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。

自从语义网的概念提出以来，越来越多的开放链接数据和用户生成内容被发布于互联网中。在此背景下，知识图谱首先由 Google 提出，其目标在于描述真实世界中存在的各种实体和概念，以及这些实体、概念之间的关联关系。知识图谱在实体层面对本体进行了丰富和扩充；本体中突出和强调的是概念以及概念之间的关联关系，描述了知识图谱的数据模式；而知识图谱则是在本体的基础上，增加了更加丰富的关于实体的信息。知识图谱在语义搜索、智能问答、知识工程、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，形成气象方面的可视化知识图谱。

本文中我们将会从原始数据源中进行数据采集，然后进行数据清洗、实体对齐、知识融合等操作，最终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 and user generated contents are published on the Web after the Semantic Web has been proposed. Knowledge graph has been fist proposed by Google, which is focused on describing various real world entities and concepts, and their relations. Knowledge graph is a new vision of ontology; a knowledge graph extends an ontology in the entity level. Ontology usually focuses on concepts and their relations, which specifies the schema of knowledge graph; and knowledge graph adds a large number of entities into ontology. Knowledge graph is widely used in semantic search, intelligent question-answering, knowledge engineering, 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](#_Toc482162481)

[1.1强对流预警研究背景 1](#_Toc482162482)

[1.1.1国内外现状、水平和发展趋势 1](#_Toc482162483)

[1.1.2经济建设和社会发展需求 3](#_Toc482162484)

[1.1.3科学技术价值、特色和创新点 4](#_Toc482162485)

[1.1.4国内外强对流天气预警的研究成果 4](#_Toc482162486)

[1.2 知识图谱研究背景 5](#_Toc482162487)

[1.3 研究内容和意义 6](#_Toc482162488)

[1.4 研究现状和存在的问题 6](#_Toc482162489)

[1.5 本文的组织 7](#_Toc482162490)

[2 理论部分 8](#_Toc482162491)

[2.1 知识图谱构建过程 8](#_Toc482162492)

[2.1.2 本体学习 8](#_Toc482162493)

[2.1.2 实体层学习 9](#_Toc482162494)

[2.1.3 数据更新 9](#_Toc482162495)

[2.2 知识图谱的可用数据源分析 9](#_Toc482162496)

[2.2.1 数据源分析 9](#_Toc482162497)

[2.2.2 数据源采集 11](#_Toc482162498)

[2.2.3 数据清洗 11](#_Toc482162499)

[3 实验部分 13](#_Toc482162500)

[3.1 基本方案 13](#_Toc482162501)

[3.2 方案介绍 13](#_Toc482162502)

[3.2.1 原始数据源 13](#_Toc482162503)

[3.2.2 互联网数据采集与清洗 13](#_Toc482162504)

[3.2.3 知识抽取 13](#_Toc482162505)

[3.2.4 知识融合与图谱构建 13](#_Toc482162506)

[3.2.5 知识图谱存储与应用 13](#_Toc482162507)

[4 结果和讨论 14](#_Toc482162508)

[5 结论和展望 15](#_Toc482162509)

[5.1 结论 15](#_Toc482162510)

[5.2 展望 15](#_Toc482162511)

[参考文献 16](#_Toc482162512)

[谢 辞 18](#_Toc482162513)

# 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

知识图谱(Mapping Knowledge Domain）[11]也被称为科学知识图谱，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，用可视化技术描述知识资源及其载体，挖掘、分析、构建、绘制和显示知识及它们之间的相互联系。

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。

随着互联网中用户生成内容（User Generated Content， UGC）和开放链接数据（Linked Open Data，LOD）等大量 RDF（Resource Description Framework）数据被发布，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）。在此背景下，知识图谱的概念由谷歌于2012年正式提出，旨在实现更智能的搜索引擎，并于2013年开始在学术界与业界普遍普及，在智能问答、情报分析、反欺诈等应用中发挥重要作用。

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，语义网络有了一个新的应用场景，即语义Web。语义Web与传统Web的一个很大区别是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。

## [1.3 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。

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。

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 研究现状和存在的问题

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。

## 1.5 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基本过程进行了描述，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 理论部分

## 2.1 知识图谱构建过程

### 2.1.2 本体学习

本体的构建步骤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。

第一层次：术语抽取

术语是知识图谱中概念、实体或属性的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。

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。

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。

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。

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。

**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。

**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。

**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。

**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。

**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

**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。

**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。

**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。

**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。

**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。

**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

**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。

**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1。

|  |  |
| --- | --- |
|  | 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> |

表1 开放链接数据集的下载地址

**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。

**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 实验部分

## 3.1 基本方案

## 3.2 方案介绍

### 3.2.1 原始数据源

### 3.2.2 互联网数据采集与清洗

### 3.2.3 知识抽取

### 3.2.4 知识融合与图谱构建

### 3.2.5 知识图谱存储与应用

# 4 结果和讨论

# 5 结论和展望

## 5.1 结论

## 5.2 展望

# 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.

# 谢 辞