-in-the-loop annotation）的思路，将数据标注和模型训练两个步骤结合了起来。当一个待消歧的实体名输入模型时，模型首先从wiki的查询接口中获取候选实体，经过与待消歧实体的相似度排序后输出候选实体列表。之后由人工从中选择出正确的实体，此时模型将从人类的选择中进一步进行学习，优化排序策略，如此迭代反复，直到训练完成。预测时则直接将排序后得到的相似度最高的候选实体作为结果输出。训练和预测的流程如下图所示：

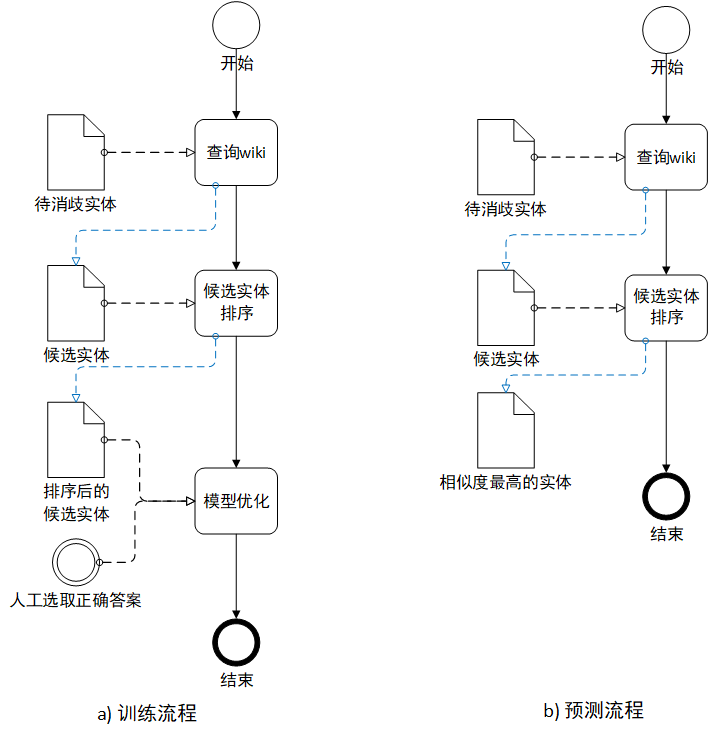


图 2‑2实体消歧训练预测流程图

模型本身也可以提前使用一定量的标准数据进行初步训练，不过在实际的生产实践中，基于机器学习的算法模型往往需要部分实际数据的训练才能够更加适应使用场景，为此需要人工标注一定量的数据。相比于传统的实体消歧模型，这种HITL的方法借用了少样本学习和强化学习的思路，一方面，通过人类手把手的训练，模型可以快速地适应应用场景；另一方面，模型在迭代过程中准确率逐步提高，也让人工标注的工作量随着模型的优化而减少。这种训练方式对于样本不足的领域具有很高的适配性，能够大大减少生产过程中的时间和人力成本。

之后将所指代的实体的属性和简介以及wikipedia的实体编号组合为基础实体对象，将实体对象保存在neo4j数据库中，该实体与所在的文章之间也建立关联，此关联须保存至数据库中。同时，在实体抽取阶段所得到的实体原句中标记实体的wiki编号，以备关系抽取算法的使用，文章中的所有实体对象组合形成文章实体列表，以备其它算法的使用。

### 文本关系抽取

文本关系抽取分为句内关系抽取和句间关系抽取两部分算法。

**句内关系：**以实体抽取和实体消歧得到的实体和实体原句为输入，筛选出含有至少两个实体的句子，并将含有超过两个实体的句子中的实体两两组合，最终得到若干个仅标记了两个实体的句子，使用基于语法标注和神经网络的开放式关系抽取算法进行句内开放式实体关系抽取，使用句中的动词、名词、代词、形容词等词语表示实体之间所具有的关系，并与两个实体的wiki编号一起组成关系三元组对象。

此处使用的关系抽取算法采用表示学习的方式，从大量的语料中学习了语法依存关系的表示，同时在语法树上使用了GCN模型，克服了近年来基于神经网络的开放式关系抽取算法无法充分使用语法信息的问题。

开放式关系抽取所得到的关系由于采用的是句子中原有的词语，相比于传统的封闭式关系抽取有着可读性强、鲁棒性强的特点。但其表示上却具有不一致性，即同一对实体间的不同关系词表述的是同一个关系，为此需要对关系进行归一化处理，将表达相同关系的关系词进行表述上的统一，减少冗余的连边。

**句间关系：**以实体抽取和实体消歧得到的实体和文章全文为输入，使用文章级关系分类算法在文章范围内进行跨句子关系抽取，判断不在同一个句子中共现的实体对之间是否也能够通过文章推断出关系，提取出的关系与两个实体的wiki编号一起组成关系三元组对象。

三元组对象主要包括两个实体的wiki编号、名称和关系的名称。句间关系得到的三元组对象本身已经进行了归一化，故将直接被存放在neo4j数据库中，句内关系则需进行进一步得关系归一化。

