Computer Science and Application 计算机科学与应用, 2018, 8(3), 242-251
Published Online March 2018 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2018.83029

Hans汉斯

Knowledge Graph Construction from Meteorology Literature

Li Li, Meiji Cui

College of Electronics and Information Engineering, Tongji University, Shanghai
Email: lili@tongji.edu.cn, cui_mj@163.com

Received: Feb. 23rd, 2018; accepted: Mar. 12th, 2018; published: Mar. 19th, 2018

Abstract

The timeliness and accuracy of weather forecast and early warning are closely related to the safety of people's life and property. Meteorology literature published online, as an important part of open data, brings both challenges and opportunities for meteorological data analysis. Compared with numerical meteorological data, works on knowledge discovery from textual meteorological data are limited. Therefore, based on knowledge graph technique, the knowledge graph of the meteorology literature is constructed to realize the intelligent applications, such as path analysis, correlation analysis, visualization, statistical analysis.

Keywords

Meteorology Literature, Knowledge Graph, Knowledge Graph Construction

气象文献知识图谱构建

李　莉，崔美姬

同济大学，电子与信息工程学院，上海
Email: lili@tongji.edu.cn, cui_mj@163.com

收稿日期：2018年2月23日；录用日期：2018年3月12日；发布日期：2018年3月19日

摘　要

气象预报与预警的及时性和准确性，与人民生命财产安全息息相关。网上发表的气象资料是开放数据的重要部分，给基于数据的气象预报和预警带来了机遇与挑战。与数值气象数据相比，气象文本数据知识发现的研究非常有限。因此，本文利用知识图谱技术，构建了气象文献知识图谱，实现了气象知识图谱的智能应用，如路径分析、关联分析、可视化、统计分析。

文章引用: 李莉, 崔美姬. 气象文献知识图谱构建[J]. 计算机科学与应用, 2018, 8(3): 242-251.
DOI: 10.12677/csa.2018.83029

关键词

气象文献，知识图谱，知识图谱构建

1. 引言

近年来，随着 DBpedia [1]等链接开放数据源(Linked Open Data) [2]的出现，谷歌于 2012 年率先提出“知识图谱(Knowledge Graph)”的概念[3]。此后，国内知识图谱产品如百度“知心”和搜狗“知立方”等也相继发布，知识图谱获得了前所未有的关注。搜索引擎公司构建知识图谱的初衷是为了利用链接数据，进一步改善搜索质量，实现基于语义的搜索，使得搜索更加人性化。这类知识图谱主要强调知识的覆盖率，通常包含大量的常识性知识，称之为通用知识图谱。与之相对应的则是行业垂直知识图谱如中医药知识图谱[4]、军事知识图谱[5]和工业产品知识图谱[6]等，主要是利用行业数据构建特定领域的知识图谱，强调知识的深度，其目的是为行业的专业人员提供辅助支持。

据统计，2014 年上海港海域因天气影响施行通航管制共计 71 次，影响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)。2015 年，“东方之星”受下击暴流影响沉没；2016 年四川广元发生因强对流导致的翻船事件，共致 15 人死亡。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。这些事件无不显示着气象预报预警及时性与准确性的重要性。随着现代气象技术的迅猛发展，各种类型的气象数据呈现爆炸式增长趋势，如何有效查询和利用各类气象数据成为气象领域的难题。与数值气象数据相比，文本气象数据的研究非常有限。因此，本文针对文本气象数据，利用知识图谱技术将看似无关的气象数据关联起来，有效管理文本气象大数据，而且将之转换为语义化的知识，为气象行业的专业人员提供辅助决策支持。气象文献知识图谱能够为气象领域提供高效准确的查询和分析，如路径分析、关联分析、可视化、统计分析等。

本文详细介绍气象知识图谱的构建过程，及其在气象领域的智能应用。全文结构如下：第二章主要介绍知识图谱相关工作；第三章详细阐述气象文献知识图谱的构建过程；第四章给出气象文献知识图谱的智能化应用实例；最后讨论所构建的气象文献知识图谱的不足并对将来的工作做进一步展望。

