港口及其临近水域强对流风险预警知识图谱的设计与实现

摘 要

现如今，灾害性天气尤其是强对流天气对于港口及其临近水域的人员生命财产安全造成了强烈的威胁。而在国内，就算是上海港这样的国际化港口对于强对流天气的预防机制依然不完善，无法紧跟国际前沿研究的步伐。因此，我们通过构建可视化的气象知识图谱为国内气象专业人员指出当前的国际研究前沿，并提供相关的资料、论文、图书以及机构，让国内能早日拥有比较完善的强对流风险预警机制，提升气象水文服务能力以及港口的助导航能力。

自从语义网这个概念被提出以来，越来越多的开放数据、开放数据库和用户生成内容被发布到互联网中。在此背景下，知识图谱概念首先由 Google 提出，用于描述世界中实际存在的各种实体以及概念，并且展示这些实体、概念之间的关联关系。知识图谱在实体层面（数据层）对本体进行了扩充；本体中强调的是概念以及概念之间的关系，描述了知识图谱的数据模式（数据模式层）。知识图谱在本体的基础上，增加了更加丰富的关于实体的信息。知识图谱在语义搜索、情报分析、智能问答、知识工程、反欺诈、数据挖掘和数字图书馆等领域有着广泛的应用。因此，在这次工程应用中，我们将知识图谱与气象两个领域结合起来，形成强对流方面的气象知识图谱。

本文使用爬虫从原始数据源中进行数据采集，原始数据源包括百度学术、维普、谷歌学术、必应学术、万方、知网等国内知名的学术网站。然后对数据进行清洗、实体对齐、知识融合等操作。对于结构化的气象方面关系型数据库，采用D2R映射来获取相应的RDF三元组。接着对得到的结构化数据进行本体/实体识别与关系抽取，将得到的关系型数据导入ES、Neo4j与MangoDB等数据库进行存储。最终使用PlantData平台导入数据形成可视化的知识图谱。

**关键词：**强对流天气，预警，知识图谱，数据采集，数据清洗

**Design and implementation of the knowledge graph of the strong convective risk warning for the port and its adjacent waters**

**ABSTRACT**

Nowadays, severe weather, especially severe convective weather, threatens the life and property safety of the people around the port and its adjacent waters. In China, even the prevention of severe convective weather mechanism of Shanghai port, the international port is still not perfect, can not keep up with the pace of international cutting-edge research. Therefore, we use the visualization of meteorological knowledge graph construction for domestic meteorological professionals pointed out the international research frontier, and providing relevant information, papers, books and institutions. So that the country can have a more perfect early warning mechanism of strong convective risk, enhance the meteorological service capabilities and port hydrological navigation capability.

More and more linked open data, open data base and user generated contents are published on the Web after the concept of Semantic Web has been proposed. The concept of knowledge graph has been fist proposed by Google, which is focused on describing various entities and concepts that exactly exists in the world, and their relations. Knowledge graph is actually a new vision of ontology; a knowledge graph extends an ontology in the entity level. Ontology usually focuses more on concepts and their relations, which specifies the schema of knowledge graph; and knowledge graph adds a large quantities of entities into ontology. Knowledge graph is widely used in semantic search, intelligence analysis, intelligent question-answering, knowledge engineering, anti-fraud, data mining, and digital library. Therefore, in this application of this project, we combine the knowledge graph with meteorology to form the visual knowledge graph of the weather.

In this paper, we will carry on the data collection from the original data source, then carry on the data cleaning, the entity alignment, the knowledge fusion and so on.

**Key words：**strong convection weather, early warning, knowledge graph, data acquisition, data cleaning

目 录

[1 引 言 1](#_Toc483278114)

[1.1强对流预警研究背景 1](#_Toc483278115)

[1.1.1国内外现状、水平和发展趋势 1](#_Toc483278116)

[1.1.2经济建设和社会发展需求 3](#_Toc483278117)

[1.1.3科学技术价值、特色和创新点 4](#_Toc483278118)

[1.1.4国内外强对流天气预警的研究成果 4](#_Toc483278119)

[1.2 知识图谱研究背景 5](#_Toc483278120)

[1.2.1知识图谱含义 5](#_Toc483278121)

[1.2.2国内外研究现状 6](#_Toc483278122)

[1.2.3知识图谱应用 7](#_Toc483278123)

[1.3 研究内容和意义 8](#_Toc483278124)

[1.3.1研究内容 8](#_Toc483278125)

[1.3.2研究意义 8](#_Toc483278126)

[1.4 研究现状及问题 8](#_Toc483278127)

[1.4.1气象知识图谱研究现状 8](#_Toc483278128)

[1.4.2知识图谱构建难点 9](#_Toc483278129)

[1.5 本文的组织 9](#_Toc483278130)

[2 知识图谱基本理论 10](#_Toc483278131)

[2.1 知识图谱构建理论 10](#_Toc483278132)

[2.1.2 本体学习 10](#_Toc483278133)

[2.1.2 实体层学习 11](#_Toc483278134)

[2.1.3 数据更新 11](#_Toc483278135)

[2.2 知识图谱的可用数据源分析 12](#_Toc483278136)

[2.2.1 数据源分析 12](#_Toc483278137)

[2.2.2 数据源采集 13](#_Toc483278138)

[2.2.3 数据清洗 14](#_Toc483278139)

[3 知识图谱构建核心技术 15](#_Toc483278140)

[3.1 实体关系抽取技术 15](#_Toc483278141)

[3.1.2 基于SVM的监督学习 15](#_Toc483278142)

[3.1.2 基于神经网络的监督学习 16](#_Toc483278143)

[3.1.3 基于Bootstrap的半监督学习 16](#_Toc483278144)

[3.1.4 基于联合聚类算法的无监督学习 16](#_Toc483278145)

[3.2 知识融合技术 16](#_Toc483278146)

[3.2.1 模式匹配 16](#_Toc483278147)

[3.2.2 实例匹配 16](#_Toc483278148)

[3.3 实体链接技术 16](#_Toc483278149)

[3.3.1 基于概率生成模型的方法 16](#_Toc483278150)

[3.3.2 基于主题模型的方法 16](#_Toc483278151)

[3.3.3 基于图的方法 16](#_Toc483278152)

[3.3.4 基于深度神经网络的方法 16](#_Toc483278153)

[3.4 知识推理技术 16](#_Toc483278154)

[3.4.1 基于符号逻辑的推理方法 16](#_Toc483278155)

[3.4.2 基于统计的推理方法 16](#_Toc483278156)

[4 方案设计 17](#_Toc483278157)

[4.1 气象知识图谱构建方案 17](#_Toc483278158)

[4.2 气象知识图谱构建流程 18](#_Toc483278159)

[4.3 方案介绍 19](#_Toc483278160)

[4.3.1 原始数据源 19](#_Toc483278161)

[4.3.2 数据采集 19](#_Toc483278162)

[4.3.3 数据清洗 23](#_Toc483278163)

[4.3.4 知识抽取 24](#_Toc483278164)

[4.3.5 数据处理 26](#_Toc483278165)

[4.3.6 图谱构建 28](#_Toc483278166)

[4.3.7 知识图谱存储 29](#_Toc483278167)

[4.3.8 知识图谱应用 35](#_Toc483278168)

[5 方案实现与结果展示 38](#_Toc483278169)

[5.1 数据源的选择 38](#_Toc483278170)

[5.2 数据采集与处理 38](#_Toc483278171)

[5.2.1 爬虫设计 38](#_Toc483278172)

[5.2.1 数据处理 38](#_Toc483278173)

[5.3 图谱存储 38](#_Toc483278174)

[5.4 图谱可视化 38](#_Toc483278175)

[6 结论和展望 39](#_Toc483278176)

[6.1 结论 39](#_Toc483278177)

[6.2 展望 39](#_Toc483278178)

[附录 40](#_Toc483278179)

[参考文献 41](#_Toc483278180)

[谢 辞 43](#_Toc483278181)

# 1 引 言

## 1.1强对流预警研究背景

国内以上海港为例

### 1.1.1国内外现状、水平和发展趋势

**（1）国际“E航海”和“智慧港口”发展趋势**

国际海事组织（IMO）近年来大力推进E航海（E-Navigation)[1]的发展理念。E航海的核心就是通过电子的数据交换手段，提高船舶的航行能力，从而保障航行安全、保护海洋环境及便利船舶航行，船岸之间的海事信息、海上服务信息及安全保障能力是E航海建设中的一个重要内容。我国航运业处于高速发展时期，港口吞吐量激增，港口外推、航路外移，船舶大型化趋势明显。如果一艘大型船舶在港口发生事故，其他船舶都不能靠港，这将导致港口瘫痪，不能正常运作，造成巨大经济损失。由于大型船舶的操纵和结构特性，对保障大型船舶安全进出港提出了严峻挑战。超大型船舶线型尺度大，受风等气象条件影响大于一般船舶，使得其在抛锚、港内航行、靠离泊位的困难和风险都明显增加。大型船舶进行引航服务的引航员登轮地点大多都安排在离港口有一定距离的外海开阔水域。这样的开阔水域一般风浪较大，对特定地点特定时间的气象条件要求较高。对于特殊船舶如LNG 运输船舶和化学品运输船舶进出港口有着严格的天气条件限制，如在雷雨、暴风雨雪等恶劣气象情况可能侵袭港口时，必须禁止LNG 运输船舶进港。在进出港航道航行、靠泊、装卸作业、在港系泊、离泊时对风速有不同的限制要求。因此开展上海港及邻近水域的气象灾害（强对流）风险预警技术研究将提升E航海工程气象水文服务能力，将提高上海港现有的助导航能力。

“智慧港口”[2]代表着未来港口的发展方向，世界先进的港口都已开始探寻向下一代港口的转变，包括汉堡、鹿特丹、新加坡、迪拜、上海港等。实现这种转变，各大港口需有效利用数字化技术和产业内外的协作，打造一个“3E级”港口，即在港口运营上卓越（Excel）、在生态圈构建上保持开放（Extend）、在可持续的创新业务上积极拓展（Explore）。根据埃森哲调研分析，在港口运营仅码头资源利用这一方面，由于衔接不畅、超时等待等造成的经济损失，以人民币测算，约70亿元；若长三角港口在现有基础上再增长15%作业效率，则每年可节约成本10亿元；码头效率的提升将减少船舶在港时间，降低10%的在港时间即可为船公司节约14.6亿元。同时，港口安全重要性凸显，港口安全已不再是单个企业的事情，直接关系到当地城市运营，重大港口甚至关系到国家安全。港口借助数字化新技术，开放协作、高度互联，发达国家开始建立港口社区系统（Port Community System），如德国的DAKOSY、荷兰的Portbase、英国的MCP Plc、法国的SOGET和西班牙的PORTIC系统，都旨在整合港口相关服务，借助海量的数据积累和大数据分析技术，提升港口的智能洞察潜在风险和智能化协作能力。

**（2）发达国家提供气象灾害风险预警的趋势**

美国海洋与大气局积极推动“地球系统大数据计划”（Big Earth Data Initiative）[3]，包括通过融合港口、海事和气象的大数据，提升港口的集疏运能力。在2015年第五届港口气象服务人员国际研讨会的会议总结中，也提出将融合电子海图、船舶自动识别系统(AIS)、气象、水文、航标动态、智能靠泊、港口信息、海事信息为一体的港口综合服务平台。

越来越多的气象部门认识到，传统的气象要素预报即便非常准确，但因为不了解用户在不同的生产管理环节中对不同气象要素风险控制、明确的气象要素指标、服务的时效节点、气象信息获取手段等方面要求，使得气象服务没有达到预期效果。

国际上的海洋气象业务，正在经历从传统预报预警向着风险预警业务转变的趋势。为了更好的改善现有的气象风险预警业务，世界气象组织正在积极倡导开展极端天气影响预报业务，美国、英国等发达国家已先后制定了极端天气影响预报发展计划，开展了极端性天气影响预报业务。美国国家大气局（NWS）正在推进的“时刻准备好应对各类天气事件的国家” （Weather Ready Nation）[4]，项目提出要大力发展基于影响预报的决策支持服务（IDSS），主要目的就是在原有预报及资料基础上，集成包括社会信息在内的尽可能多的相关信息，并从气象大数据中挖掘更有价值的信息，生成具有更大附加值的精细化预报产品，最终为决策者提供包括预报可信度、防御指引在内的风险预警产品。对于海洋气象业务来说，则是要在气象卫星、雷达、自动气象站的观测和数值加经验外推预报的基础上，集成包括水文、航标动态、码头靠泊、港口信息、海事等在内的相关信息，从气象与海事大数据中挖掘更有价值的信息，生成具有满足“智慧港口”卓越运营和“E航海”助航需求的精细化预报产品，最终为港口管理决策者提供基于上海港及邻近水域的包括气象灾害（强对流）预警、预报信息的可信度、气象灾害（强对流）防御指引在内的风险预警产品。

**（3）国内港口、海事和气象服务相关现状**

上海港与目前世界其他先进的港口都已开始探寻向下一代智慧港口的转变，致力于打造一个“3E级”港口。其中港口运营上卓越（Excel）包括：码头运营智能化、智能桥吊、智能车辆调度、智能泊位和智能闸口等，均需要海事和气象部门提供智能化的保障服务。根据GB/T14405-2011 《通用桥式起重机》、GB/T14406-2011 《通用门式起重机》、《岸边集装箱起重机（桥吊）安全技术操作规程》、《轮胎式集装箱龙门吊起重机安全技术操作规程》等规定，对码头运营相关作业设施的气象条件提出了明确的阈值规定。

上海市气象局承担了区域级海洋气象服务业务，是全国三大海洋气象中心之一，在国内率先开展海洋气象服务业务，但是在用户生产运营环节对气象服务需求的了解和针对港口和邻近水域安全的气象灾害预报能力上，均存在不足。

首先是定点、定时的精细化预报能力有待提高，尤其是针对海上强对流的在预报阈值和时效性方面与用户需求存在较大差距。如暂停集装箱作业的风速阈值是20米/秒，暂停桥吊作业的风速阈值是12米/秒，但目前上海气象部门提供的大风预报，仅在预警中提醒有出现7-9级大风的可能。造成预报能力不足的原因来自很多方面，包括对海上强对流发生发展的机理研究。目前对上海地区海风锋对强对流天气的触发机制研究尚处于研究阶段；不同天气背景下，基于热力、动力因子的强对流天气类型、结构特征、生命长度等展望预报技术的研究还处于不断深入的过程，需要基于用户特定区域和特定时间要求，搜集更多个例开展模式研究和参数改进。现有海洋气象服务产品时效的精细化程度不够，针对港口及邻近水域的气象服务产品的时间分辨率目前为12小时，对于生命史在几小时内的突发强对流天气系统没有有效的手段监测、报警，监控、外推其发生、发展趋势，还没有为港口管理用户提供急需的0-2、0-12小时的短时临近的气象灾害预警预报决策服务。

其次是海洋气象灾害风险预警专业服务机制尚未建立。气象部门对海事部门的气象服务需求了解不足，针对性的服务产品粗略匮乏，尚未开展以防御和减轻强对流等海洋气象灾害为目标的风险区划评估，更没有相关的海洋气象灾害风险预警业务。海洋气象灾害风险预警业务的建立需要气象信息与海事信息的深度融合，将气象信息嵌入到海事管理相关运营环节，如码头运营调度，大型船舶停靠泊、引水船作业等，这不同于气象部门现有的传统预报预警业务，属于当今前沿研究领域。