### 文本关系归一化

关系的归一化同样借鉴了HITL的思想，算法的核心为关系词的相似度计算模型，我们将本系统的关系归一化操作分为两个阶段，其流程如下图所示：

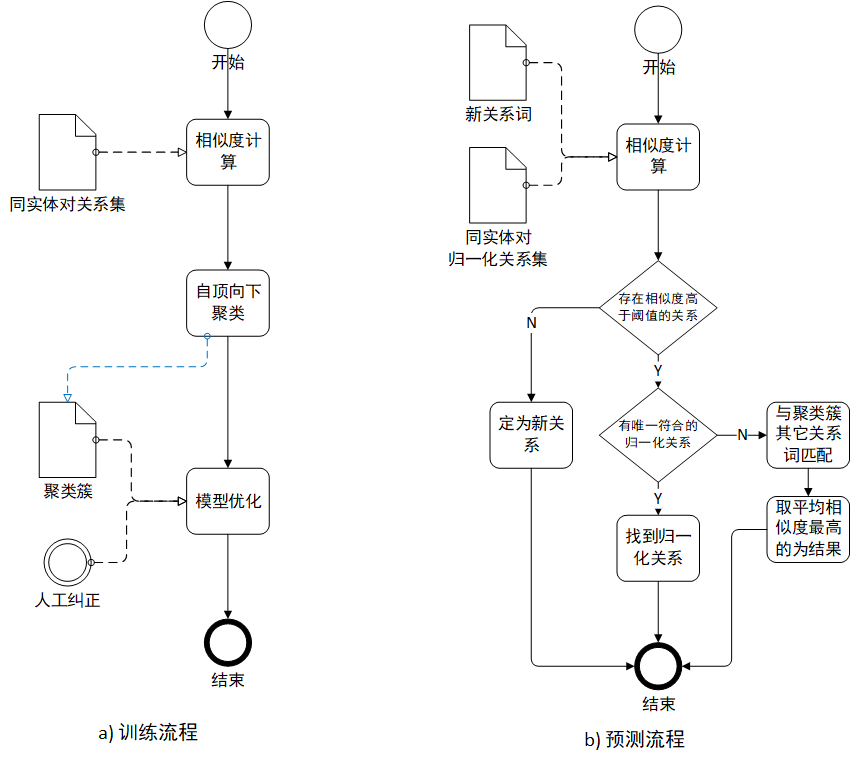


图 2‑3关系归一化训练预测流程图

第一个阶段为训练阶段，在本阶段，系统会将该领域一定规模的历史数据进行开放式关系抽取的处理，所得到的<实体, 关系, 实体>三元组存入文件中备用。当所有的历史数据处理完成后，对于同一个实体对之间的关系，模型会根据每个关系间的相似度进行自上而下的层次聚类，并由人工对聚类错误的点进行重新分配，模型则从人工修改后的结果中进一步优化关系的相似度算法，如此迭代进行，直到训练结束，其余的数据均以聚类的方式进行划分，将每个聚类簇的中心点作为归一化的表述，此时整个聚类簇的归一化关系即为该中心点。

第二个阶段则为预测阶段，与实体链接问题类型，当基础数据的处理完成，系统投入使用并更新数据时，需要将新挖掘出的关系词链接到原有的归一化关系或者新的关系。具体来说，从文章中提取的新关系词将首先与该实体对间已有的归一化关系（聚类簇中心点关系）进行相似度计算：若存在唯一的一个归一化关系与新的关系词相似度大于某个阈值，将新的关系词归类于该归一化关系；若出现超过一个归一化关系与新的关系词相似度大于阈值，则分别计算每个归一化关系所属的聚类簇中其它关系与新关系词的相似度的平均值，取平均值最大者为新关系词的归一化关系；若相似度与所有归一化关系均低于某个阈值，则定为新型关系，新关系词本身就是的归一化关系。

在上述流程中得到的归一化关系将与实体对组成三元组，存储在neo4j数据库中，其它关系与归一化关系将被存储在Hbase数据库中。

### 文本事件抽取和消歧

通过对于在同一时间收集到的所有新闻文章进行基于聚类的事件抽取，得到大量事件及其所发生的大致时间。将多篇新闻文章与其所蕴含的事件之间进行关联。与某个事件关联的若干新闻中出现次数最多的若干个实体（即与这些新闻关联最多的若干个实体），将被认为是事件的关键实体，与事件之间建立关联。通过对事件的时间、地点、人物、起因、经过、结果等属性的分析，将其与数据库中已经有的主题事件进行关联，或者与已有的事件组合成为新的主题事件。事件的属性、与实体的关联、所属的主题事件都将被存入Hbase中。

### 视觉实体抽取

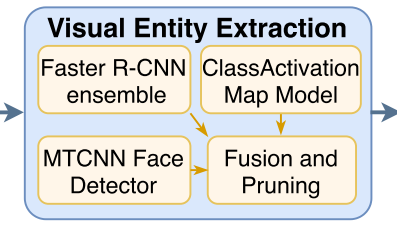


图 2‑4视觉目标检测模块图

视觉上主要需要检测的实体有人体、物体、人脸。

**人体和物体：**采用Faster R-CNN预训练模型抽取图片或视频关键帧中的object，主要思路是先通过启发式方法或RPN网络生成一系列稀疏的候选框，接着对候选框进行分类与回归，综合效率以及检测效果，采用R-CNN系列实例分割与检测算法。

**人脸：**人脸检测是一种在应用场景广泛的算法，可以识别数字图像中的人脸。人脸检测有时也代指人类在视觉场景中定位人脸的过程。在人脸检测中常用且目前比较前沿的算法就是MTCNN（Multi-task Cascaded Convolutional Networks）

**ClassActivationMap(CAM) Model：**生成类激活热力图，辅助特定情境的实体和事件抽取。

### 视觉实体链接和消歧

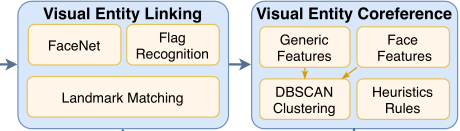


图 2‑5视觉实体链接和消歧模块图

视觉上，运用特定应用场景下的分类标签，训练模型专门用于几类标签的视觉实体识别，可通过将开放式图片搜索结果保存到知识库中，来进行识别分类器的单独训练。对于提取到的特征，可以利用DBSCAN Clustering的思想，采用启发式规则聚类进行实体消歧。

### 跨模态融合

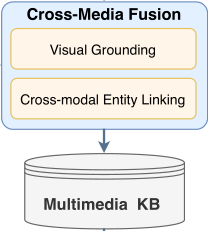


图 2‑7 跨模态融合模块图

为了充分利用多种模态的互补性，可以将多种模态的数据库组合成连贯的多媒体知识库。对于文本中提取的每个实体，将其和整个句子输入到ELMo或BERT等文本特征提取模型中，提取其相关信息及上下文特征，与相关图像特征进行比较，得出相关性得分和热力图，给定阈值以获得足够相关的图像并分配给已知的最重叠的视觉实体或新建一个视觉实体，使用一个类似NIL集群来链接匹配的文本和视觉实体。

## 知识库存储层

使用neo4j图形数据库和Hbase数据库，支持百万级节点和关系的存储管理。

原始数据通过数据中心进行临时存储，在与数据服务的实体信息功能交互后，仅将实体和关系信息存放到neo4j中，其余的实体和关系的附加信息，事件信息等一律存入Hbase，以避免过长的处理流程可能导致的数据丢失，并提高数据库读写效率。在相应数据分析处理完成后，得到的结果先存入数据库，在用户希望查看时再进行读取，除去了因实时处理而造成的响应时延，大大提高了用户的使用体验。

### Neo4j

Neo4j是一个高性能的,NOSQL图形数据库，它将结构化数据存储在网络上而不是表中。它是一个嵌入式的、基于磁盘的、具备完全的事务特性的Java持久化引擎，但是它将结构化数据存储在网络(从数学角度叫做图)上而不是表中。Neo4j也可以被看作是一个高性能的图引擎，该引擎具有成熟数据库的所有特性。程序员工作在一个面向对象的、灵活的网络结构下而不是严格、静态的表中——但是他们可以享受到具备完全的事务特性、企业级的数据库的所有好处。