2. 知识图谱相关工作

2012 年，谷歌最先提出知识图谱的概念，知识图谱作为一个知识库旨在提高其搜索引擎的能力[3]。知识图谱本质上是一种语义网，旨在描述真实世界中存在的各种实体或者概念，以及它们之间的关系[7]。其中，每个实体或者概念都有一个唯一的标识符，每个“属性——值”对用来描述实体的内在特性，而关系用来连接两个实体，刻画它们之间的关联。知识图谱因其可扩展性、易集成性、标准化、知识语义化等特性，受到了国内外学者们的广泛关注。尤其是大型的搜索引擎公司陆续建立了自己的通用知识图谱以及少量的行业垂直知识图谱，推动了知识图谱构建的研究。

2.1. 通用知识图谱

通用知识图谱主要面向通用领域，其中包含大量的常识性知识，主要是结构化的百科知识，强调知

识的覆盖面，主要服务对象为普通用户。

最初的通用知识图谱大多由人工手动构建而成，如 Word Net [7]。由于人工构建费时费力且构建规模有限，这种知识图谱构建方式逐渐被取代。此后，通用知识图谱的构建逐渐转向利用现有的结构化信息，如 DBpedia [8]从维基百科中抽取了信息框中的所有信息和统计信息，这种方式提高了知识图谱的知识覆盖面，但是无法保证其正确性。YAGO [9]则仅仅从维基百科中抽取出其自定义的属性，从而提高了准确性，但是由于自定义的属性量有限，牺牲了知识图谱的广度。近年来，知识图谱构建大多基于网页数据，使用增量迭代的方式自动学习得到高质量的三元组，这种开放式的知识抽取方法得到了广泛的关注，如 Know It All [10]、NELL [11]。

随着英文知识图谱运动的开展，中文知识图谱项目也相继展开。zhishi.me [12]和 SSCO [13]等中文知识图谱项目从互动百科、百度百科和中国维基百科等中文百科网站抽取出结构化的中文信息。Xu 等[14]提出 CN-DBpedia，能够连续不断地自动生成和更新知识库，是一种永不停歇的中文知识抽取系统。与英文知识图谱不同，中文知识图谱在构建的过程中存在着一下几点特殊性：① 中文的开放链接数据和开放知识库相对缺乏；② 中文的在线百科不如英文维基百科丰富；③ 缺乏完整的词典库，无法用于知识图谱构建过程中的机器学习起步和评估学习的结果；④ 中文与英文的语言特性不同，适用于英文的文本抽取和学习的方法在中文中效果不佳。因此，想要构建有影响力的中文知识图谱可从以上四个方面突破。

2.2. 行业垂直知识图谱

与通用知识图谱相对应，行业垂直知识图谱是面向某一特定领域，其主要是基于行业的内部数据和行业知识库，强调的是知识的深度和准确性，其潜在客户是行业人员。

Yu 等[15]为了解决中医中存在的“知识孤岛”问题，构建了中医预防保健知识图谱，整合了专业术语、文献、数据库以及其他一些知识资源，能够实现知识可视化、知识检索、知识推荐等功能，有助于中医知识的共享、解释和使用。

程文亮[16]构建了商业领域的企业知识图谱，从上市公司的公报数据中抽取了企业之间的商业往来，并从新闻文本数据中抽取了企业所发生的重大事件和事件摘要，并且对比了知识图谱和社交网络在图结构上的统计特征。

葛斌等[5]构建了军事领域的知识图谱，能够整合大量分散孤立的情报，使得计算机理解情报语义，并在语义层面将大量情报关联起来，从而挖掘出关联情报中的新知识。

目前，行业垂直知识图谱得到了各行各业的关注，但是还没有专门针对气象领域的知识图谱。气象领域数据量大且数据丰富，尤其是网络上的大量文本气象数据没有得到充分的利用。因此，本文利用文本气象数据，构建气象领域的文献知识图谱，为气象文本数据的分析与应用提供支持。

3. 气象文献知识图谱的构建

3.1. 气象文献知识图谱应用框架

气象文献知识图谱主要基于文献网站以及新闻网站的气象文本数据，如维普、万方以及百度新闻，利用知识图谱技术对气象文本数据进行管理和知识抽取，最终构建的气象文献知识图谱能够实现一些智能应用，包括文本数据的路径分析、关联分析、可视化和统计分析等。气象文献知识图谱应用框架如图 1 所示。