最后海洋气象信息发布技术落后，发布手段单一，导致用户获取信息便捷性和及时性不够。目前海洋气象信息发布手段比较落后，停留在利用电台语音广播和手机短信等单一途径，海洋气象信息发布到用户的时间滞后，无法发挥气象灾害预警预报信息的社会经济效益。海事部门对港口的信息服务已经走在世界前列，交通运输部东海航海保障中心与上海海事大学合作，推出的海e行电子航行示意图平台APP，已经成为航行警告信息发布的一个全新途径；海上安全信息数字播发系统（NAVDAT）是世界最新的海上安全信息的传输系统，具有速率高、覆盖广的特点,是E航海的重要通信设施，交通运输部东海航海保障中心研发了全球首座NAVDAT播发台，于2016年1月起试运行，覆盖范围达到250海里以上。如能通过本项目集成海事和气象的大数据和信息发布技术，将有助于上海港实现“3E级”港口的转型发展。

### 1.1.2经济建设和社会发展需求

港口和邻近水域的作业安全对气象条件非常敏感，受气象灾害直接影响的沉船、桥吊倾覆、雷击事故的发生，会造成高达数百万的经济损失，甚至有严重的人员伤亡。2003年韩国釜山港受强风影响，集装箱码头上52台装卸桥坍塌或出轨，直接经济损失达到5800万美元；2013年深圳孖洲岛突发强天气，致使8号泊位作业的2名工人死亡、一名工人受伤；2015年6月1日“东方之星”受下击暴流影响沉没；2016年3月, Pacific Victor轮从天津赴上海途中，在长江口等泊期间，因为遭受恶劣天气的影响，致使舱内3000吨卷钢严重移位，不得不停靠上海罗泾港进行重新绑扎、积载；2016年6月4日广元又发生因强对流导致的沉船事件。由此可见，国内外港口及相关水域因突发的强天气导致港口、航运作业的安全事故时有发生，进而引发港口管理问题。

上海港是我国沿海的主要枢纽港，我国对外开放，参与国际经济大循环的重要口岸。上海港以3653.7万标箱的吞吐量，连续6年稳居世界第一，到2017年总吞吐量将突破4000万标准箱，是目前全球港口年吞吐量的十分之一。据统计，2014年因天气恶劣而影响上海港口航道运行累计达1456.5小时（约占全年20%的通航时间），影响洋山港区作业累计达540.5小时（约占全年8%的作业时间）。随着上海国际航运中心的建设，港口进出口贸易的增加，气象灾害的风险和导致的潜在的经济损失也将与日俱增。

在影响港口作业的诸多气象灾害中（台风、强对流、暴雨、海雾、高温等），强对流天气的突发性最强，其带来的雷击和大风灾害对港口运行影响非常严重，同时预报时效和精准度要求高，难度非常大。上海港地处东海之滨，受海陆下垫面热力性质不同等因素影响，强对流灾害频发，也是目前气象预报预警业务中的薄弱环节，上海港港口作业和近海航运对强对流天气预报技术的提高有着迫切需求。

### 1.1.3科学技术价值、特色和创新点

通过对上海地区强对流天气的触发机制研究，发展不同天气背景下，基于热力、动力因子的强对流天气类型、结构特征、生命长度等展望预报技术，是提高海洋气象短临强对流预报能力的核心技术。将陆地上成熟的不断发展的强对流天气系统的监测与分析技术，临近（0-2小时）预报技术，短时（0-12小时）预报技术，叠加港口针对性观测，包括沿海风力实况和港口实时天气，融合港口的卫星、雷达重点监测，建立海洋气象灾害风险预警决策平台，这在国内属首创。

人工神经网络[5]是一种模仿及延伸人脑功能的新型信息处理系统, 它是由大量全局、局部或稀疏连接的简单处理器组成的非线性动力学网络系统。利用人工神经网络技术，对大量数据的整合（包括数值模拟预报产品，集合预报产品和实时观测数据），有望增加海洋气象预报的时效性和准确率。机器学习可以使计算机程序随着经验积累自动提高性能。基于对综合大数据的深度学习，将持续改善风险评估模型。

以“互联网+”和“E航海”理念提升航海气象保障服务社会民生的综合能力和水平。实现 “AIS+”[6]精准化和定制化服务，深化海上宽带通信技术应用，开展空间测绘技术、全球海上遇险与安全通信系统现代化及“E航海”关键技术研究。船舶自动识别系统（Automatic Identification System), 简称AIS系统）由岸基设施和船载设备共同组成，是一种新型的集网络技术、现代通讯技术、计算机技术、电子信息显示技术为一体的数字助航系统和设备。AIS系统的使用简化了船岸间的信息交互，增强了船舶交通管理系统、船舶报告制的功能，使航海界进入了数字时代。交通运输部东海保障中心已经在上海沿海和内河地区建成并运行由28个岸台组成的AIS岸基网络系统。海上安全信息数字播发系统（NAVDAT）[7]是世界最新的海上安全信息的传输系统，具有速率高、覆盖广的特点,是E航海的重要通信设施。东海航海保障中心研发了全球首座NAVDAT播发台，于2016年1月起试运行。AIS设备配置船舶数量多，而NAVDAT覆盖范围可达250海里以上，两者的结合应用可以覆盖更广泛的船舶用户，覆盖更大的海域面积，延长船舶对突发性气象预警信息的应对时间。

交通运输部东海保障中心与上海海事大学合作研制的海e行电子航行示意图平台APP，利用国际标准电子海图的光栅瓦片数据流技术，实现了电子海图的全覆盖，在此基础上成功叠加开发了航行警告图形化显示模块，可以图形化地显示各类航行警告信息，减轻了船舶驾驶人员的劳动强度，提高了信息标注的准确性。目前注册用户超过5万人次，实际在线用户超过6000人次，已经成为电子海图信息发布和港口航道服务信息发布的一个全新途径。

针对上海港气象风险开发强对流预报预警产品、基于人工神经网络技术提供风险预警和决策支撑、在国际上率先实现利用AIS岸基系统、NAVDAT广播系统和航海保障专用APP软件结合发布海洋气象信息均是气象和海洋灾害研究学科的前沿方向。

### 1.1.4国内外强对流天气预警的研究成果

较强的热力不稳定和适宜的动力环境是强对流发展的基础 ,造成灾害的强对流一般是一种深厚对流 ,深对流指数和对流有效位能可反映对流上升运动的潜势和强度 ,对流有效位能还隐含地反映了对流层大气总体垂直热力结构。下沉对流有效位能和大风指数反映了对流下沉运动和下击暴流潜势 ,对流下沉和中层干空气的入侵高度、干燥程度及对流层中下层的稳定度和湿度有关。强风暴特别是超级单体一般都具有很高的螺旋性 ,高螺旋度有利于风暴生命的维持 ,而风暴相对螺旋度则对风暴发生及风暴类型有一定的预示。粗里查逊数反映了对流能量和环境场动力之间的平衡关系 ,能量螺旋度指数反映了动力和能量对强对流天气发展的共同效应 ,它们都综合了动力和热力两方面的因子 ,对强风暴及其类型的预报有指示意义。风暴强度指数和瑞士雷暴指数成功地把动力和对流能量参数结合起来 ,在实际研究和业务工作中这种方法值得借鉴。随着高分辨率中尺度和风暴模式的发展 ,模式输出的对流动力和能量参数[8]将有广泛的应用前景。

廖玉芳等3人[9]提出的基于单多普勒天气雷达的强对流天气预警方法对发生在常德新一代天气雷达探测区内的强对流天气(雷雨大 风、冰雹、龙卷)在雷达回波强度场及速度场上的表现特征进行了归类分析,既用实例证实了美国强对流天气在多普勒天气雷达产品上的表现特征同样适用于中国, 又对其表现特征进行了完善。在此基础上建立了适用于当地的基于单多普勒天气雷达产品的强对流天气预报方法,并根据雷达实时体扫资料对强对流天气进行语音、文字自动报警以及对风暴移向、移速进行预报作了实用性的探讨。在多普勒天气雷达产品应用还处在探索、起步阶段的我国,该方法为制作强对流天气预报提供了较系统的预报思路。

杨引明等2人[10]对边界风廓线雷达在强对流天气预警方面做了大量研究。由于高时空分辨率的特征,风廓线雷达资料在局地暴雨、冰雹等夏季强对流天气预报中的应用正越来越受到人们的重视,上海LAP-3000边界层风廓线雷达由风廓线仪(Profiler)和无线电声波探测系统(RASS)两部分构成,其中风廓线仪用于探测大气边界层内各高度上的水平风和垂直速度,无线电声波探测系统RASS主要探测单站各高度上的温度。LAP-3000边界层风廓线雷达资料时间和空间分辨率较高,能有效揭示常规天气资料难以分析的一些大气动力和热力特征,在短时强对流天气预报中有较好的业务应用前景。

## 1.2 知识图谱研究背景

### 1.2.1知识图谱含义

知识图谱(Mapping Knowledge Domain）[11]，又名科学知识图谱，在图书情报界称为知识域可视化或知识领域映射地图，是显示知识发展进程与结构关系的一系列各种不同的图形，其使用可视化技术描述知识资源及其载体，挖掘、分析、处理、构建、绘制和显示知识及知识之间的相互联系。

知识图谱本质上是语义网络，是一种基于图的数据结构，由节点(Point)和边(Edge)组成。在知识图谱里，每个节点表示现实世界中存在的“实体”，每条边为实体与实体之间的“关系”。知识图谱是关系的最有效的表示方式。通俗地讲，**知识图谱就是把所有不同种类的信息（Heterogeneous Information）连接在一起而得到的一个关系网络。知识图谱提供了从“关系”的角度去分析问题的能力。**

具体来说，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。它把复杂的知识领域通过数据挖掘、信息处理、知识计量和图形绘制而显示出来，揭示知识领域的动态发展规律，为学科研究提供切实的、有价值的参考。迄今为止，其实际应用在发达国家已经逐步拓展并取得了较好的效果，但它在我国仍属研究的起步阶段。

### 1.2.2国内外研究现状

随着互联网中用户生成内容、开放链接数据等大量 RDF数据被发布到互联网上，互联网又逐步从仅包含网页与网页之间超链接的文档万维网（Web of Document）转变为包含大量描述各种实体和实体之间丰富关系的数据万维网（Web of Data）。

在此基础上，知识图谱这个概念最早由Google提出，主要是用来优化现有的搜索引擎。不同于基于关键词搜索的传统搜索引擎，知识图谱可用来更好地查询复杂的关联信息，从语义层面理解用户意图，改进搜索质量。

之后，各大搜索引擎纷纷推出了自己的知识图谱方案，如百度、必应等。比如在微软搜索引擎必应的搜索框里输入Bill Gates的时候，搜索结果页面的右侧还会出现Bill Gates相关的信息比如出生年月，家庭情况等等。

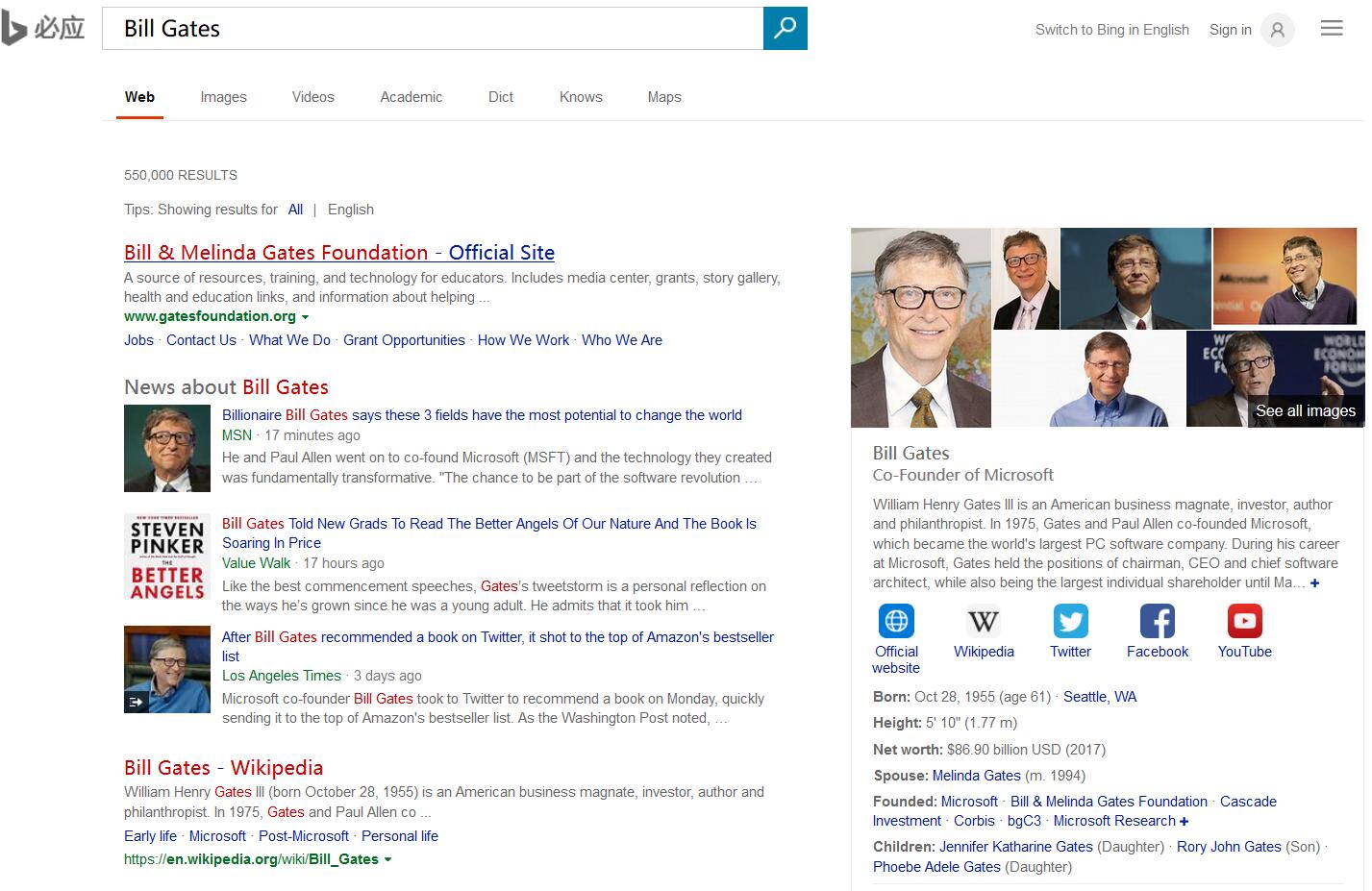


图1.1 必应搜索界面

知识图谱本质上是一种叫做语义网络的知识库，即具有有向图结构的一个知识库，其中图的节点代表实体或者概念，而图的边代表实体/概念之间的各种语义关系，比如说两个实体之间的相似关系或包含关系。

语义网络是19世纪50年代末60年代初提出，可以看成是一种用于存储知识的数据结构，即基于图的数据结构，这个图既可以是有向图也可以是无向图。使用语义网络，可以很方便的将自然语言的句子用图来表达和存储，用于机器翻译、问答系统和自然语言理解。

进入20世纪，语义网络又有了新的应用场景，即语义Web。语义Web与传统Web的一个很大区别就是网络用户可以上传各种图结构的数据（采取的是W3C的标准RDF），数据之间建立链接并建立链接数据。连接数据项目汇集了很多高质量知识库，比如Freebase、DBpedia和Yago，都来源于人工编辑的大规模知识库-维基百科。