Neo4j提供了大规模可扩展性，在一台机器上可以处理数十亿节点/关系/属性的图，可以扩展到多台机器并行运行。相对于关系数据库来说，图数据库善于处理大量复杂、互连接、低结构化的数据，这些数据变化迅速，需要频繁的查询——在关系数据库中，这些查询会导致大量的表连接，因此会产生性能上的问题。Neo4j重点解决了拥有大量连接的传统RDBMS在查询时出现的性能衰退问题。通过围绕图进行数据建模，Neo4j会以相同的速度遍历节点与边，其遍历速度与构成图的数据量没有任何关系。

### HBase

HBase是一个分布式的、面向列的开源数据库，该技术来源于 Fay Chang 所撰写的Google论文“Bigtable：一个结构化数据的分布式存储系统”。就像Bigtable利用了Google文件系统（File System）所提供的分布式数据存储一样，HBase在Hadoop之上提供了类似于Bigtable的能力。HBase是Apache的Hadoop项目的子项目。HBase不同于一般的关系数据库，它是一个适合于非结构化数据存储的数据库。另一个不同的是HBase基于列的而不是基于行的模式。

HBase是Apache Hadoop中的一个子项目，Hbase依托于Hadoop的HDFS作为最基本存储基础单元，通过使用hadoop的DFS工具就可以看到这些数据存储文件夹的结构,还可以通过Map/Reduce对HBase进行操作。

HBase不同于一般的关系数据库,它是一个适合于非结构化数据存储的数据库。所谓非结构化数据存储就是说HBase是基于列的而不是基于行的模式，这样方面读写你的大数据内容。

HBase是介于Map Entry(key & value)和DB Row之间的一种数据存储方式。就有点类似于现在流行的Memcache，但不仅仅是简单的一个key对应一个value，你很可能需要存储多个属性的数据结构，但没有传统数据库表中那么多的关联关系，这就是所谓的松散数据。简单来说，你在HBase中的表创建的可以看做是一张很大的表，而这个表的属性可以根据需求去动态增加，在HBase中没有表与表之间关联查询。你只需要告诉你的数据存储到Hbase的哪个column families 就可以了，不需要指定它的具体类型：char, varchar, int, tinyint, text等等。但是你需要注意HBase中不包含事务此类的功能。Apache HBase 和Google Bigtable有非常相似的地方，一个数据行拥有一个可选择的键和任意数量的列。表是疏松的存储的，因此用户可以给行定义各种不同的列，对于这样的功能在大项目中非常实用，可以简化设计和升级的成本。

由于我们项目的源数据包含各种各样的数据，并且都是非结构化的数据，所以HBase很适用于我们系统的源数据存储**。**

## 数据分析处理

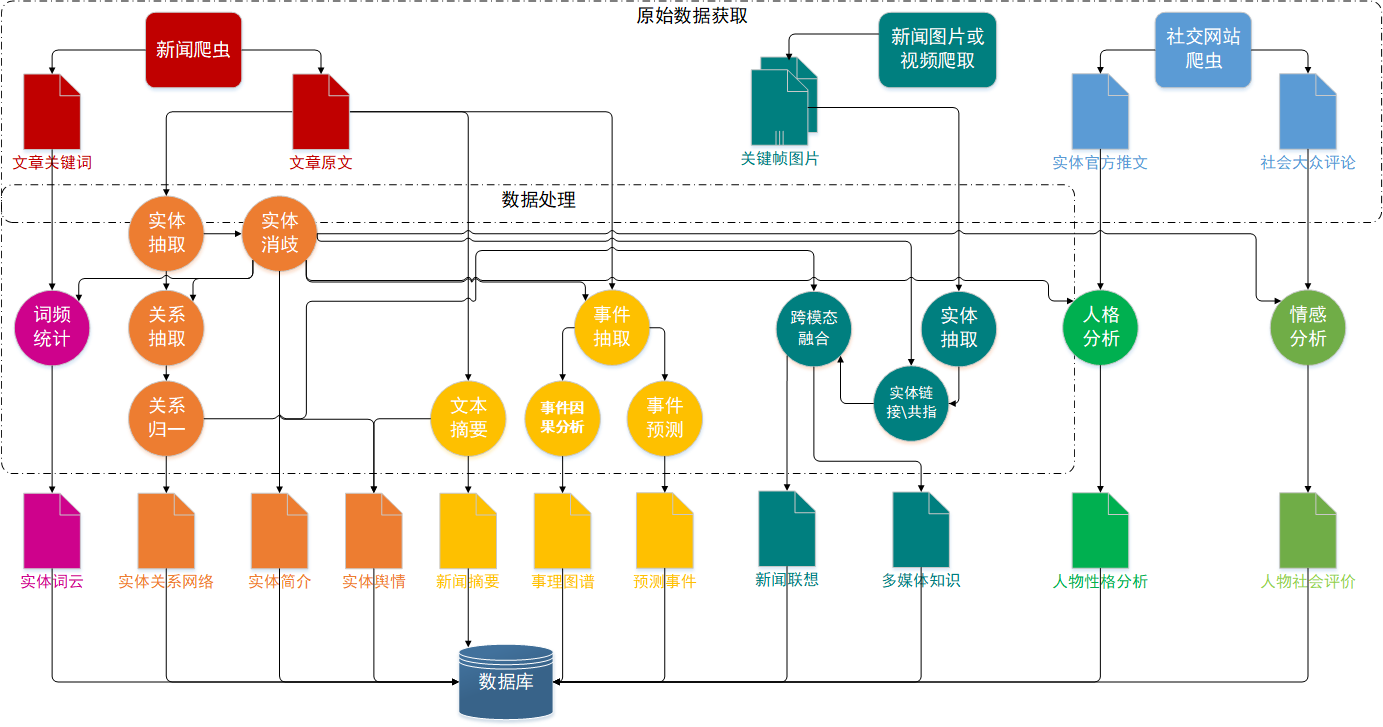


图 2‑1数据分析处理流程图（基础服务支持）

本层的处理流程如上图所示，目前的设计中只考虑了基础服务的支持，群体画像和部分事件分析的功能实现难度较高，故优先考虑了系统的核心算法以及与组员研究方向相关的算法实现。