3.2. 气象文献知识图谱构建

3.2.1. 气象文献知识图谱构建基本方案

气象文献知识图谱的应用方案基本框架如图 2 所示，主要分为 6 层结构。

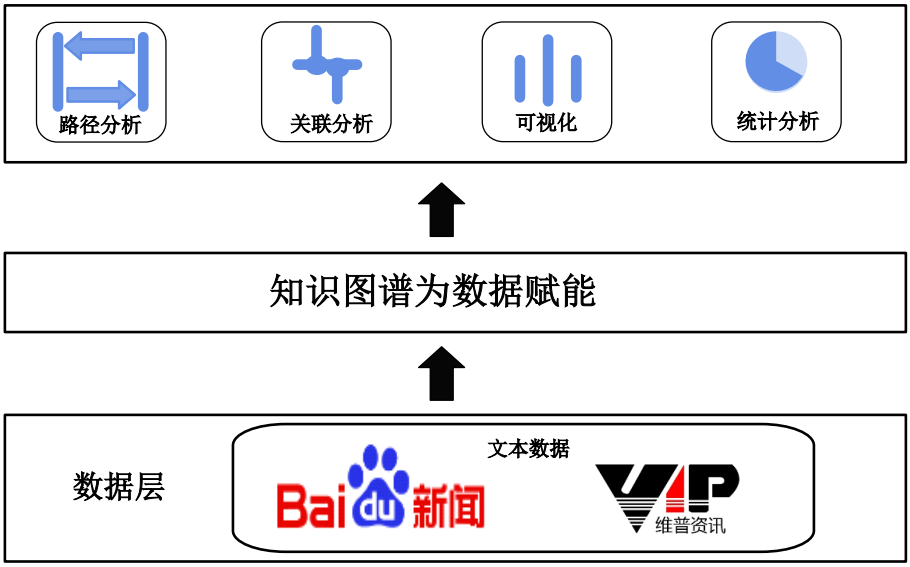

Figure 1. The application framework of meteorology literature knowledge graph
图 1. 气象文献知识图谱应用框架图