与早期的语义网络相比，知识图谱具有自己的特点。首先，知识图谱强调的是实体之间的关系，实体的属性值以及概念之间的层次关系；其次，知识图谱的其中一个重要来源是百科，特别是抽取得到百科中半结构化的数据，与早期语义网络靠人工构建不一样，通过百科获取种子知识，然后通过知识挖掘技术可以快速构建大规模、高质量知识图谱；最后，知识图谱强调不同来源知识的融合以及知识的清洗技术。

### 1.2.3知识图谱应用

**知识计算**

主要是根据图谱提供的信息得到更多隐含的知识，如通过本体或者规则推理技术可以获取数据中存在的隐含知识；通过知识计算知识图谱可以产生大量的智能应用如可以提供精确的用户画像为精准营销系统提供潜在的客户；提供领域知识给专家系统提供决策数据，给律师、医生、公司 CEO 等提供辅助决策的意见；提供更智能的检索方式，使用户可以通过自然语言进行搜索；当然知识图谱也是问答必不可少的重要组建。

**链接预测**

预测实体间隐含的关系，同时使用社会计算的不同算法在知识网络上计算获取知识图谱上存在的社区，提供知识间关联的路径；通过不一致检测技术发现数据中的噪声和缺陷。

**知识推理**

推理能力是人类智能的重要特征，使得我们可以从已有的知识中发现隐含的知识，一般的推理往往需要一些规则的支持。例如“朋友”的“朋友”，可以推理出“朋友”关系，“父亲”的“父亲”可以推理出“祖父”的关系。再比如张三的朋友很多也是李四的朋友，那我们可以推测张三和李四也很有可能是朋友关系。当然，这里会涉及到概率的问题。当信息量特别多的时候，怎么把这些信息（side information）有效地与推理算法结合在一起才是最关键的。常用的推理算法包括基于逻辑（Logic） 的推理和基于分布式表示方法（Distributed Representation）的推理。随着深度学习在人工智能领域的地位变得越来越重要，基于分布式表示方法的推理也成为目前研究的热点。

**反欺诈**

反欺诈在金融领域是风控中非常重要的一道环节。基于大数据的反欺诈的难点在于如何把不同来源的数据（结构化，非结构）整合在一起，并构建反欺诈引擎，从而有效地识别出欺诈案件（比如身份造假，团体欺诈，代办包装等）。而且不少欺诈案件会涉及到复杂的关系网络，这也给欺诈审核带来了新的挑战。 知识图谱，作为关系的直接表示方式，可以很好地解决这两个问题。 首先，知识图谱提供非常便捷的方式来添加新的数据源，这一点在前面提到过。其次，知识图谱本身就是用来表示关系的，这种直观的表示方法可以帮助我们更有效地分析复杂关系中存在的特定的潜在风险。

**失联客户管理**

现实中，不少借款人在借款成功后出现不还款现象，而且玩“捉迷藏”，联系不上本人。即便试图去联系借款人曾经提供过的其他联系人，但还是没有办法联系到本人。这就进入了所谓的“失联”状态，使得催收人员也无从下手。而且这部分人群并没有以关联联系人的身份出现在我们的知识图谱里。如果我们能够挖掘出更多潜在的新的联系人，就会大大地提高催收成功率。

**精准营销**

一个聪明的企业可以比它的竞争对手以更为有效的方式去挖掘其潜在的客户。在互联网时代，营销手段多种多样，但不管有多少种方式，都离不开一个核心，即分析用户和理解用户。知识图谱可以结合多种数据源去分析实体之间的关系，从而对用户的行为有更好的理解。比如一个公司的市场经理用知识图谱来分析用户之间的关系，去发现一个组织的共同喜好，从而可以有针对性的对某一类人群制定营销策略。只有更好的、更深入的（Deep understanding）理解用户的需求，才能更好地去做营销。

## [1.3 研究内容和意义](#生成函数法及其优势)

随着大数据时代的到来，知识工程受到了广泛关注，如何从海量的数据中提取有用的知识，是大数据分析的关键。知识图谱技术提供了一种从海量文本和图像中抽取结构化知识的手段，从而具有广阔的应用前景。

### 1.3.1研究内容

本文将介绍基于原始数据层、互联网信息采集与清洗层、知识抽取层、知识融合层、图谱存储层、图谱应用层等方面来构建一个完整的气象知识图谱，通过可视化的图谱了解强对流天气预警方面研究前沿，了解在该方面的气象类型、气象专家、科研机构，获取相关论文、期刊、图书的名称、作者、所属机构等。

### 1.3.2研究意义

构建气象知识图谱或者说是强对流天气预警方面的知识图谱是很有意义的。

如今，在国内，关于强对流天气的预警方面的研究远落后于国外，而每年，由于对于灾害天气预警不及时所造成的直接间接人员伤亡、经济财产损失都很大，所以，应该及时了解国际上关于灾害性天气尤其是强对流天气预防预警方面的前沿科技，借鉴其中的可行性方案，学习并制定出符合中国国情的相关预警机制。

通过可视化的气象知识图谱，相关气象专业人员就可以及时了解并掌握国际先进的关于灾害性天气尤其是强对流天气预警方面的科研成果与技术手段，紧紧跟随该方面的研究前沿，改变中国现如今的落后局面。

## 1.4 研究现状及问题

### 1.4.1气象知识图谱研究现状

对于气象灾害方面，特别是强对流方面，现如今有关的数据库不多，其中的开放型数据库就更少了，这对构建知识图谱所需要的数据获取构成了阻碍。因此，需要从互联网外部数据，如气象网站、百科网站等通过网络爬虫来获取所需要的主要数据。

在此之前，很少人使用知识图谱手段来分析气象灾害特别是强对流预警方面的前沿科技，因此，在这个应用场景下，如何使用知识图谱、如何通过知识图谱进行分析是一个比较新的课题，这方面可供借鉴的经验、论文与书籍都很少。这就需要创新的思路将气象应用场景与知识图谱有机结合。

### 1.4.2知识图谱构建难点

**知识的自动获取**

由于手工获取知识来建立一个可以被计算机应用的知识库是一件费力费时并且效率极低的工作，因此知识自动获取相关的理论与方法相继被提出。但是这方面的工作一直是知识库构建过程中的难点之一。目前，国内，黄瑞提出了基于KDD算法的知识自动获取算法。但是综合来说，在知识自动获取研究领域国内处于落后状态。

**面向知识的表示学习**

如今的知识库通常使用表格形式，在此形式下，需要设计专门的图算法来使用并存储这些知识库，这样做不仅费时费力，而且存在数据稀疏问题。因此，基于深度学习的表示学习方法被提出，旨在将语义信息表示成低维的稠密实值向量。而知识表示学习的主要面向对象为知识库中的实体以及实体之间的关系，能有效解决数据稀疏问题并加快获取知识的速度、提升知识融合的效率。目前，在国内清华大学的刘知远等人对这方面的深度学习算法做了大量研究与创新。

**知识推理**

知识推理是指在计算机或智能系统中，模拟人类的智能推理方式，依据推理控制策略，利用形式化的知识进行机器思维和求解问题的过程。杨劲松等人提出了一种基于的Petri网的逆向推理方法。

## 1.5 本文的组织

第一章主要介绍研究背景，大概描述研究内容及其意义，介绍了构建强对流天气预警知识图谱所存在相关问题。

第二章主要从方法学上对知识图谱的构建过程的基本理论进行了描述，并且说明了知识图谱构建的相关数据源。

第三章主要描述设计方案与设计思路。

第四章主要展示可视化的图谱并作出相关的结论。

第五章对本文工作进行了总结，并提出了未来的研究方向。

# 2 知识图谱基本理论

## 2.1 知识图谱构建理论

### 2.1.2 本体学习

本体的构建步骤分为自上而下与自底向上。这里介绍通用知识图谱构建过程，即自底向上。自底向上依次包含术语、同义词、概念、分类、关系以及公理与规则。

1. 第一层次：术语抽取

术语是知识图谱领域中概念、实体或属性在语言学上的表示形式，术语抽取的目标是找到用于表示概念、实体或属性的相关术语或标记集合。术语抽取通常采用基于语言学规则的方法和使用统计的方法，需要经过分词与词性标注、形态学分析与使用统计学方法比较分析语料中的术语分布特征等步骤。

中文术语抽取方法与英文类似，但这方面研究较少。陈文亮等人[12]使用 Bootstrapping 机器学习技术，提出了一个从较大规模的语料中获取领域词汇的学习模型“FWB Model”。文献[13]中结合使用非线性函数与“成对比较法”，考虑词汇的位置和词频两个因素，计算候选术语的权重并从中自动选取术语。

1. 第二层次：同义关系抽取

同义关系抽取的目标是寻找那些代表同一概念、实体或属性的术语。主要方法有基于词典的方法、基于词法模式的方法和浅层语义分析。

而随着网络上开放链接数据的增加，基于开放链接数据和在线百科的方法也随之提出。Milne 等人[14]从维基百科中成功抽取了一个领域的辞典库，抽取的关系包括同义关系、反义关系和分类学习等。陆勇等人[15]使用百度百科作为抽取的目标，使用了字符串相似度、模式匹配和 PageRank 链接分析等方法结合，从中文百科语料库中自动获取同义词。

1. 第三层次：概念抽取

概念主要包含三个层次：内涵、外延与词汇发现。概念抽取的主要方法包括基于语言学的方法、基于统计的方法和混合方法。

同样，在互联网开放链接数据、用户生成数据以及垂直网站中也存在大量的数据的情况下，Cui等人[16]提出了一种从维基百科中获取本体构建所需要的概念方法。Zirn等人[17]则使用了一种方法从百科的分类体系中区分概念和实例的方法。

1. 第四层次：分类学关系抽取

分类学关系通常指的是概念之间的层次关系。通常采用基于词法模式、基于共现分析、基于语言学、基于开放链接数据和在线百科等方法。

近年来，基于开放链接数据和在线百科的方法收到越来越多的关注，文献[18]从维基百科的分类系统中构建了一个精确度非常高的分类系统。

1. 第五层次：公理和规则学习

公理与规则的地位在概念与关系之上，使得语义更加丰富，但是目前由于技术原因，对于该方面的研究较少，多数都是基于模板的抽取方法。

### 2.1.2 实体层学习

在本体上层构建完之后，需要为概念添加实例，称之为本体填充。知识图谱中，实体是最基本的元素，是“图”的基本构成。实体层的学习首先是从语料中识别发现实体，其次对实体的数据进行填充，而在此过程中，需要进行实体对齐。

1. 实体学习

实体学习又称实体识别，其目标是识别出语料中的人名、地名与组织机构名等一些专有名词。孙镇等人[19]对命名实体识别的研究进行了综述，列出了命名实体识别的主要方法。主要包括基于规则与词典、基于统计以及混合的方法。

1. 实体数据填充

通常对于一个实体，不仅只有名称，还有一系列的描述方式，包括实体描述、图片、同义名、属性等。

通常情况下，用户把实体发布到互联网的同时，也同时会发布一些描述实体的信息，因此， 这些互联网结构化或半结构化数据已经成为实体详细信息学习的主要来源。

另外，实体的信息还表示于海量的非结构化文本中。在传统的非结构化内容中，通常使用大段的文本对所描述的目标实体进行说明。由此可见，文本也是学习实体数据的重要来源。

1. 实体对齐

在构建知识图谱时，特别是当从多个数据源获取数据（实体）时，需要把这些描述同一个实体的数据进行对齐，称为实体对齐。其中的关键技术有实体链接、数据融合与知识融合。

实体链接的目标是把文本中包含的实体与知识库中的实体对应起来[20]，常用的链接目标为YAGO和DBpedia。

数据融合的目标是对来自不同数据源的数据进行整合。Castanedo[21]依据整合时使用和考虑的指标对数据融合的方法进行了归纳和分类。

### 2.1.3 数据更新

当数据源有更新或学习的方法有更新时，不可避免地要对知识图谱进行更新。知识图谱的更新主要分为两个层面的更新，数据模式层（本体）的更新和数据层的更新。

1. 数据模式层的更新

指的是本体元素的变化，比如概念的添加与删除、上下位关系的变更及概念属性的更新。通常情况下，数据模式层的更新是在人工干预下进行的，因为概念属性更新会影响到它的所有子概念。

1. 数据层的更新

指的是实体数据的更新，包括实体的添加与删除，修改实体的描述、属性等。通常，数据层的更新影响小，以自动的方式完成。

## 2.2 知识图谱的可用数据源分析

### 2.2.1 数据源分析

1. 结构化数据

结构化数据为存储在关系数据库或是面向对象数据库中的数据。关系数据库采用的是经典的关系模型，其结构非常简单，二维形式的表格非常直观，易于理解。

值得注意的是，存储于关系数据库中的数据基本都是面向行业的。因此，主要用于构建行业知识图谱。

1. 机器可读的开放本体或词典

机器可读的本体或词典通常是由人工构建的，因此具有良好的可靠性。英文领域中，用于本体构建最多的当属WordNet和Cyc与OpenCyc。而在中文领域中，最常用的是知网[22]和同义词词林[23]。

WordNet是一个由名词、动词、形容词和副词组成的一个同义词网络，是由普林斯顿大学建立和维护的英语词典[24]。它根据词条的含义将它们分组，每一个具有相同意义的字条组称为一个同义词集合，每一个同义词集合都有简短的概要和定义。

Cyc是一个通用知识库，将各领域知识综合在一起，于1984年建立。OpenCyc是它的公开发布版本。

知网是一个常识知识库，描述了汉英词语概念，并展示了概念之间与属性之间的关系。包括：上下位关系、同义关系、反义关系等。

同义词词林是1983年出版的一本词典，包括14706个词语，但之后一直没更新，哈工大实验室发布了其扩展版并增加了大量新词。

1. 开放链接数据与开放数据库

开放链接数据与开放数据库属于半结构化数据，通常由图的形式存储，其中最知名的是DBpedia[24]、YAGO[25]与Freebase[26]。在中文领域，Zhishi.me[27]是一个从三大中文百科：百度百科、互动百科与维基中文中映射得到的中文开放链接数据集。

DBpedia的目标是从维基百科中抽取结构化的数据并把这些数据公开发布到互联网中，基于所抽取的结构化的数据，它允许用户查询一些复杂的问题。DBpedia 包含了其它知识库的优点，包括：覆盖了多个领域，真实描述了社区的协议，能够自动地与维基百科保持同步，而且是多语言的。

YAGO是一个轻量级的可扩展本体，它具有很好的覆盖面和可靠的质量。YAGO 中包含实体间的层次关系和非层次关系，这些知识都自动地从维基百科中抽取而来，然后通过基于启发式规则的方法与 WordNet 进行合并。

Freebase 是一个由元组数据组成的大型人工协作知识库，其中的主要内容是由人工协作创造的。它使用自有的元信息查询语言（ Metaweb Query Language，MQL）提供基于 HTTP 的图形查询 API。

1. 行业知识库与行业垂直网站

行业知识库和行业垂直网站中的数据属于半结构化数据，它们描述的目标是特定的领域中的实体，描述范围方面比较窄，但数据一致性和完整性方面较通用的知识库完善。这些知识库或垂直网站通常具备自己的结构，在抽取其中的知识时，通常仅需要按照其结构解析即可。