### 实体抽取

通过基于神经网络的实体抽取算法，以序列标注的形式标记出新闻正文中出现的人物、组织、地点实体，及其所处的句子，以备进一步的处理。

### 实体消歧

使用实体抽取所得到的实体名，在wiki的http接口中进行查询，获取若干个wiki中候选实体的部分属性和简介，通过基于神经网络的实体链接算法，将这些候选实体的名称、属性、简介的语言学特征与待消歧实体的名称、所在句子、所在文章的语言学特征进行匹配，从候选实体中找出待消歧实体所指代的实体。

为适应实际应用的需要，实体消歧模型借用了HITL（Human-in-the-loop annotation）的思路，将数据标注和模型训练两个步骤结合了起来。当一个待消歧的实体名输入模型时，模型首先从wiki的查询接口中获取候选实体，经过与待消歧实体的相似度排序后输出候选实体列表。之后由人工从中选择出正确的实体，此时模型将从人类的选择中进一步进行学习，优化排序策略，如此迭代反复，直到训练完成。预测时则直接将排序后得到的相似度最高的候选实体作为结果输出。训练和预测的流程如下图所示：

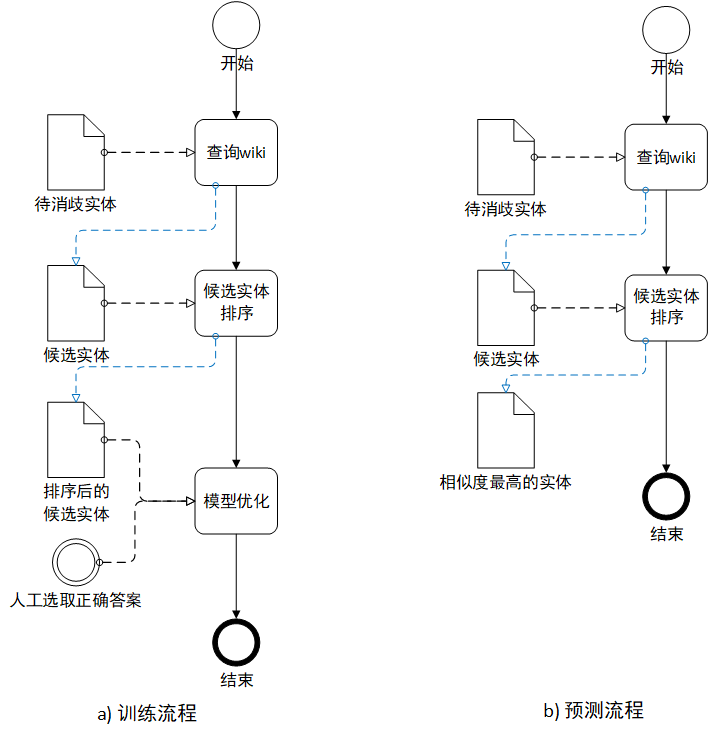


图 2‑2实体消歧训练预测流程图

模型本身也可以提前使用一定量的标准数据进行初步训练，不过在实际的生产实践中，基于机器学习的算法模型往往需要部分实际数据的训练才能够更加适应使用场景，为此需要人工标注一定量的数据。相比于传统的实体消歧模型，这种HITL的方法借用了少样本学习和强化学习的思路，一方面，通过人类手把手的训练，模型可以快速地适应应用场景；另一方面，模型在迭代过程中准确率逐步提高，也让人工标注的工作量随着模型的优化而减少。这种训练方式对于样本不足的领域具有很高的适配性，能够大大减少生产过程中的时间和人力成本。

之后将所指代的实体的属性和简介以及wikipedia的实体编号组合为基础实体对象，将实体对象保存在neo4j数据库中，该实体与所在的文章之间也建立关联，此关联须保存至数据库中。同时，在实体抽取阶段所得到的实体原句中标记实体的wiki编号，以备关系抽取算法的使用，文章中的所有实体对象组合形成文章实体列表，以备其它算法的使用。

### 关系抽取

关系抽取分为句内关系抽取和句间关系抽取两部分算法。

**句内关系：**以实体抽取和实体消歧得到的实体和实体原句为输入，筛选出含有至少两个实体的句子，并将含有超过两个实体的句子中的实体两两组合，最终得到若干个仅标记了两个实体的句子，使用基于语法标注和神经网络的开放式关系抽取算法进行句内开放式实体关系抽取，使用句中的动词、名词、代词、形容词等词语表示实体之间所具有的关系，并与两个实体的wiki编号一起组成关系三元组对象。

此处使用的关系抽取算法采用表示学习的方式，从大量的语料中学习了语法依存关系的表示，同时在语法树上使用了GCN模型，克服了近年来基于神经网络的开放式关系抽取算法无法充分使用语法信息的问题。

开放式关系抽取所得到的关系由于采用的是句子中原有的词语，相比于传统的封闭式关系抽取有着可读性强、鲁棒性强的特点。但其表示上却具有不一致性，即同一对实体间的不同关系词表述的是同一个关系，为此需要对关系进行归一化处理，将表达相同关系的关系词进行表述上的统一，减少冗余的连边。

**句间关系：**以实体抽取和实体消歧得到的实体和文章全文为输入，使用文章级关系分类算法在文章范围内进行跨句子关系抽取，判断不在同一个句子中共现的实体对之间是否也能够通过文章推断出关系，提取出的关系与两个实体的wiki编号一起组成关系三元组对象。

三元组对象主要包括两个实体的wiki编号、名称和关系的名称。句间关系得到的三元组对象本身已经进行了归一化，故将直接被存放在neo4j数据库中，句内关系则需进行进一步得关系归一化。

### 关系归一化

关系的归一化同样借鉴了HITL的思想，算法的核心为关系词的相似度计算模型，我们将本系统的关系归一化操作分为两个阶段，其流程如下图所示：

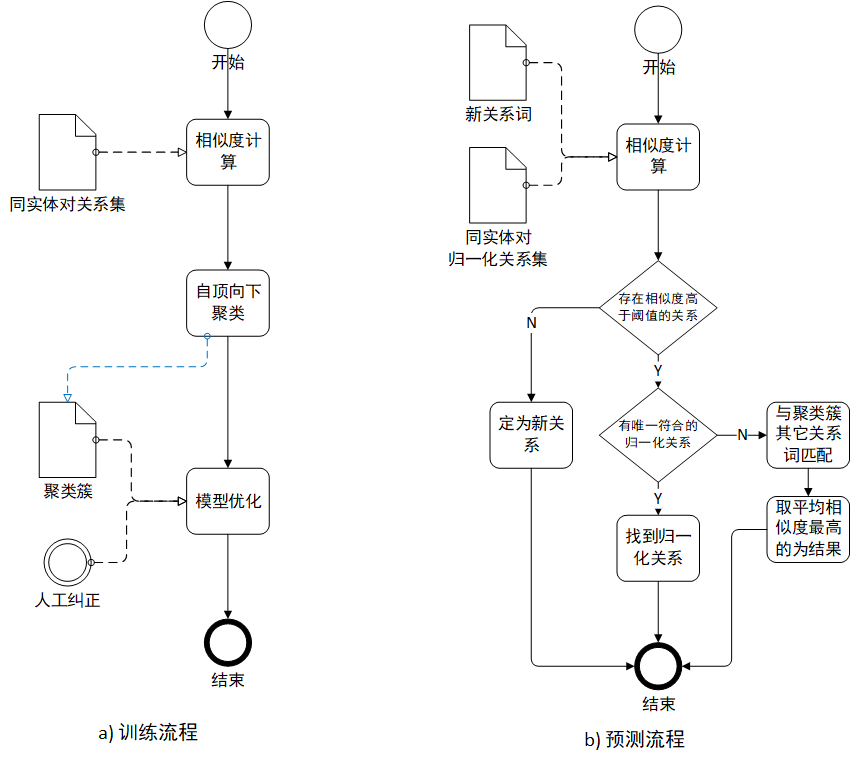


图 2‑3关系归一化训练预测流程图

第一个阶段为训练阶段，在本阶段，系统会将该领域一定规模的历史数据进行开放式关系抽取的处理，所得到的<实体, 关系, 实体>三元组存入文件中备用。当所有的历史数据处理完成后，对于同一个实体对之间的关系，模型会根据每个关系间的相似度进行自上而下的层次聚类，并由人工对聚类错误的点进行重新分配，模型则从人工修改后的结果中进一步优化关系的相似度算法，如此迭代进行，直到训练结束，其余的数据均以聚类的方式进行划分，将每个聚类簇的中心点作为归一化的表述，此时整个聚类簇的归一化关系即为该中心点。

第二个阶段则为预测阶段，与实体链接问题类型，当基础数据的处理完成，系统投入使用并更新数据时，需要将新挖掘出的关系词链接到原有的归一化关系或者新的关系。具体来说，从文章中提取的新关系词将首先与该实体对间已有的归一化关系（聚类簇中心点关系）进行相似度计算：若存在唯一的一个归一化关系与新的关系词相似度大于某个阈值，将新的关系词归类于该归一化关系；若出现超过一个归一化关系与新的关系词相似度大于阈值，则分别计算每个归一化关系所属的聚类簇中其它关系与新关系词的相似度的平均值，取平均值最大者为新关系词的归一化关系；若相似度与所有归一化关系均低于某个阈值，则定为新型关系，新关系词本身就是的归一化关系。

在上述流程中得到的归一化关系将与实体对组成三元组，存储在neo4j数据库中，其它关系与归一化关系将被存储在Hbase数据库中。

### 多媒体目标检测

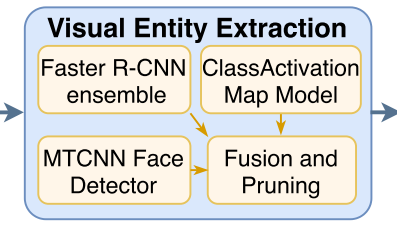


图 2‑4视觉目标检测模块图

视觉上主要需要检测的实体有人体、物体、人脸。

**人体和物体：**采用Faster R-CNN预训练模型抽取图片或视频关键帧中的object，主要思路是先通过启发式方法或RPN网络生成一系列稀疏的候选框，接着对候选框进行分类与回归，综合效率以及检测效果，采用R-CNN系列实例分割与检测算法。

**人脸：**人脸检测是一种在应用场景广泛的算法，可以识别数字图像中的人脸。人脸检测有时也代指人类在视觉场景中定位人脸的过程。在人脸检测中常用且目前比较前沿的算法就是MTCNN（Multi-task Cascaded Convolutional Networks）

**ClassActivationMap(CAM) Model：**生成类激活热力图，辅助特定情境的实体和事件抽取。

文本上的实体抽取可以依赖2.3.1模块进行。

### 多媒体实体链接和共指

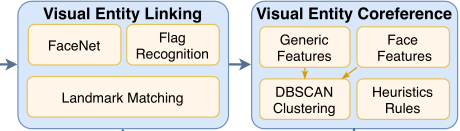


图 2‑5视觉实体链接和共指模块图

视觉上，运用特定应用场景下的分类标签，训练模型专门用于几类标签的视觉实体识别，可通过将开放式图片搜索结果保存到知识库中，来进行识别分类器的单独训练。对于提取到的特征，可以利用DBSCAN Clustering的思想，采用启发式规则聚类进行实体共指。

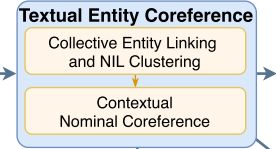


图 2‑6文本实体链接和共指模块图

文本方面，对可链接到知识库中已存在实体的相关信息，添加共指信息。对于无法链接到知识库的相关信息，可以采用NIL Clustering的思路，以启发式规则形成NIL集群，每个集群指向同一个实体。

### 多媒体数据融合

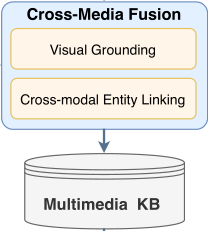


图 2‑7 跨模态融合模块图

为了充分利用多种模态的互补性，可以将多种模态的数据库组合成连贯的多媒体知识库。对于文本中提取的每个实体，将其和整个句子输入到ELMo或BERT等文本特征提取模型中，提取其相关信息及上下文特征，与相关图像特征进行比较，得出相关性得分和热力图，给定阈值以获得足够相关的图像并分配给已知的最重叠的视觉实体或新建一个视觉实体，使用一个类似NIL集群来链接匹配的文本和视觉实体。

### 事件发现

通过对于在同一时间收集到的所有新闻文章进行基于聚类的事件抽取，得到大量事件及其所发生的大致时间。将多篇新闻文章与其所蕴含的事件之间进行关联。与某个事件关联的若干新闻中出现次数最多的若干个实体（即与这些新闻关联最多的若干个实体），将被认为是事件的关键实体，与事件之间建立关联。通过对事件的时间、地点、人物、起因、经过、结果等属性的分析，将其与数据库中已经有的主题事件进行关联，或者与已有的事件组合成为新的主题事件。事件的属性、与实体的关联、所属的主题事件都将被存入Hbase中。