1. 在线百科

在线百科是互联网中公开存在的最大数量的用户生成数据集合，这些数据具备一定的结构，属于半结构化数据。其中最知名的当属维基百科，中文领域则还有互动百科和百度百科。百科中的内容虽然是以HTML的形式的存在，但其中包含了许多结构化的信息，比如文章标题、分类标签、分类系统、信息模块、重定向、消歧页面和摘要。

1. 文本

文本可以说是最丰富的数据源，在互联网网页、开放链接数据、开放知识库、在线百科中，甚至是数据库中结构化数据的某些字段中，均存在大量的文本。文本最大的特点就是没有结构，因此从中学习知识的难度非常高。

### 2.2.2 数据源采集

1. 结构化数据采集

对于数据库中的结构化数据的采集，通常是由关系数据库中批量导入。而对于机器可读的开放本体和词典通常会提供下载的方式和接口。

1. 开放链接数据采集

DBpedia、YAGO、Freebase等开放链接数据集都提供了可供下载的仓库地址。如表2.1。

表2.1 开放链接数据集的下载地址

|  |  |
| --- | --- |
|  | 下载地址 |
| DBpedia | <http://wiki.dbpedia.org/Downloads39> |
| YAGO | <http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/yago-naga/yago/dawnloads/> |
| Freebase | <http://developers.google.com/freebase/data> |

1. 百科采集

维基百科提供了下载库地址，其中是XML格式的文件。

对于互动百科与百度百科，由于并没有提供相关的下载库，因此需要编写网络爬虫从中进行采集。在互动百科，有一个功能强大的分类系统，因此，可以根据分类系统编写相应的爬虫逐层爬取信息。而百度百科中的文章都是数字标号的，因此可以以数字为基准累积爬取数据。

1. 文本信息采集

对于通用知识图谱而言，互联网网页是最主要的文本信息来源，采集这些文本需要先把网页从互联网中采集然后提取其中的内容。

一般而言，需要通过编写相应的爬虫代码来采集互联网中的文本信息。常见的爬虫系统策略有：深度优先遍历、广度优先遍历、反向链接数、Partial PageRank、OPIC、大站优先等。

### 2.2.3 数据清洗

对于采集或者是下载到的开放链接数据、开放知识库数据或是百科中的数据，他们通常会有各自的格式，因此可以采用构建面向站点的包装器方法。

但是对于基于网络爬虫的网页本文数据，由于其结构各异，不能使用包装器。所以在提取文本时，通常会移除所有HTML标签，但这样会留下许多噪声信息，因此需要把这些噪声信息进行剔除。

Chang 等人[28]对从网页是抽取信息的方法进行了总结，通常包括：基于包装器的方法、基于模板的方法和基于统计的方法。

早期研究中，人们基本使用包装器来提取网页正文，在这里，包装器是一个针对目标中的数据制定了抽取规则的计算机程序。Muslea[29]等人提出了一个名为“STALKER”的包装器自动生成算法，算法的思想基于层次化信息抽取。

基于模板的方法包括生成模板与信息抽取。Ji[30]等人提出了一种从网页的HTML标签树中自动生成模板的技术。这个方法有一个很大的局限性，此方法需假设所有的网页都是使用相同的模板生成，而且抽取的网页也必须是一个集合，因此此方法使用不多。

基于统计的方法是通过对网页内容进行统计分析以找到当中的内容，Lin和Ho[31]通过假设网页的正文内容都位于表格标签<table>中，然后为<table>标签定义了“熵”，“熵”依据网页集合中单词的分布权重来衡量的；最后他们依据<table>标签的“熵”来确定正文内容的位置。但是，由于这个方法对网页做了过多的假设并且仅使用了网页的少部分特征作为依据，因此在准确率与通用性方面均有很大的局限性。

胡芳槐[32]提出了一种基于启发式规则的正文内容抽取方法，通过发现4个规则：正文内容通常比其他内容与标题的相似度更高；正文内容通常与标题比较接近；正文内容通常比网页的其他部分长度更长；正文内容中除了换行标签<br/>和段落标签<p>外，其他标签数量通常比较少。从而定义了一些概念：内容相似度、相对位置权重、文字数量、标签数量与标签影响因子作为启发式规则指标。在抽取正文时，将网页中每个信息块的指标值相乘，最后得分高的那块即是正文信息块。使用该方法抽取网页正文的准确率达到了98%以上。

# 3 知识图谱构建核心技术

知识图谱的构建目的是为了得到大量的机器可读的数据，而在互联网高速发展的今天，网络上充斥着大量的非结构化文本、半结构化网页以及结构化数据，知识图谱的构建即利用上述数据构建可视化的知识库。图3.1详细介绍了构建知识图谱的技术地图。

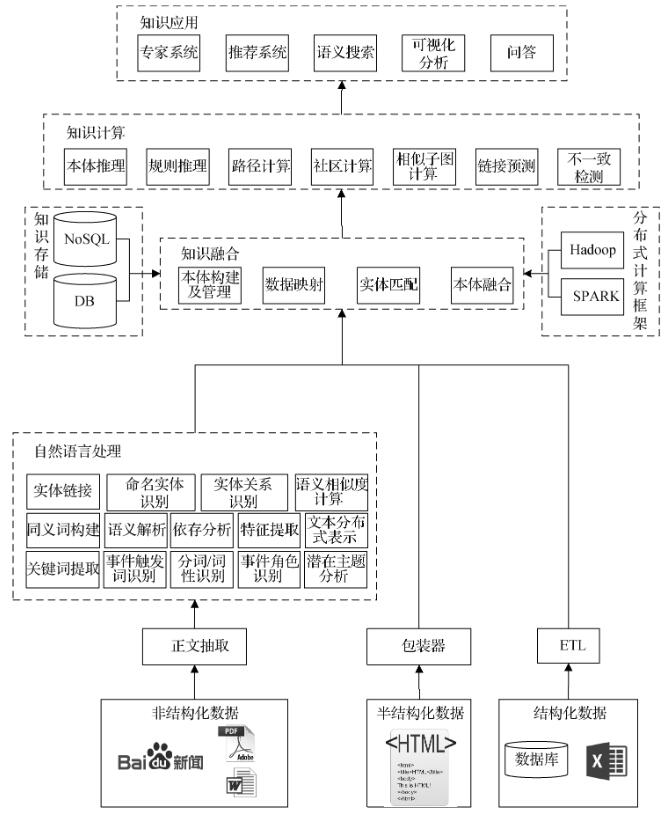


图3.1知识图谱技术地图

## 3.1 实体关系抽取技术

### 3.1.1 基于SVM的监督学习

### 3.1.2 基于CRF的监督学习

### 3.1.3 基于神经网络的监督学习

### 3.1.4 基于Bootstrap的半监督学习

### 3.1.5 基于联合聚类算法的无监督学习

## 3.2 知识融合技术

### 3.2.1 模式匹配

### 3.2.2 实例匹配

## 3.3 实体链接技术

### 3.3.1 基于概率生成模型的方法

### 3.3.2 基于主题模型的方法

### 3.3.3 基于图的方法

### 3.3.4 基于深度神经网络的方法

## 3.4 知识推理技术

### 3.4.1 基于符号逻辑的推理方法

### 3.4.2 基于统计的推理方法

# 4 方案设计

## 4.1 气象知识图谱构建方案

强对流方面气象知识图谱的构建及其应用方案的基本框架如下图4.1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；

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；

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；

图谱存储层：基于MongoDB的大规模三元组知识存储为所构建的强对流方面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。

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；



图4.1 气象知识图谱基本框架

## 4.2 气象知识图谱构建流程

气象知识图谱构建流程如图4.2所示。

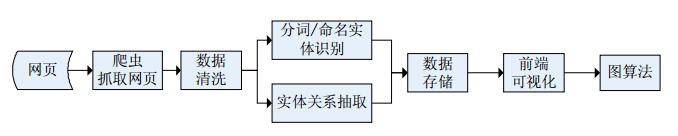


图4.2 气象知识图谱构建流程

流程介绍如下：

* 选择合适的网站设计与之对应的爬虫结构抽取网站内容（行业网站或关系数据库则使用只是抽取技术）；
* 对上一步进行爬取的网站内容进行清洗，将其中本文需要的内容从中抽取出来，比如将论文题目中<div>标签中提取出来，将论文作者从<p>标签中提取出来等；
* 对上一步经过清洗的数据进行进一步的处理，包括分词、命名实体识别与实体关系的抽取，命名实体识别包括将每个概念的中的实体从清洗过后的数据中提取出来，在这个过程中还将对其进行数据融合与实体对齐。数据融合指将从不同数据源提取的相同数据进行融合；实体对齐又称同义关系抽取，即将具有同义关系的实体进行合并，其中基于支持向量机（SVM）与基于CRF（Conditional random fields）的方法应用较为广泛。
* 接下来将处理好之后的结构化数据存入数据库中，现阶段比较受欢迎的是MangoDB与Neo4j这两个数据库。
* MongoDB 是一个基于分布式文件存储的数据库。由 C++ 语言编写。旨在为 WEB 应用提供可扩展的高性能数据存储解决方案。MongoDB 是一个介于关系数据库和非关系数据库之间的产品，是非关系数据库当中功能最丰富，最像关系数据库的。
* Neo4j是图数据库中一个主要代表，其开源，且用Java实现。经过几年的发展，已经可以用于生产环境。其有两种运行方式，一种是服务的方式，对外提供REST接口；另外一种是嵌入式模式，数据以文件的形式存放在本地，可以直接对本地文件进行操作。
* 之后就是图谱的可视化，也可称之为图谱的应用。其中有知识导航、关联分析、数据访问。

## 4.3 方案介绍

### 4.3.1 原始数据源

1. 结构化数据源

结构化数据主要来源于行业内部关系型数据库中的关系型数据，或者是行业知识库以及行业垂直网站中的数据。

1. 非结构化数据源

非结构化数据源是本次知识图谱构建的主要数据来源，因为行业知识库与企业内部关系型数据库不多而且大部分并非开源即不对网络用户开放。本文中气象知识图谱的非结构化数据主要来源于一些相关网站。主要包括的网站见表4.1。

表4.1数据源网站

|  |  |
| --- | --- |
| 学术期刊网站 | 知网、万方、维普、百度学术、谷歌学术等 |
| 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、互动百科、维基百科等** |

### 4.3.2 数据采集

数据采集主要包括采集强对流领域关键词、气象领域专家、气象领域科研机构以及学术论文等，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，具体如表4.2所示。

表4.2数据采集源网站

|  |  |
| --- | --- |
| 百度学术 | http://xueshu.baidu.com/ |
| 谷歌学术 | **http://scholar.google.com/schhp?hl=zh-CN** |
| 维普 | **http://www.cqvip.com/** |
| 万方 | **http://www.wanfangdata.com.cn/** |
| 知网 | **http://cnki.net/** |
| 中国天气网 | **http://www.weather.com.cn/** |

数据的采集最主要的工作是爬虫代码的设计与实现。

爬虫的设计理念主要分为深度优先遍历、宽度优先遍历（BFS）、带偏好的爬虫。

1. 深度优先遍历
2. 定义

假设给定图G的初态是所有顶点均未曾访问过。在G中任选一顶点v为初始出发点(源点)，则深度优先遍历可定义如下：首先访问出发点v，并将其标记为已访问过；然后依次从v出发搜索v的每个邻接点w。若w未曾访问过，则以w为新的出发点继续进行深度优先遍历，直至图中所有和源点v有路径相通的顶点(亦称为从源点可达的顶点)均已被访问为止。若此时图中仍有未访问的顶点，则另选一个尚未访问的顶点作为新的源点重复上述过程，直至图中所有顶点均已被访问为止。

图的深度优先遍历类似于树的前序遍历。采用的搜索方法的特点是尽可能先对纵深方向进行搜索。这种搜索方法称为深度优先搜索(Depth-First Search)。相应地，用此遍历方法就很自然地称之为图的深度优先遍历。

1. 实现思路

（1）访问顶点v；

（2）从v的未被访问的邻接点中选取一个顶点w，从w出发进行深度优先遍历；

（3）重复上述两步，直至图中所有和v有路径相通的顶点都被访问到。

1. 伪代码

递归实现

（1）访问顶点v；visited[v]=1；//算法执行前visited[n]=0

（2）w=顶点v的第一个邻接点；

（3）while（w存在）

if（w未被访问）

从顶点w出发递归执行该算法；

w=顶点v的下一个邻接点；

非递归实现

（1）栈S初始化；visited[n]=0；

（2）访问顶点v；visited[v]=1；顶点v入栈S

（3）while(栈S非空)

x=栈S的顶元素(不出栈)；

if(存在并找到未被访问的x的邻接点w)

访问w；visited[w]=1；

w进栈;

else

x出栈；

1. 宽度优先遍历
2. 定义

图的广度优先遍历BFS算法是一个分层搜索的过程，和树的层序遍历算法类同，它也需要一个队列以保持遍历过的顶点顺序，以便按出队的顺序再去访问这些顶点的邻接顶点。

1. 实现思路

（1）顶点v入队列。

（2）当队列非空时则继续执行，否则算法结束。

（3）出队列，取得队头顶点v；访问顶点v并标记顶点v已被访问。

（4）查找顶点v的第一个邻接顶点col。

（5）若v的邻接顶点col未被访问过的，则col入队列。

（6）继续查找顶点v的另一个新的邻接顶点col，转到步骤（5）。

直到顶点v的所有未被访问过的邻接点处理完。转到步骤（2）。

广度优先遍历图是以顶点v为起始点，由近至远，依次访问和v有路径相通而且路径长度为1，2，……的顶点。为了使“先被访问顶点的邻接点”先于“后被访问顶点的邻接点”被访问，需设置队列存储访问的顶点。

1. 伪代码

（1）初始化队列Q；visited[n]=0；

（2）访问顶点v；visited[v]=1；顶点v入队列Q；

（3） while（队列Q非空）

v=队列Q的对头元素出队；

w=顶点v的第一个邻接点；

while（w存在）

如果w未访问，则访问顶点w；

visited[w]=1；

顶点w入队列Q；

w=顶点v的下一个邻接点。

1. 带偏好的爬虫
2. 定义

很多时候，我们在URL队列中选择需要抓取的URL时，不一定要按照“先进先出”的原则，而是可以对URL进行选择，将需要的或是重要的URL挑选出来，又称为“页面挑选”（Page Select）。

1. 实现思路

影响网页重要性的因素有很多，比如链接的欢迎度、链接的重要度、平均链接深度、网站质量与历史权重等。

链接的欢迎度主要是由反向链接即指向当前URL的链接（backlinks）的数量和质量决定的。将之定义为IB（P）。

链接的重要度是一个函数，它仅考察字符串本身因此是一个关于字符串的函数，比如认为“.com”的URL重要度比“.cc”的URL重要度要高。将之定义为IL（P）。

以宽度优先遍历的原则计算出爬取网站的平均链接深度，认为离种子站点越近的网页的重要性越高。将之定义为ID（P）。

定义一个网页的重要性为I（P），则网页的重要度可以由如下公式（4.1）决定：

|  |  |  |
| --- | --- | --- |
|  |  | ( 4.1 ) |

其中参数X、Y可以调整IB（P）与IL（P）的权重，ID（P）是由宽度优先遍历算法决定的，这里不作为权重指标。

1. 实现方式

使用优先级队列来实现TODO表，并把每个URL重要性作为队列元素的优先级。

1. 示例

一个遍历图如图4.3所示，深度优先遍历算法的遍历顺序为：

A→B→G→C→H→D→E→I→J→K→F→L→M

宽度优先遍历算法的遍历顺序为：

A→B→C→E→F→G→D→H→I→L→M→J→K

假设图4.3中节点的重要性为D>B>C>A>F>K>E>H>G>I>J>L>M；则整个遍历过程如表4.3所示。

表4.3带偏好的爬虫遍历表

|  |  |
| --- | --- |
| TODO优先级队列 | Visited表 |
| A | **NULL** |
| BCFE | **A** |
| CFEG | **AB** |
| DFEHG | **ABC** |
| FEHG | **ABCD** |
| EHGL | **ABCDF** |
| HGIL | **ABCDFE** |
| GIL | **ABCDFEH** |
| IL | **ABCDFEHG** |
| JL | **ABCDFEHGI** |
| KL | **ABCDFEHGIJ** |
| L | **ABCDFEHGIJK** |
| M | **ABCDFEHGIJKL** |
| NULL | **ABCDFEHGIJKLM** |

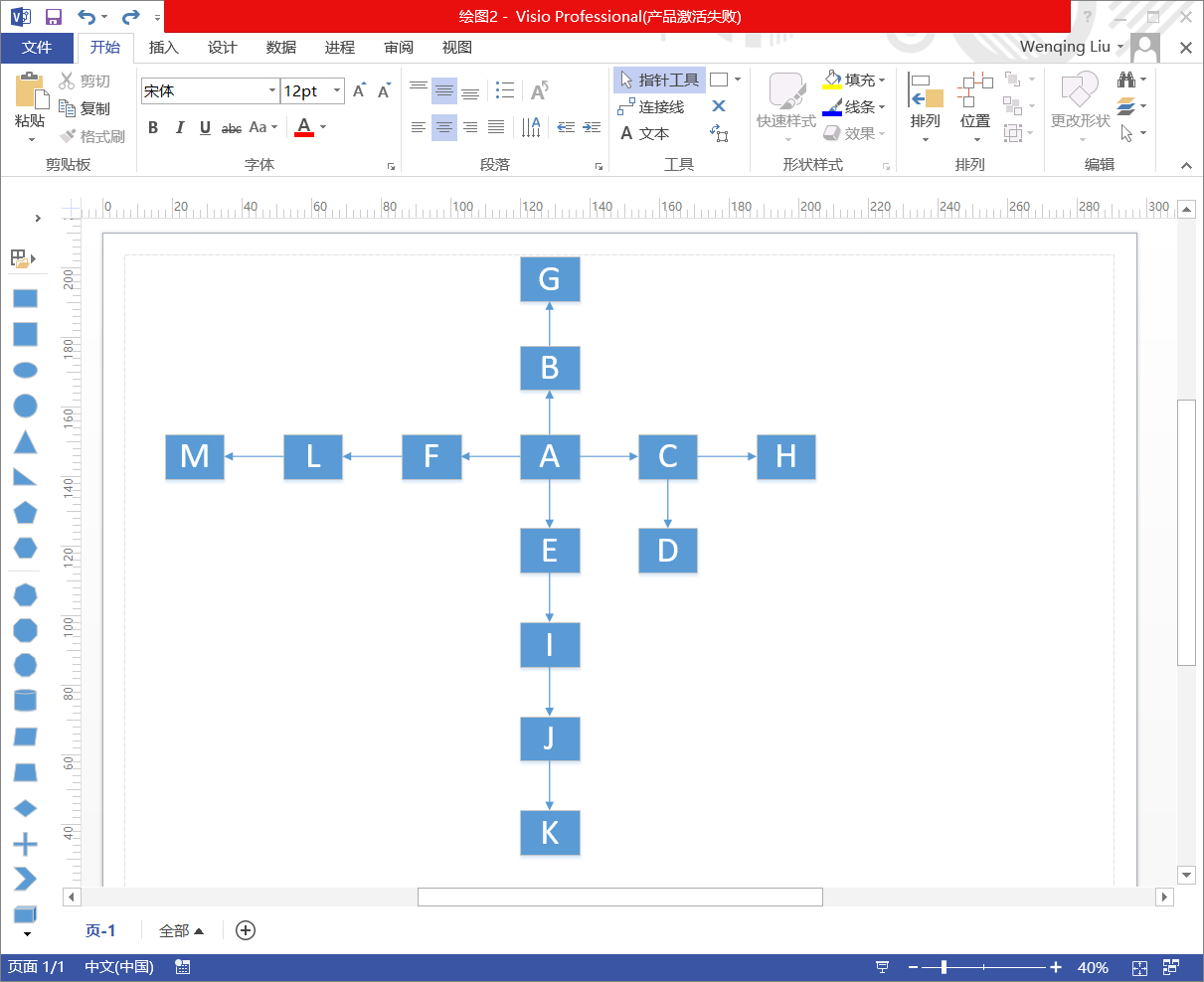


图4.3遍历图示例

### 4.3.3 数据清洗

数据清洗指将通过爬虫算法从网站的各个网页提取的HTML页面中我们需要的字段提取出来。图4.4是我们在百度学术网站上会抽取的一个网页，图4.5使我们使用爬虫抽取后得到的网页源码中的一部分，是一个HTML格式的文件，很显然并不是所有的字段都是有用的，只有其中一些网页标签中的字段是本文抽取的目标。



图4.4待抽取的网页示例



图4.5抽取的网页HTML源码

如图4.5，很显然，<div.sc\_content>标签下的<div.sc\_info>中的<a>标签中的字段使我们所需要的文章作者名字，他们可以作为图谱中的气象专家实体节点。

因此，数据清洗后得到的就是本文所需要的第一步数据，即未经过进一步处理的实体与概念 的集合。

### 4.3.4 知识抽取

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。

1. D2R映射

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图4.6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>气象论文</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>*

*<mappings>*

*<mapping>*

*<table>weather</table> <!--源数据表-->*

*<concept>气象专家</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>*

*<mappings>*

*<mapping>*

*<table>weather</table> <!--源数据表-->*

*<concept>气象机构</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>*

图4.6 D2RML映射示例

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。

1. 行业类网站抽取

行业类网站一般拥有自己独有的特点与结构，例如DrugBank、IMDB等。在该类型网站抽取知识时仅需要对其结构进行解析并编写相对应的包装器（wrapper）即可。

包装器的生成步骤分为：

1. 对Web页面进行预处理,去除网页噪声。使用基于DOM树的去噪算法来充分利用关键字,能够有效的保证Web页面中有效数据的完整性；
2. 对经过预处理和去噪后的Web页面,把它们的HTML源代码看作字符串,用Ukkonen算法构造其后缀树，目的是找到Web页面内的连续重复子串；
3. 利用后缀树来找到HTML源代码中的连续重复子串。由于Deep Web页面内的数据记录构成了连续重复子串,可以用嵌套模式来对Web页面进行建模。得到HTML页面内的连续重复子符串后,就可以进一步归纳出能够描述Web页面嵌套模式的正则表达式；
4. 生成Web页面嵌套模式的正则表达式,即包装器。

### 4.3.5 数据处理

在图谱构建之前需要对采集或抽取到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、关联关系学习与知识融合。

1. 实体及概念识别

实体识别指将上一步经过清洗的数据进行进一步处理，由于清洗后的得到的某些字段中包含了一些噪声信息，比如“这篇文章的作者是\*\*\*”，这里我们只需要作者的名字而其他的字符都是冗余信息。需要设计一些特定的正则表达式（比如上述示例只提取后三个字符）模式或定义一些启发式规则将我们需要的字符从字段中提取出来。

1. 实体合并

实体合并指对来自同一个数据源的相同实体进行合并，比如在百度学术上抽取学术论文时会抽取出两篇不同来源的相同学术论文。如图4.7。



图4.7同一来源相同实体示例

1. 关系抽取

实体对齐可以理解为同义关系抽取，因此实体对齐、上下位关系学习与关联关系学习可以统称为关系抽取。目前关系抽取比较常用的方式是基于支持向量机（SVM）、基于CRF以及基于深度学习神经网络的方法。不过本文不需要使用如此复杂的方法，这些方法在处理数量庞大而且概念种类繁多的情况下较为实用，而本文中的概念及其属性都已经知道，爬虫也是基于已知的关键词编写的。因此本文的关系抽取可以使用较为简单的底层代码进行实现。

由于本文的概念已知，不存在上下位关系，因此只需要进行实体对齐与关联关系抽取。而实体对齐是指将具有同义关系的实体进行合并，比如“荷花”与“菡萏”，但是本文中的实体是学术论文、气象专家、科研机构，这些实体之间一般不存在同义关系，所以本文不进行同义关系的抽取。

综上，本文主要进行的关系抽取是指关联关系的抽取，即实体或概念之间的关系，比如某一个气象专家发表了某些论文，他又是属于某一个研究机构的。这样就可以将实体连接起来，以一度关系或二度关系来表示它们之间的关系，有助于之后形成可视化的知识图谱。

1. 知识融合

知识融合是指将来自不同数据源的数据进行整合，主要指相同实体的合并。比如一篇来自百度学术的论文与一篇来自谷歌百科的论文有相同的论文题目，如图4.8与图4.9所示。则可以将他们进行合并，如果它们的信息有一些出入，比如论文的作者、来源等出现不同，则根据这些网站的置信度选择置信度较高的网站的论文。置信度如下：

谷歌学术>知网>百度学术>万方>维普



图4.8百度学术中检索到的论文



图4.9谷歌学术中检索到的论文

### 4.3.6 图谱构建

在互联网上经过爬虫提取并清洗的数据需要经过处理成为RDF三元组，而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：

1. 定义强对流方面气象知识图谱概念：本文定义强对流方面的气象知识图谱的相关概念，包括强对流方面专业论文、强对流方面气象专家、强对流方面研究机构；
2. 定义专家、科研机构、学术论文等概念的基本属性，比如气象专家的学术领域、就职单位、论文数量等；气象学术论文的出版方、被引量、来源等；气象机构的论文数量、所在地等等；
3. 添加实体节点；包括专家节点、科研机构节点、专业论文等。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。
4. 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。

### 4.3.7 知识图谱存储

本文方案中的知识图谱存储由四部分组成，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。

在这里，我们可以选择MangoDB与Neo4j这两个业内比较常用的数据库。同时基于Elastic Search构建气象图谱的分布式文件索引

1. **MangoDB**

MongoDB 是由C++语言编写的，是一个基于分布式文件存储的开源数据库系统。在高负载的情况下，添加更多的节点，可以保证服务器性能。

MongoDB 旨在为WEB应用提供可扩展的高性能数据存储解决方案。MongoDB 将数据存储为一个文档，数据结构由键值(key=>value)对组成。MongoDB文档类似于JSON对象，如图4.10所示。字段值可以包含其他文档，数组及文档数组。



图4.10 JSONObject示例

MangoDB特点：

1. MongoDB的提供了一个面向文档存储，操作起来比较简单和容易；
2. 你可以在MongoDB记录中设置任何属性的索引 (如：FirstName="Sameer",Address="8 Gandhi Road")来实现更快的排序；
3. 你可以通过本地或者网络创建数据镜像，这使得MongoDB有更强的扩展性；
4. 如果负载的增加（需要更多的存储空间和更强的处理能力），它可以分布在计算机网络中的其他节点上这就是所谓的分片；
5. MongoDB支持丰富的查询表达式。查询指令使用JSON形式的标记，可轻易查询文档中内嵌的对象及数组；
6. MongoDB使用update()命令可以实现替换完成的文档（数据）或者一些指定的数据字段；
7. MongoDB中的Map/Reduce主要是用来对数据进行批量处理和聚合操作；
8. Map和Reduce。Map函数调用emit(key,value)遍历集合中所有的记录，将key与value传给Reduce函数进行处理；
9. Map函数和Reduce函数是使用Javascript编写的，并可以通db.runCommand或mapreduce命令来执行MapReduce操作；
10. GridFS是MongoDB中的一个内置功能，可以用于存放大量小文件；
11. MongoDB允许在服务端执行脚本，可以用Javascript编写某个函数，直接在服务端执行，也可以把函数的定义存储在服务端，下次直接调用即可；
12. MongoDB支持各种编程语言:RUBY，PYTHON，JAVA，C++，PHP，C#等多种语言；
13. MongoDB安装简单。
14. **Neo4j**

现实中很多数据都是用图来表达的，比如社交网络中人与人的关系、地图数据、或是基因信息等等。RDBMS并不适合表达这类数据，而且由于海量数据的存在，让其显得捉襟见肘。NoSQL数据库的兴起，很好地解决了海量数据的存放问题，图数据库也是NoSQL的一个分支，相比于NoSQL中的其他分支，它很适合用来原生表达图结构的数据。

图4.11说明，相比于其他NoSQL，图数据库存放的数据规模有所下降，但是更能够表达复杂的数据。

通常来说，一个图数据库存储的结构就如同数据结构中的图，由顶点和边组成。

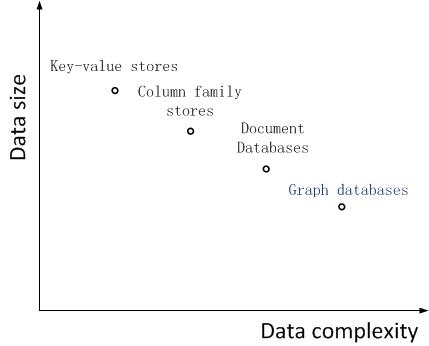


图4.11各NoSQL数据库数据量与数据复杂度趋势

接下来就从Neo4j的几个主要特性出发，结合代码对其逐一作出介绍。它们包括：数据模型、索引、事务、遍历和查询、以及图算法。

接下来本文都会围绕图4.12进行举例。

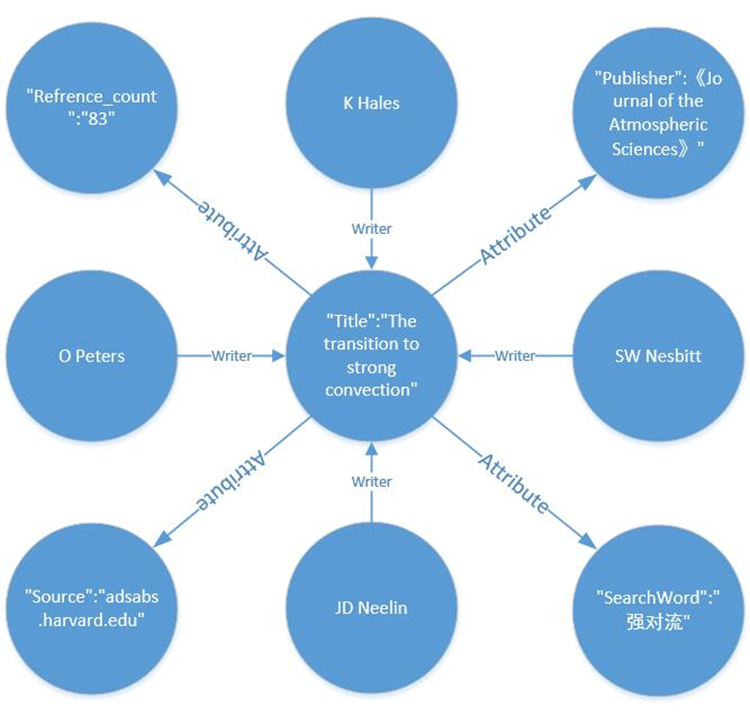


图4.12示例

1. 数据模型

Neo4j被称为property graph，除了顶点（Node）和边(Relationship，其包含一个类型)，还有一种重要的部分，属性。无论是顶点还是边，都可以有任意多的属性。属性的存放类似于一个hashmap，key为一个字符串，而value必须是Java基本类型或是基本类型数组，比如说String、int或者int[]都是合法的。