### 事件因果分析

从包含事件的新闻数据中，使用基于规则的方法对事件之间的因果关系进行抽取，得到大量的前置和后置事件对。对于这些事件对进行泛化，除去事件中的特殊属性，抽象出其中的共性，得到大量泛化前置和后置事件集合。对于泛化后的因果事件进行统计，按照事件对出现的频率构建事理图谱，并将泛化后的事件和因果关系存入neo4j数据库中。

### 事件预测

事件预测采用封闭式的事件预测方法，主要研究社会事件在政治领域的影响。当系统获取到一个社会事件时，使用模型预测该事件的发生对于若干个政治经济相关的后置事件的影响。目前的设计中这些事件包括：政府支持率提高/降低，事件人物社会评价提高/降低，社会稳定性提高/降低，民众生活满意度提高/降低。事件预测也将作为事件属性的一部分存入Hbase中。

### 多文本摘要

在得到事件和新闻文章的关联之后，首先将同一事件的相关文章的前部分（总结部分，一般是文章的第一段）结合在一起，组成一篇文章。然后对该文章进行分句、分词处理。将分好词的数据作为模型的输入，最终模型输出关于该事件的摘要。事件与其事件摘要和文章一起组合成事件对象，事件对象需要保存至数据库。

**算法创新：**多文本摘要采用预训练语言模型从大量的语料库中学习到通用的知识。自然语言处理任务在预训练语言模型下都取得了更好的效果。目前的预训练语言模型的趋向于更大、更深、更多参数，但是这对硬件性能，尤其是显存大小有着更高的要求，同时也导致了算法存在鲁棒性差的问题，这使得这些学术成果难以被应用于生产实践。本算法采用预训练语言模型ALBERT。ALBERT采用参数共享等策略，模型的参数量仅为BERT的十分之一，却取得更好的效果。最近对于Transformer-block类型的预训练模型的研究表明：每个隐含层都蕴含丰富的语言，如词序、语法、语义知识。所以算法采用融合层根据所有句子表示计算出ALBERT每一层的权重，融合ALBERT每一层隐含的知识作为最后输出的补充。最终挑选出作为摘要的句子。

### 人格分析

根据实体消歧产生的所有文章的实体列表，经去重后对每个人物实体，手动搜集其社交网站的账号主页，并爬取账号下发布的官方动态，以文件的形式进行保存。再从官方动态中分析并获取人物的大五人格分值，其中，大五人格包括：开放性、尽责性、外倾性、宜人性和神经质性，其通过五种特质涵盖了人格描述的所有方面。人格分析的五维分值将存入关系型数据库的人物实体表中。

针对获取到的每一条动态数据，将属于同一用户的所有动态进行合并，且动态之间使用<sep>分隔符标记。首先对原始合并后的文本，进行数据清洗，使用相关预处理工具（如NLTK）等去除非英文字符，网址链接字符串，用户名字符，表情符等，进而还原缩略词如we’ll →we will等以便后续统计特征信息。然后针对清洗后的文本统计相关特征，用于算法模块的分析。

**算法创新：**在预测人格特质分数此项任务上，现有的研究大多数是根据用户文本抽取其词汇特征，根据词频（如LIWC）等特点，来回归预测五人格分值，但此类方法忽略了用户说话的语义等信息，预测准确度有所偏差。为了弥补上述缺陷，出现了相当一部分研究基于深度学习的方法，运用RNN、CNN等网络捕获文本的语义信息等，再者则结合了LIWC等心理词典特征提升了预测的准确度，但是以往这些方法仅仅是学习文本表层蕴含的语义或词汇特征，其是以传统的文本处理角度去解决该任务，而人格分析不仅仅与浅层的词汇语义等相关联，还与个体发文的惯用的语言表达风格等有着紧密联系，这些也是推断个体人格的重要线索，所以人立方系统中使用的人格分析算法巧妙地结合了语言风格表达等领域知识，将其引入到人格预测这一任务上来，极大地增强了算法的预测能力，也具有很好的可解释性。该算法首先借助美国著名的社会心理学家James W. Pennebaker所著的《语言风格的秘密》一书，人工抽取了词语与风格之间的关系，构建一个词语到风格的层级图，再而根据获取到的数据集中每一个用户统计的特征值（+1或-1），构建与已有心理领域知识图的关联，用户通过这些特征间接与知识图连接，从而形成一个用户-词-语言风格的层次图，采用层次图注意力机制学习用户的表征，具体来说，先将与用户间接关联的语言风格层图输入到GAT网络中，聚集相对重要的语言风格特征，更新语言风格层节点特征，再根据语言风格层与词层节点的关联情况，也通过一个GAT网络更新语言风格-词子图的节点特征，使得与用户关联的词聚集到深层语言风格特征信息，最后使用用户所关联的词节点特征的池化来表征用户，从而接入到全连接层预测五人格分数。

### 社会评价

与人格分析类似，动态搜集其社交网站的账号主页，爬取主页下的大众评论，此外也可以根据关键词、话题等方式搜索其它的大众评论，大众评论以文件的形式保存。再对评论进行情感分析，以获得人物的社会评价，社会评价可以有多种维度。人物的社会评价分值将存入关系型数据库的人物实体表中。

## 应用服务层

### 知识图谱问答

知识图谱问答属于可视化层的用户交互部分。

输入问题后，根据开放式关系抽取得到的关系划分问题类别，以确定支持问答命令类型。之后根据问题的类型结合实体识别来确定目标实体类型或查询关键词，最终将转化为知识图谱数据库查询来获得结果。

# 数据库设计（暂定）

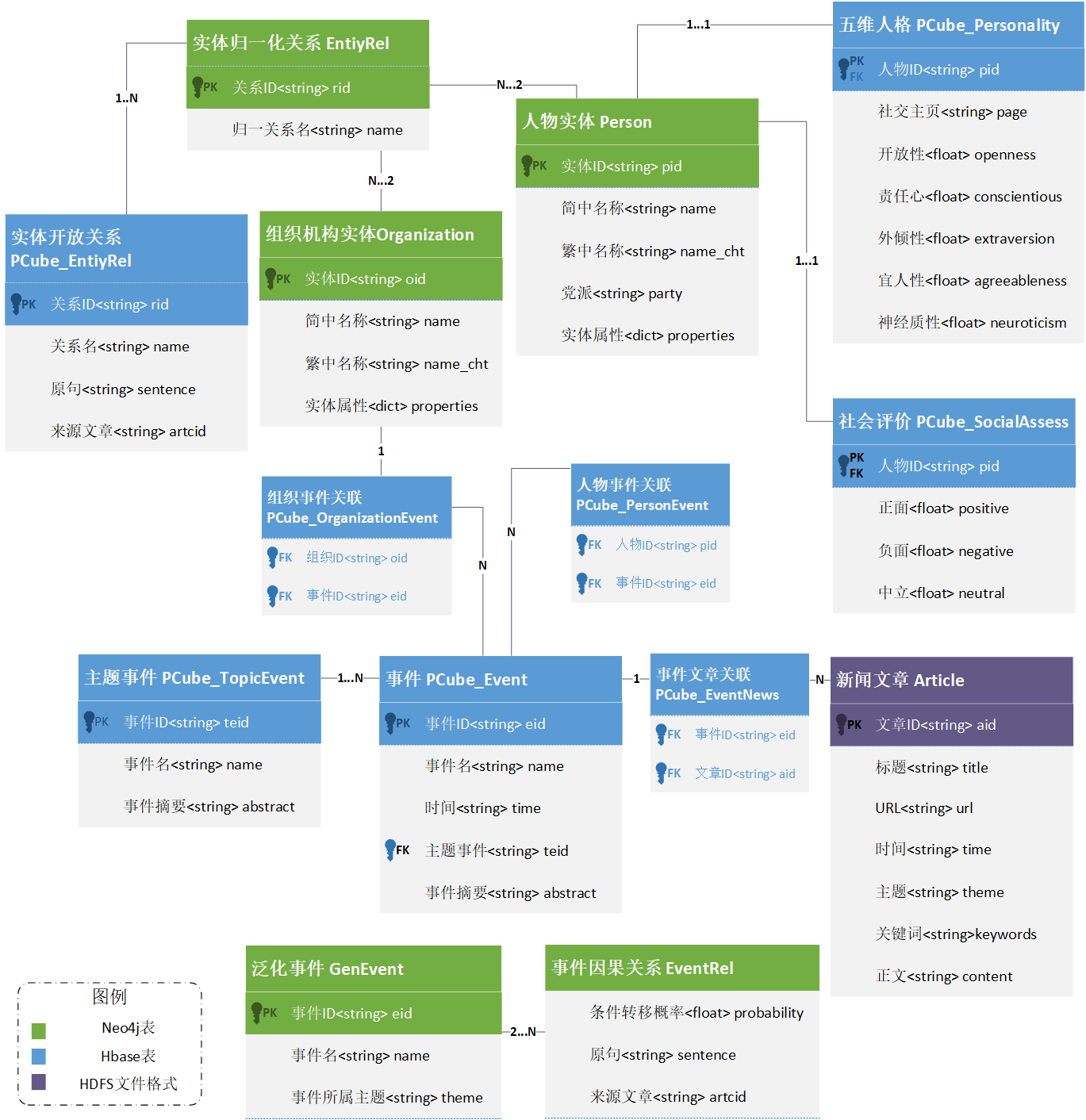


图 3‑1数据库设计图

# 接口设计（暂定）

## 原始数据获取接口

### 新闻爬虫接口

启动爬虫过程，获取新闻数据。

param date: 爬虫目标新闻的最早日期（可选，默认上前一个月）

return: [（news对象）{‘newsId’: 新闻id, ‘title’: 新闻标题, ‘url’: 新闻URL, ‘language’:新闻语言, ‘content’: 新闻内容, ‘topic’: 新闻主题, ‘keyWord’: 新闻关键词}, …]

### 社交网站爬虫接口

启动爬虫过程，获取社交网站数据。

param importantPerson: 爬虫目标重要人物列表

return: [（personPage对象）{‘personId’: 人物id, ‘url’: 主页URL, ‘kind’:网站名称, ‘post’: 帖子内容, ‘comment’: [评论1,评论2, … ] }, …]

## 数据存储接口

### HDFS写入

将爬虫的结果保存为HDFS文件。

param news/personPage: 爬虫结果

return: ‘success’/’error’

### Neo4j写入

param entityAlign: 消歧实体

return: ‘success’/’error’

## 数据服务接口

### 实体抽取接口

实体抽取，序列标注。

param text: 输入文本

return:[‘entity’: {‘entityId’: 实体ID, ‘entityKind’: 实体类型(person/organization/place), ‘entityName’: 实体名, ‘locatedSentece’: 所处句子, ‘entityModal’: 实体模态}, …]

### 实体消歧接口

实体对齐。

param entityName: 输入实体名

return: ‘entityAlign’: {

‘entityId’: 候选实体ID,

‘entityKind’: 候选实体类型(person/organization/place),

‘entityName’: 候选实体名,

‘locatedSentece’: 所处句子

‘locatedArticle’: 所处文章

‘wikiId’: wiki实体编号

}

### 关系抽取接口

param entity: 实体, isSameSentence: true/false 标识是否为句内关系

return: [ ‘entityRelation’: {entity1, relation: 提取出的关系, entity2}, …]

### 关系归一化接口

param [‘entityRelation’…]: 关系三元组列表

return: [ ‘entityRelation’: {entity1, relation: 归一化后关系, entity2}, …]

### 多媒体实体抽取接口

param [picture, …]: 输入图片序列

return:[‘visualEntity’: {‘entityId’: 实体ID, ‘entityKind’: 实体类型(person/organization/place), ‘locatedPicture’: 所处图片及位置}, …]

### 多媒体实体链接和共指接口

**视觉实体链接**

param [‘visualEntity’, …]: 实体抽取输出序列

return: [‘entity’: {‘entityId’: 实体ID, ‘entityKind’: 实体类型(person/organization/place…), ‘entityName’: 实体名, ‘locatedPicture’: 所处图片及位置, ‘entityModal’: 实体模态}, …]

**多媒体实体共指**

param [‘entity’, …]: 实体列表

return: [‘singleModalKnowledge’: {‘knowledgeId’: 知识ID, ‘knowledgeKind’: 知识类型(entity/relation/event), ‘knowledgeName’: 知识名, ‘knowledgeModal’: 知识模态(visual/textual)}, …] 可调用实体消歧及相关接口，生成每个模态的知识

### 多媒体数据融合接口

param [‘singleModalKnowledge’: {‘knowledgeId’: 知识ID, ‘knowledgeKind’: 知识类型(entity/relation/event), ‘knowledgeName’: 知识名, ‘knowledgeModal’: 知识模态(visual/textual)}, …]: 单模态知识列表

return: [ ‘multiMediaKnowledge’: {‘knowledgeId’: 知识ID, ‘knowledgeKind’: 知识类型(entity/relation/event), ‘knowledgeName’: 知识名, ‘knowledgeType’: 粗/细粒度类型}, …]多媒体知识列表