可以发现，图4.12中共有5个节点，论文题目与4个作者，他们之间的联系为“writer”，关系属性定义为“count”，即作者数量。论文实体有四个属性，分别为“SearchWord”、“Source”、“Publisher”、“Refrence\_count”。

要生成上图所示数据，首先可以定义所有边的类型。如图4.13。

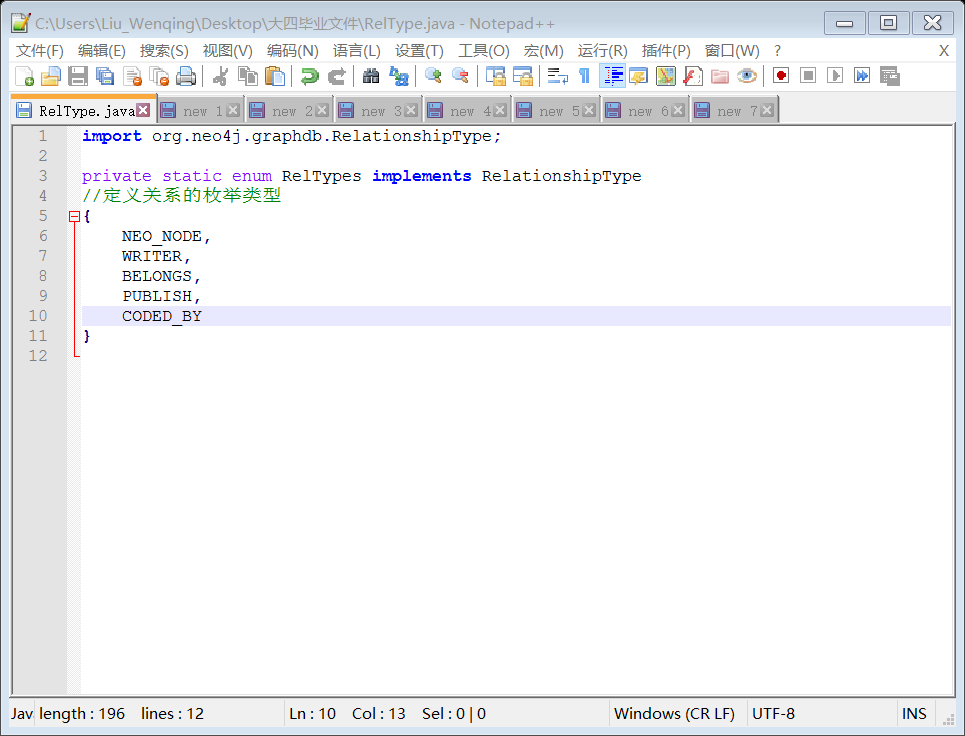


图4.13定义变得类型

接下来的代码创建了起始顶点和4个作者顶点，并创建了它们之间的边。其余数据生成类似。如图4.14.

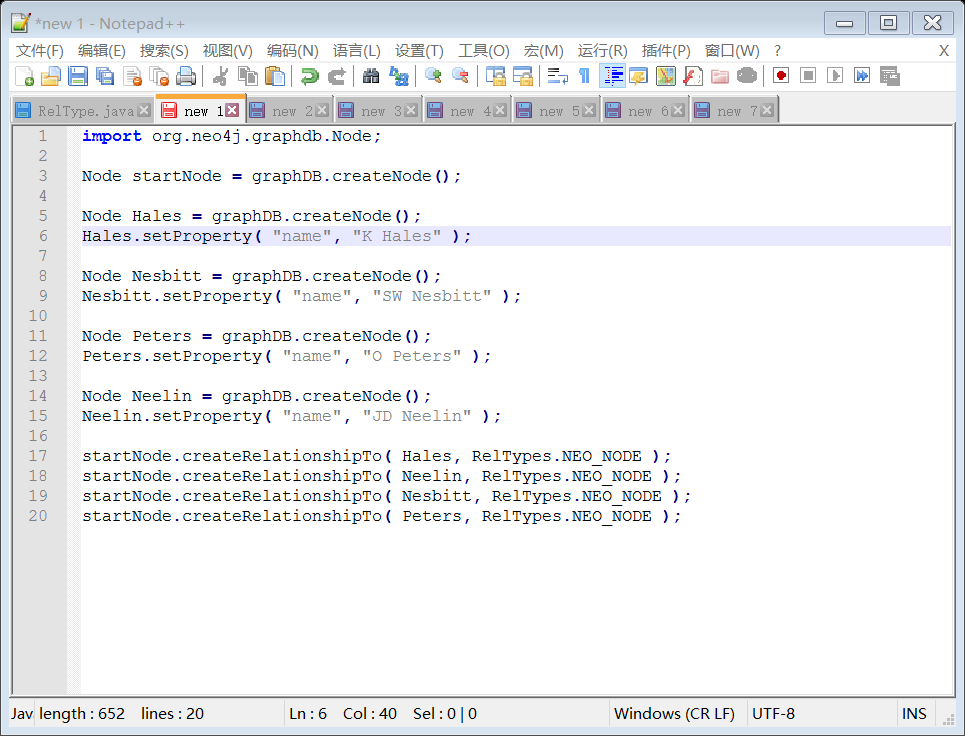


图4.14创建边

1. 索引

Neo4j支持索引，其内部实际上通过Lucene实现。现在可以创建一个索引叫nodes，来索引所有拥有name属性的顶点，这样我们就可以查询名字为“K Nales”的节点了。以下代码创建了这个索引，并索引了“K Nales”顶点，并通过名字得到了它。如图4.15.

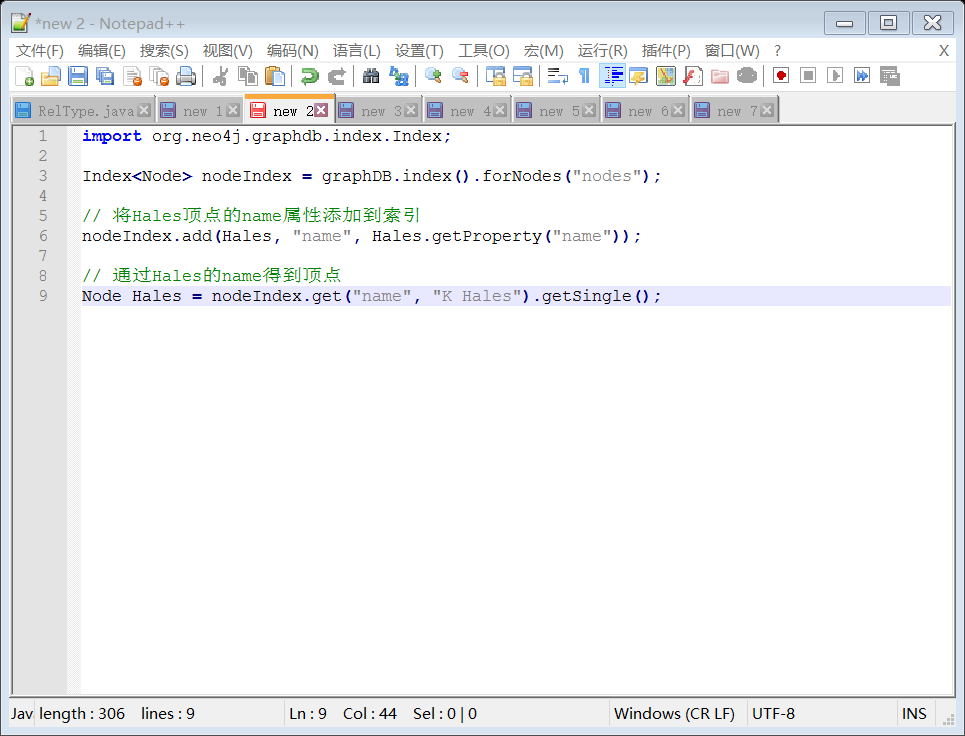


图4.15创建索引

1. 事务

Neo4j完整支持事务，即满足ACID性质。示例代码如图4.16：

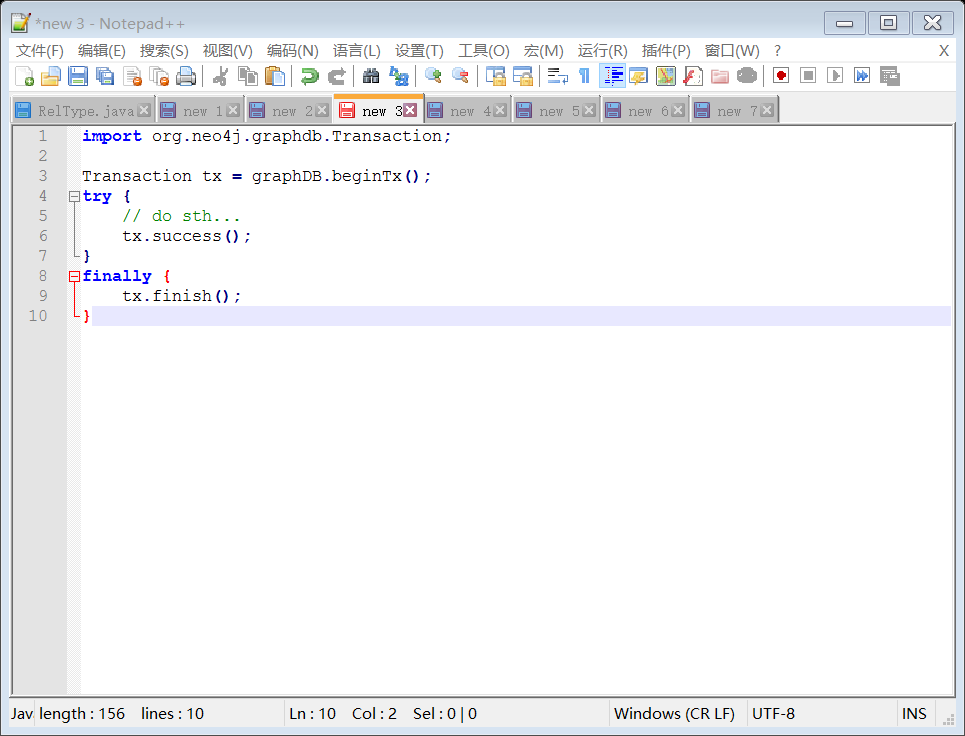


图4.16事务示例

1. 遍历和查询

遍历是图数据库中的主要查询方式，所以遍历是图数据中相当关键的一个概念。可以用两种方式来进行遍历查询：第一种是直接编写Java代码，使用Neo4j提供的traversal框架，如图4.17；第二种方式是使用Neo4j提供的描述型查询语言即Cypher。

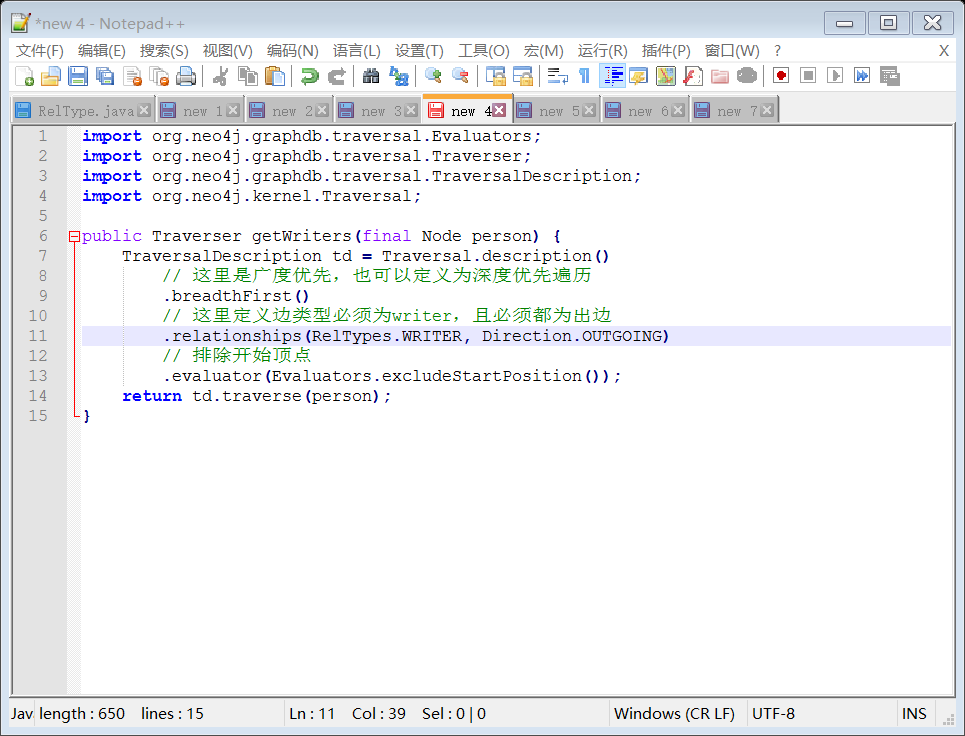


图4.17基于traversal的遍历算法

这里得到了Traverser对象后并没有立即执行遍历，而是在真正迭代结果时才进行延迟查询。以下的代码图4.18打印出一个顶点的所有朋友以及朋友的朋友。

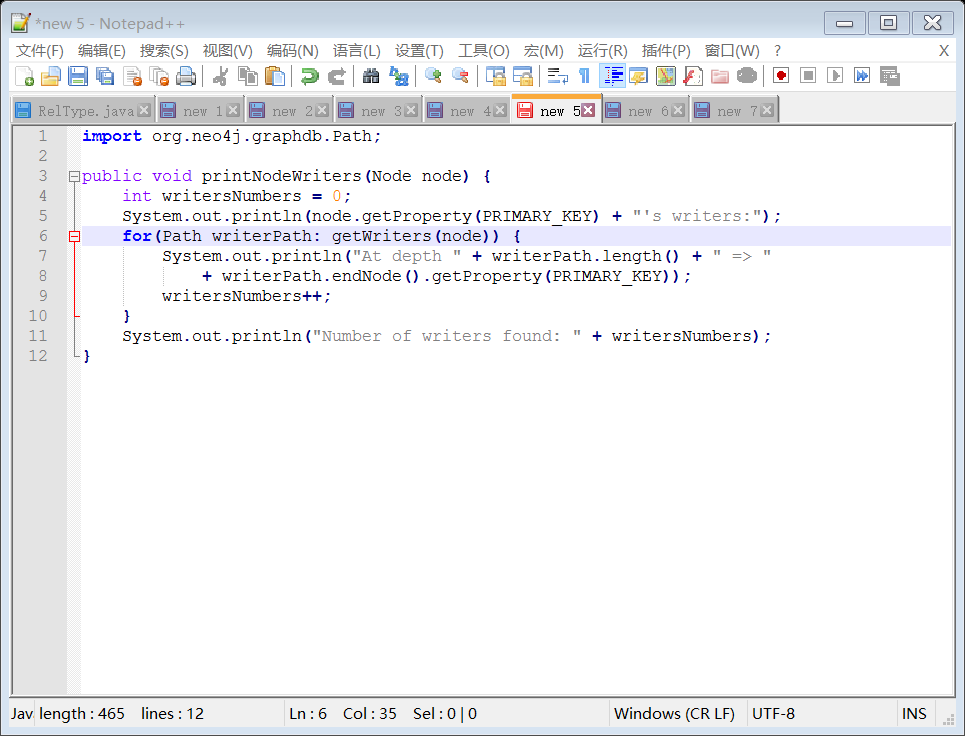


图4.18得到顶点的所有邻节点

第二种方法Cypher则直观得多，它是一种描述型的语言。图4.19是达到同样目标的Cypher命令。

简单做个说明。start说明从n节点开始，它通过查询nodes索引得到。match主要用来匹配图中顶点和边的关系，这里n和f顶点之前的relationship通过方括号表达，“:KNOWS”说明了边类型，“\*..”表示可以有任意多的边，如果只要求有两条，则是[:KNOWS\*2]。注意到这里的箭头表示是出边方向。return返回结果，distinct去除了重复访问到的顶点。可以看到还是很直观的。

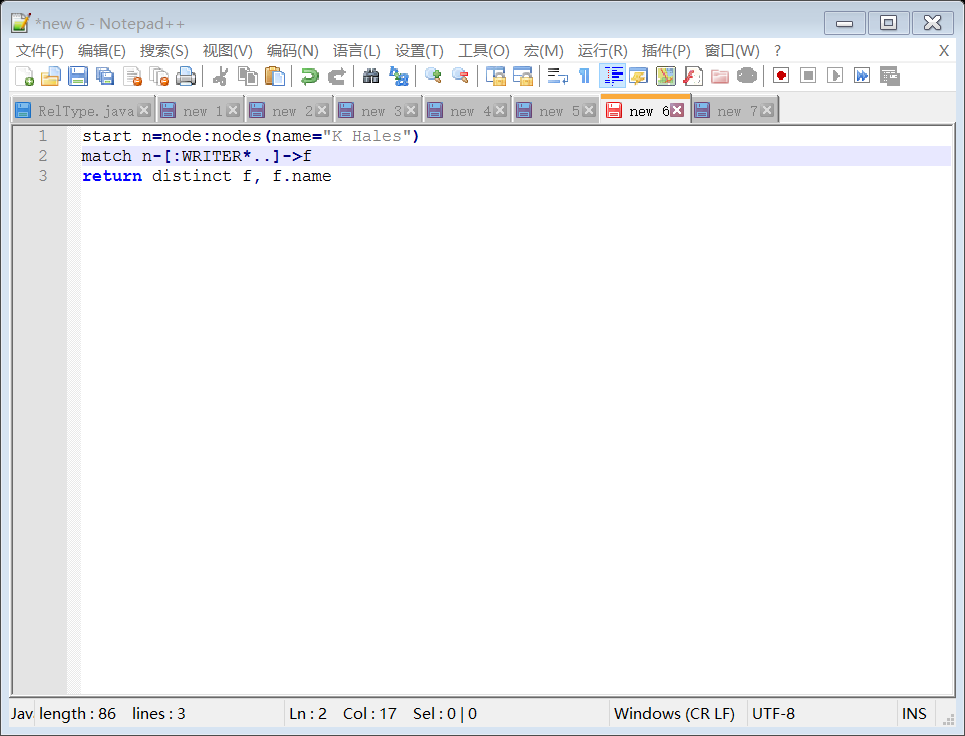


图4.19基于Cypher的遍历方法

1. 图算法

Neo4j实现的三种图算法：最短路径、Dijkstra算法以及A\*算法。图4.20是最短路径算法的简单例子。

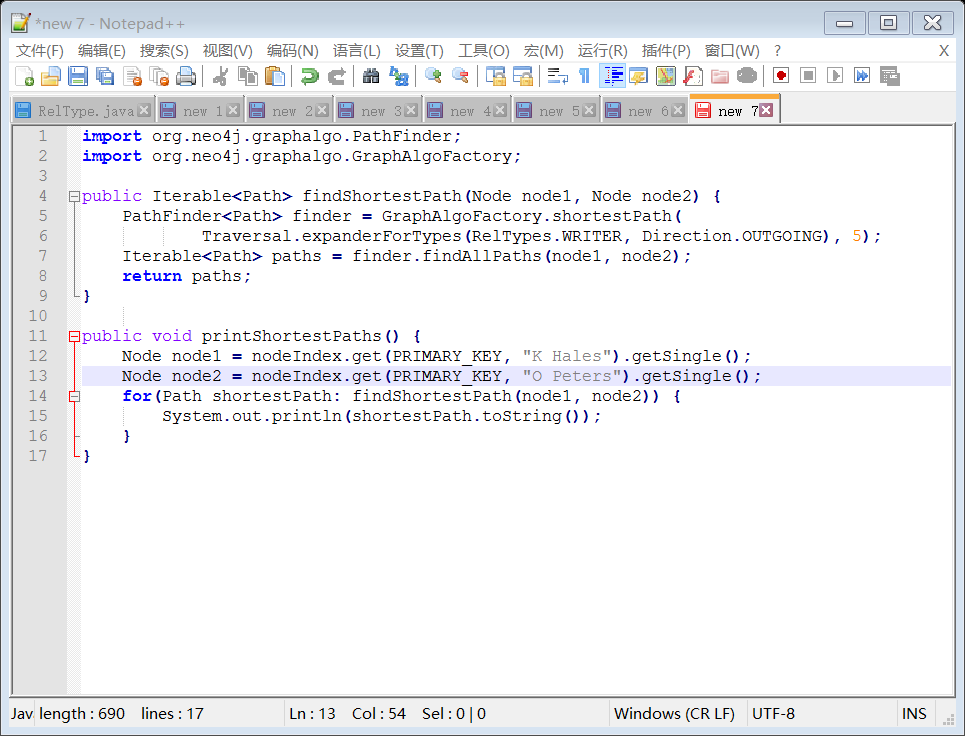


图4.20最短路径算法

1. **ElasticSearch**

Elasticsearch是一个基于lucene所编写的分布式的搜索引擎，能够达到实时搜索，稳定高效。

Elasticsearch有下面这些特点

1. 配置简单,对于初学者来说几乎是下载后零配置启动，易用性很高，即下即用。
2. 协议通用,通过http协议以json格式交互数据，而且还可以以加载外部插件的形式让其支持其他的协议。
3. 分布式索引，Elasticsearch把一个完整索引切成多个分片（数据是逻辑切分，不是单纯的文件块物理切分），分布到各个节点上去，构成分布式的检索。
4. 无中心化的访问方式，虽然一个Elasticsearch集群内部也有主从节点的区分，但是外部看集群是一个整体，和任意一个节点通信都是等价的。
5. 容错性的考虑，首先Elasticsearch集群可以为每个分片设置多个副本，以防某个分片损坏或者丢失，而且副本能分担查询压力。在有节点加入或退出时会根据机器的负载对索引分片进行重新分配，挂掉的节点重新启动时也会进行数据恢复。
6. 自动发现节点机制，集群首先通过广播方式找到当前网络上的其他节点，然后通过多播方式对所有节点进行通信，同时每个节点即是数据的接受者也是数据的提供者（P2P的思想）。