### 事件抽取接口

param text: 文章

return: [ ‘event’:

{‘eventId’: 事件ID,

‘eventName’: 事件名称,

‘eventTime’: 事件发生时间,

‘associatedNews’: [newsId, …],

‘associatedEntity’: [entityId, …]

}, …]

### 事件分析接口

param news 新闻数据

return: [ ‘eventGraph’: {event1, relation, event2}, …]

param news 新闻数据

return: ‘eventPrediction: 事件预测

### 多文本摘要接口

param event 事件

return: {

‘eventId’:事件ID,

‘eventSummary’: 得到的事件摘要,

‘articleId’: {articleTopic, articleURL} 事件包含的文章对象

}

### 人格分析接口

针对清洗后的文本统计相关特征，用于算法模块的分析，具体需统计以下特征：

|  |  |  |  |
| --- | --- | --- | --- |
| 特征 | 特征描述 | 值 | 备注 |
| WC | 字符总数 | -1或1 | 根据用户使用该类词的情况进行赋值，若高于所有用户平均水平则为1，反之则为-1 |
| Pronoun | 代词 |
| I | 第一人称单数 |
| we | 第一人称复数 |
| you | 第二人称 |
| he | 第三人称单数 |
| They | 第三人称复数 |
| Article | 冠词 |
| Prep | 介词 |
| Aux-verb | 助动词 |
| Usual-adverb | 常见副词 |
| Conj | 连词 |
| Negate | 否定词 |
| Verb | 动词 |
| Numeric | 数词 |
| Pos-emotion | 积极情绪词 |
| Neg-emotion | 消极情绪词 |
| Social | 社交有关词汇包括家庭、朋友及两性等 |
| Cognition | 认知过程相关词汇 |
| Insight | 洞察力词汇 |
| Causality | 因果性词汇 |
| diff-verb | 差异动词 |
| Assumption | 猜想性词汇 |
| Certainty | 可能性词汇 |
| Except | 转折词汇 |
| Belonging | 归属性词汇 |
| Achievement | 成就相关词汇 |
| Power | 权力相关词汇 |
| past-tense | 过去时态 |
| now-tense | 现在时态 |
| future-tense | 将来时态 |
| behave-sen | 动作短语 |
| Space | 空间 |
| Time | 时间 |
| big-word | 大词，政治哲学等领域词 |

param : 处理完成后的数据形式

以用户数为单位，一条数据格式如下:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| userId | userName | fmtPost | WC | pronoun | I | … | bigWord |
| str | str | str | int | int | int | int | int |

主要包含用户ID，用户名，处理过的该用户帖子文本，以及上述统计的各类特征等。

数据样例：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| userId | userName | fmtPost | WC | pronoun | I | … | bigWord |
| u123 | inggwen | post1<sep>post2<sep> | -1 | 1 | 1 | … | -1 |

return:

根据模型得到用户的五人格分值，主要输出数据形式如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| userId | userName | opn | con | ext | agr | neu |
| str | str | float | float | float | float | float |

### 社会评价接口

param comment(personPage, …)大众评论

return: {‘evaluation’: 社会评价}

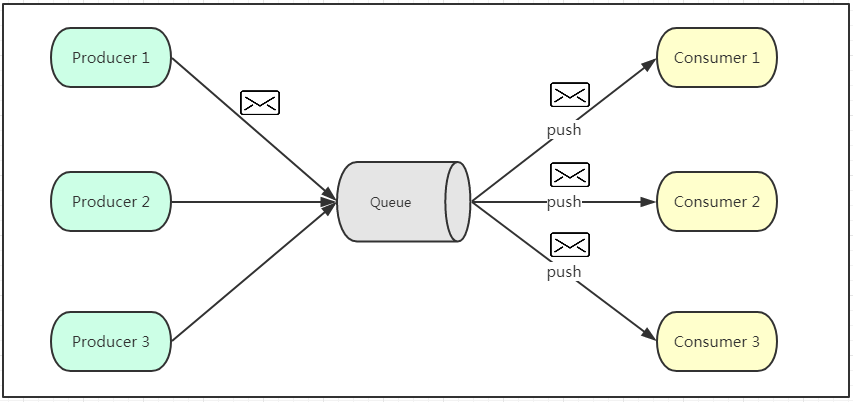
### 知识图谱问答接口

param question问题

return: {‘answerKG’: 知识图谱回答}

# 可提升增量（暂定）

## 数据获取与处理



使用Kafka消息队列进行数据获取与处理，乃至并行化模型的算法测试部分。在Kafka订阅生产者将消息放入消息队列后，队列会将消息推送给订阅过该类消息的消费者。由于是消费者被动接收推送，所以无需感知消息队列是否有待消费的消息。加入Kafka消息队列后，能够有以下的提升：

* 消息系统：解耦和生产者和消费者、缓存消息等。
* 流式处理：比如spark streaming和storm

## 多媒体目标检测及数据融合

在实体链接共指、目标检测等任务上，用多媒体知识进行辅助。视觉上可以用DBSCAN Clustering的思想，采用启发式规则聚类进行实体共指。文本上对可链接到知识库中已存在实体的相关信息，添加共指信息；对于无法链接到知识库的相关信息，可以采用NIL Clustering的思路，以启发式规则形成NIL集群，每个集群指向同一个实体。目标检测方面除了传统的视觉和文本方法，还可以使用ClassActivationMap(CAM)进行辅助

## 群体画像

能够合理进行党派划分与关系提取，目前正在进行数据准备，将由黄彦龙与张振浩进行相关调研与算法实现。

## 情感分析（社会评价）

将根据外网数据进行人物的评论等信息的获取，根据数据分析评论的相关情感及人物的社会评价。

# 人员安排（暂定）

## 算法部分（包含算法模型集成与对应数据获取及处理）

人格分析：朱洋甫

事理图谱：张威

知识问答：张鹤藐

群体画像：张振浩、黄彦龙

情感分析（社会评价）：宋威

实体、关系相关算法：杨卉帆

多媒体相关算法：宋威、杨卉帆

## 后端部分（主要为数据通道，保证并发量与鲁棒性）

开发人员：张振浩、黄彦龙、曹晨雨（晏成昊review代码）

## 前端部分（主要为数据的进一步处理与可视化展示）

开发人员：杨卉帆、宋威（张鹤藐）

## 其他（主要为新技术支持与底层维护）

负责人员：曹晨雨、晏成昊