### 4.3.8 知识图谱应用

图谱的应用包括：

1. 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；

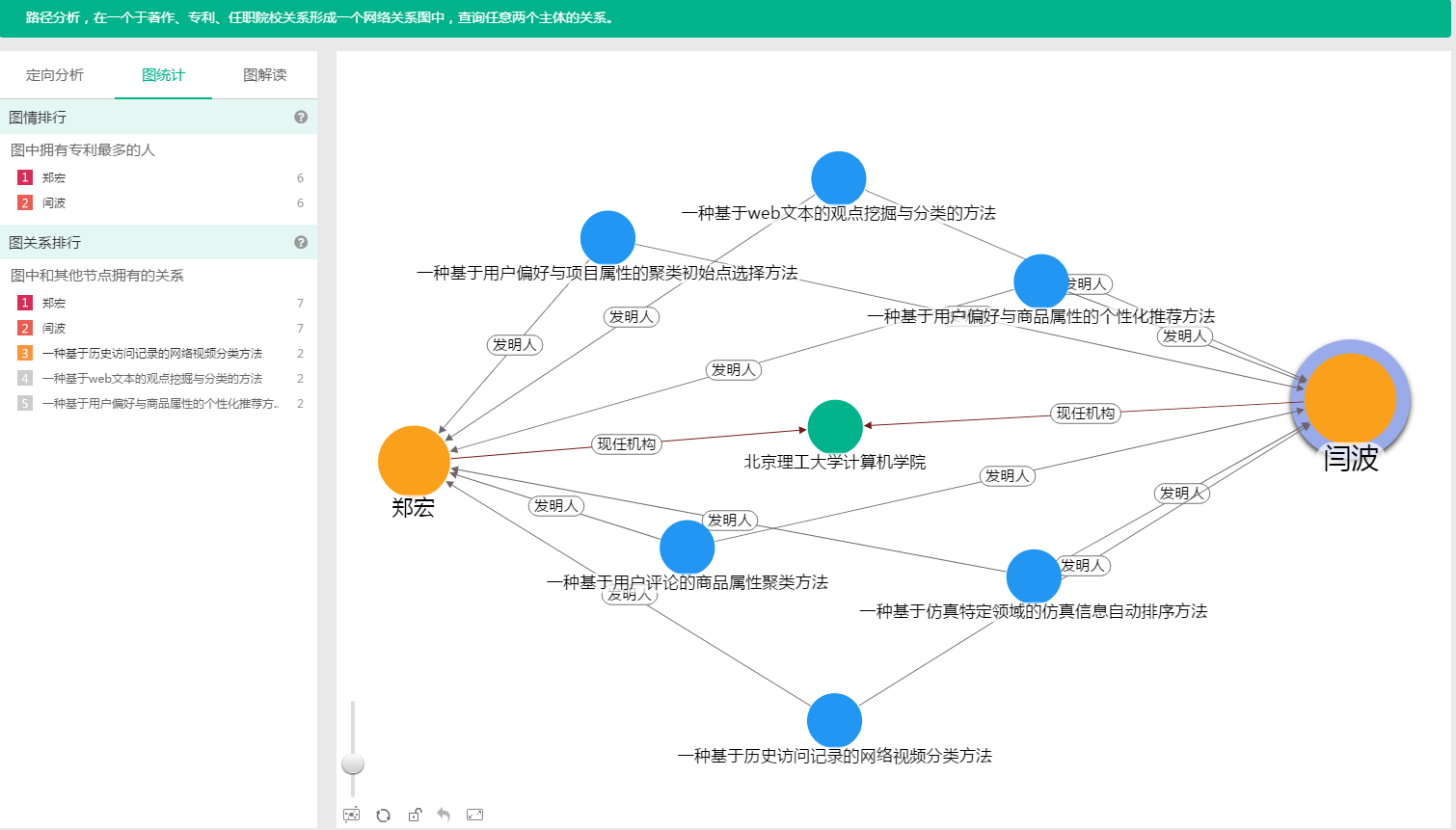


图4.21路径分析示例

1. 关联分析：发现多个实体与实体之间的关联关系；

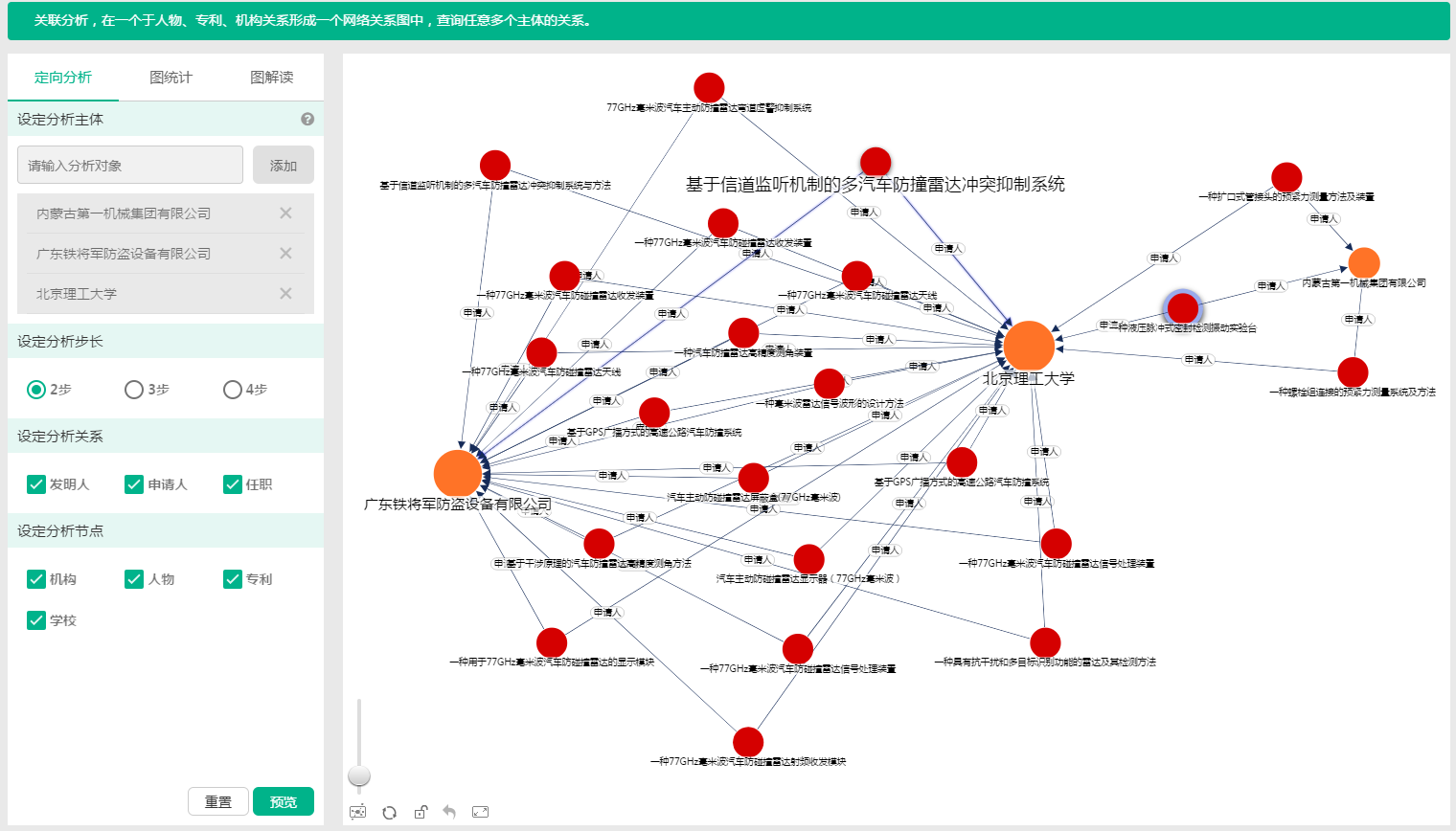


图4.22关联分析示例

1. 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；

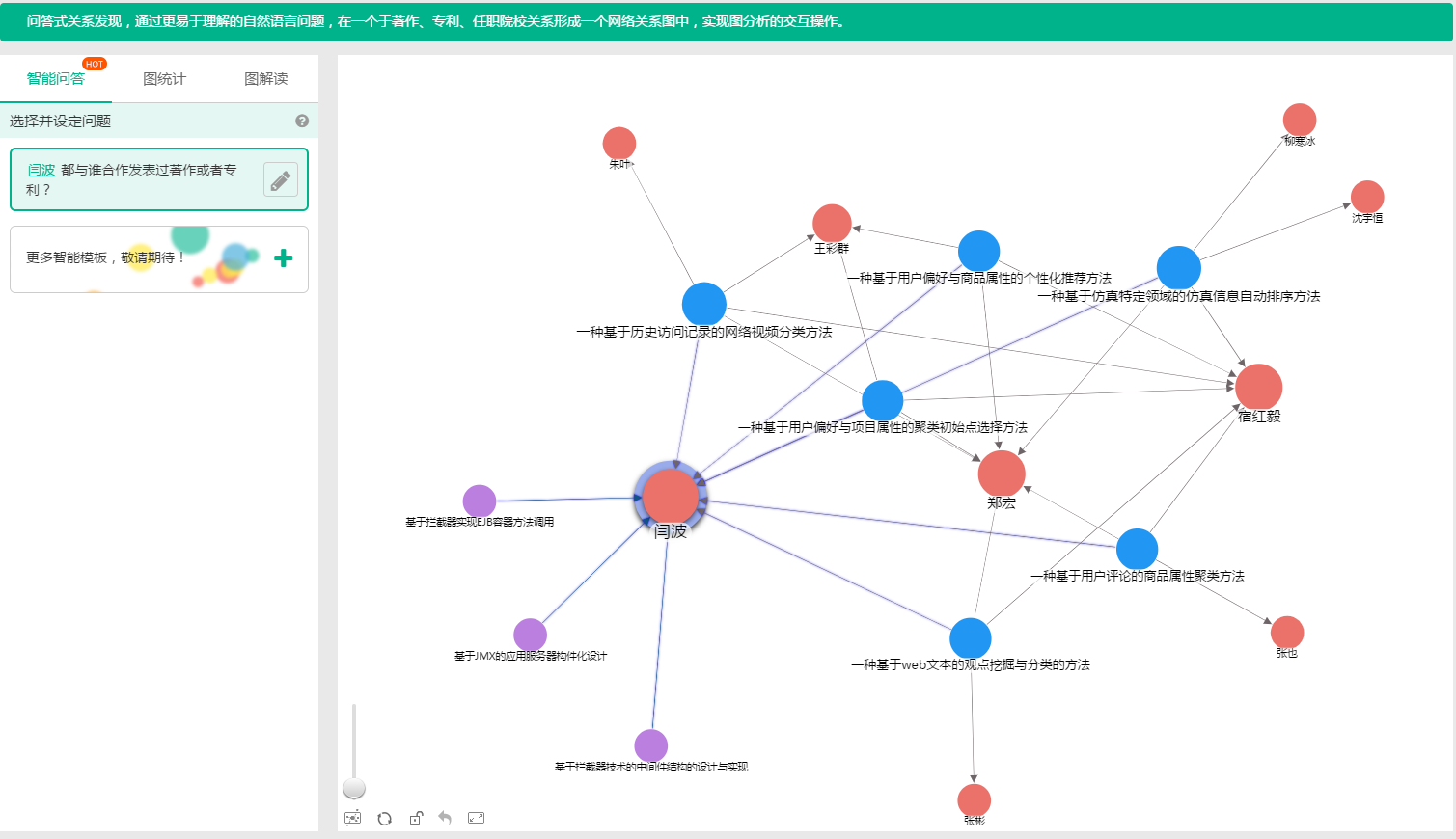


图4.23智能问答示例

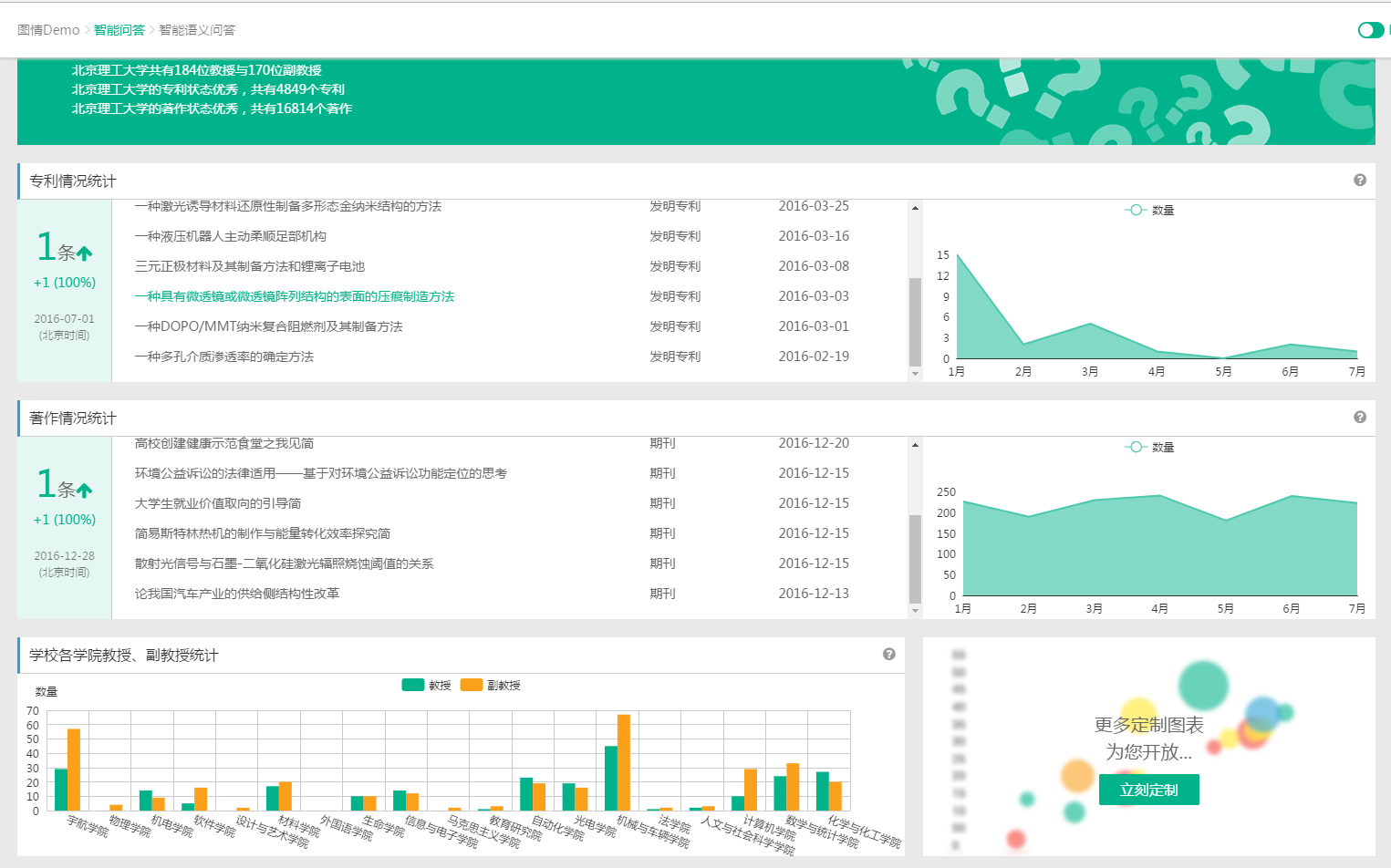


图4.24语义问答示例

1. 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4.25资源统计示例

# 5 方案实现与结果展示

## 5.1 数据源的选择

## 5.2 数据采集与处理

### 5.2.1 爬虫设计

### 5.2.1 数据处理

## 5.3 图谱存储

## 5.4 图谱可视化

# 6 结论和展望

## 6.1 结论

## 6.2 展望

# 附录

# 参考文献

1. 鲍建波. “e-航海(e—NAVIGATION)”概念的发展[J]. 中国海事, 2007(11):48-51.
2. 李向文. 新一代智慧港口[J]. 中国信息界-e制造, 2013(12):36-37.作者名. 题名[N].
3. Song J. DBAR Initiative: Big Earth Data for “Belt and Road” Development[J]. Bulletin of the Chinese Academy of Sciences, 2016(2).
4. Scharfenberg K. NOAA's Weather-Ready Nation: Progress and Plans[C]// AGU Fall Meeting. AGU Fall Meeting Abstracts, 2014.
5. 朱大奇. 人工神经网络研究现状及其展望[J]. 江南大学学报(自然科学版), 2004, 3(1):103-110.
6. 刘世长. 浅谈AIS系统在海事监管中的作用[J]. 中国水运月刊, 2010, 10(3):29-30.
7. 王翔, 徐照荣. 对海上数字广播系统(NAVDAT)技术的研究[J]. 航海, 2017(1):38-40.
8. 李耀东, 刘健文, 高守亭. 动力和能量参数在强对流天气预报中的应用研究[J]. 气象学报, 2004, 62(4):401-409.
9. 廖玉芳, 潘志祥, 郭庆. 基于单多普勒天气雷达产品的强对流天气预报预警方法[J]. 气象科学, 2006, 26(5):564-571.
10. 杨引明, 陶祖钰. 边界层风廓线雷达在强对流天气预报中的应用研究[C]// 中国气象学会2004年年会. 2004.
11. 漆桂林, 高桓, 吴天星. 知识图谱研究进展[J]. 情报工程, 2017, 3(1):004-025.
12. 陈文亮, 朱靖波, 姚天顺. 基于 Bootstrapping 的领域词汇自动获取. 第7届全国计算语言学联合学术会议 (JSCL 2003). 北京: 清华大学出版社, 2003: 67−72.
13. 郑家恒, 卢娇丽. 关键词抽取方法的研究. 计算机工程, 2005, 31(18): 194−196.
14. D. Milne, O. Medelyan, and I. H. Witten. Mining domain-specific thesauri from wikipedia: a case study. In: Proceedings of WI, 2006: 442-448
15. 陆勇, 章成志, 侯汉清. 基于百科资源的多策略中文同义词自动抽取研究. 中国图书馆学报, 2010(1): 56−62.
16. G. Cui, Q. Lu, W. Li, and Y, Chen. Mining Concepts from Wikipedia for Ontology Construction. Web Intelligence and Intelligent Agent Technologies. 2009(3): 287 - 290.
17. C. Zirn, V. Nastase, and M. Strube. Distinguishing between instances and classes in the wikipedia taxonomy. In: Proceedings of the 5th European semantic web conference on the semantic web, 2008: 376-387.
18. P. Cimiano, A. Hotho, and S. Staab. Comparing conceptual, divisive and agglomerative clustering for leaxning taxonomies from text. In: Proceedings of the European Conference on Artificial Intelligence (ECAI), 2004: 435-439.
19. 孙镇, 王惠临. 命名实体识别研究进展综述. 现代图书情报技术. 2010(6): 42-47.
20. B. Hachey, W. Radford, J. Nothman, M. Honnibal, and J. R. Curran. Evaluating Entity Linking with Wikipedia. Artificial Intelligence. 2013, 194: 130-150.
21. F. Castanedo. A Review of Data Fusion Techniques. The Scientific World Journal, 2013.
22. HowNet. <http://www.keenage.com/>
23. 梅家驹, 高蕴奇. 《同义词词林》 . 上海辞书出版社, 1983.
24. C. Fellbaum, WordNet: an electronic lexical database, MIT Press, London, 1998.S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: a nucleus for a web of open data. In: Proc. of the 6th Int. The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference, 2007: 722-735.
25. F.M. Suchanek, G. Kasneci, and G. Weikum. YAGO: a core of semantic knowledge unifying wordNet and wikipedia. In: Proceedings of the 16th International Conference on World Wide Web, 2007: 697–706.
26. Freebase. <https://www.freebase.com/>.
27. X. Niu, X. Sun, H. Wang, S. Rong, G. Qi, and Y. Yu. Zhishi.me: weaving Chinese linking open data. In: Proceedings of the 10th international conference on the semantic web, 2011, II: 205-220.
28. C. H. Chang, M. Kayed, M. R. Girgis, K. Shaalan. A survey of Web Information Extraction Systems. In: IEEE Transactions on Knoledge and Data Engineering, 2006: 1411--1428.
29. I. Muslea, S. Minton, and C. Knoblock. Hierarchical Wrapper Induction for Semistructured Information Sources. J. Autonomous Agents and Multi-Agent systems. 2001, 4: 93-114.
30. X.W. Ji, J.P. Zeng, S.Y. Shang, and C.R Wu. Tag Tree Template for Web Information and Schema Extraction. J. Expert Systems with Applications, 2010, 37, 8492-8498.
31. S.H. Lin and J.M. Ho. Discovering Informative Content Blocks from Web Documents. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2002.
32. 胡芳槐. 基于多种数据源的中文知识图谱构建方法研究[D]. 华东理工大学, 2015.
33. 刘知远, 孙茂松, 林衍凯,等. 知识表示学习研究进展[J]. 计算机研究与发展, 2016, 53(2):247-261.
34. 黄瑞. 基于KDD的知识自动获取及其应用[D]. 南京信息工程大学, 2005.
35. 杨劲松, 凌培亮. 一种模糊Petri网的逆向知识推理方法设计实现[J]. 计算机科学, 2009, 36(12):158-160.
36. 沈迅. 基于Web页面嵌套模式的包装器生成系统的设计与实现[D]. 北京邮电大学, 2010.

# 谢 辞

大学四年将要结束，毕业设计完成之后，我也无比真切地意识到，自己就要毕业了。

回想这四年，有太多人给予了我帮助，首先， 我必须感谢我的父母， 他们一直以来都支持着我的选择，从高考填志愿，到选择考研、读研究生，他们一直在经济和精神方面给予了我很大的鼓励和支持。

其次，我要感谢我的本科生导师李莉教授，李老师对我本次毕业设计的完成提供了巨大的帮助，从选题到遇到困难时不厌其烦的为我排忧解难。师恩如海，衔草难报。李老师丰富的专业知识、 严谨的科学态度、乐观开朗的精神都令我深深折服。

再其次，要感谢上海海翼知信息科技有限公司CEO丁军博士，他作为我本次毕业设计的校外导师在技术上给予了我巨大的帮助，从整个方案的设计到最后底层的实现都给了我很多良好的建议。作为一个创业公司的创始人，我在丁博士身上获益匪浅，他身上的精益求精、坚持不懈的精神深深影响了我。

然后，我要感谢上海海翼知科技有限公司的相关技术人员在本次毕业设计中给予我的巨大帮助，他们是王志宏、李春晓与李一斌，他们都是优秀的JAVA技术人员。非常感谢他们能在百忙之中抽出时间指导我在编程上遇到的困难。

接着，感谢各位自动化专业的任课老师， 在每堂课上我都学到了很多。无论我以后进入哪个行业，从事什么工作，我都无悔于自己高考结束后的选择。

还要感谢实验室的各位同学们，这几个月我们一起努力，互相学习，互相进步。感谢你们当我在做毕业设计迷茫的时候给予我最温暖的鼓励与支持。

接下来，还要感谢所有鼓励过我的人，批评过我的人，帮助过我的人，没有你们，就没有今日的我。回想往昔，我深深意识到自己的每一步都离不开他人的帮助。我也希望自己将来能帮助别人多一些， 做出自己的贡献， 让我们的社会变得更好。

最后，我也要感谢自己，感谢这四年来的坚持和努力，感谢自己在遇到困难时没有放弃，感谢自己始终为了自己的理想而奋斗，感谢自己曾经做出的那些决定和选择。

完成了这篇论文，我的大学生涯也将告一段落，而新的征程也已开始。我在以后的日子里将继续保持同济学子的优良作风，为母校争光，也不负我四年来的努力。