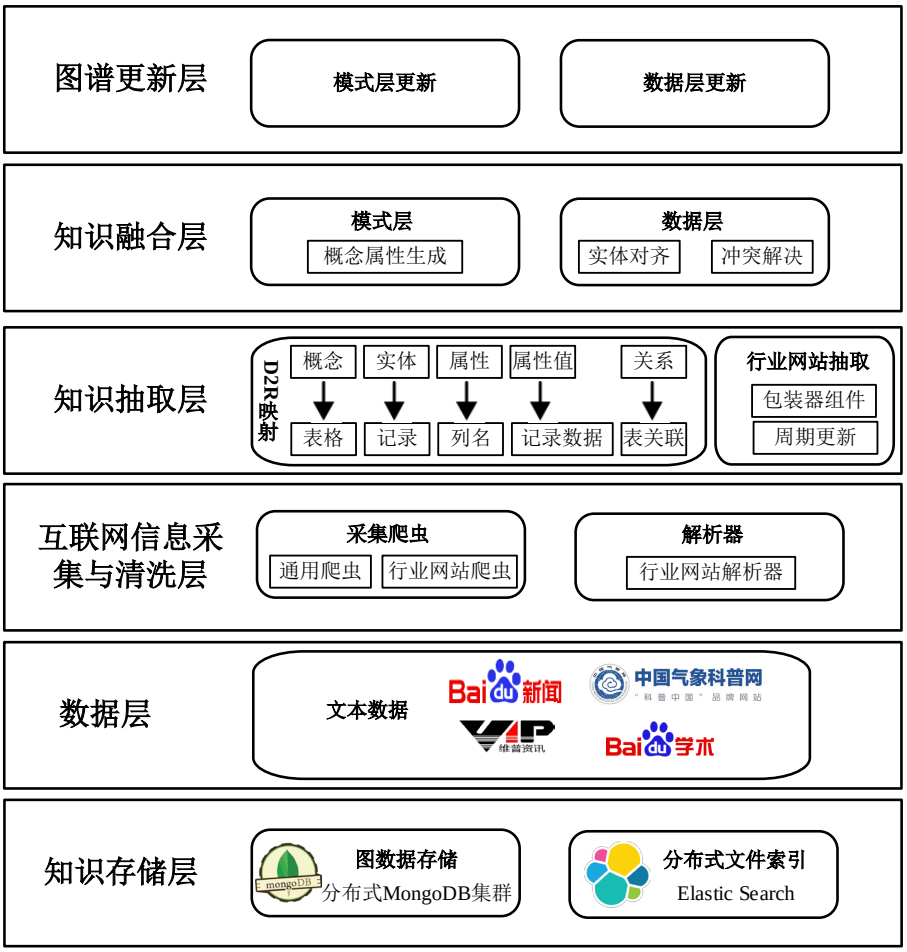

Figure 2. The framework of meteorology literature knowledge graph
图 2. 气象文献知识图谱框架图

原始数据层：为了构建气象文献知识图谱，原始数据主要包括气象行业网站、百科、论文资源、新闻资讯等互联网文本数据，主要来源网站是气象科普、百度百科、维普资讯、万方以及百度新闻等。

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：知识抽取层主要是指 D2R (Relational Data base to RDF, 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模式的文件映射(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包装器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：知识融合层包括模式融合和数据融合。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，包括不同数据源造成的冲突和不同构建方法造成的冲突，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 Mongo DB 的大规模三元组知识存储为所构建的气象知识图谱提供百亿级别以上三元组知识的存储与更新，保证知识图谱的流式处理流程中的效率；一方面支持底层数据经过知识图谱学习过程不断对知识图谱进行补充和更新，另一方面为上层高速的数据访问提供支撑。同时提供基于 Elastic Search 构建气象图谱的分布式文件索引。

图谱更新层：图谱更新层主要包括数据层更新和模式层更新。数据层的更新是指实体数据的更新，包括实体的添加和删减，修改实体的属性值等；模式层更新是指知识图谱本体中元素的变更，包括概念的增加、修改和删除，以及概念属性的更新[17]。

3.2.2. 气象文献知识图谱构建具体过程

在构建气象知识图谱之前，在气象领域专家的帮助下定义模式层，即定义知识图谱的图结构。在此，为了叙述方便，给出一个简单的图结构定义，包括实体类型、关系类型和实体属性，如表 1，表 2 所示。

Table 1. Definition of entity types, relationships
表 1. 实体类型、实体关系定义

实体类型	关系类型
article	article_author
keyword	article_institution
institution	article_keyword
publisher	article_publisher
author	author_institution

Table 2. Definition of entity attributes
表 2. 实体属性定义

属性	
机构属性	attr_AchieveNum_institution, attr_CitedNum_institution, attr_H-Index_institution
论文属性	attr_author, attr_institution, attr_keyword, attr_publisher, attr_Refcount, attr_year
期刊属性	attr_AchieveNum_publisher, attr_CitedNum_publisher, attr_H-Index_publisher
作者属性	attr_AchieveNum_author, attr_CitedNum_author, attr_H-Index_author

结合气象文献知识图谱，实体类型包括五类，分别是文章，关键词，机构，发表单位，作者，每个实体分配唯一的标识符；关系类型也包括五类，分别是文章—作者、文章—机构、文章—关键词、文章—发表单位、作者—单位；实体属性包括四类，分别是机构属性、文章属性、期刊(发表单位)属性以及作者属性。气象文献知识图谱的模式可以随时更改，只需修改相应的概念或者属性即可。

上海市气象局专家一共提供了 5 大类气象关键词，分别是气象现象关键词、气象预报关键词、气象预警关键词、气象监测关键词以及气象服务关键词。为了爬取到更加全面的气象文献数据，增加了相应的英文关键词，表 3 中给出了部分中英文关键词。

数据采集主要包括采集气象相关论文、气象领域专家、气象领域科研机构、气象新闻事件等，采集来源包括新闻类网站和学术期刊类网站等，具体说明如下表 4 所示。

根据上述关键词以及相关网站，编写网络爬虫，爬取气象文本数据，其流程如图 3 所示。

步骤一：初始化网址(URL)队列、未访问表(Unvisited)和访问表(Visited)，未访问表和访问表中装在网址信息；

步骤二：判断 Unvisited 是否为空或者是否爬取到了足够的网页文本数据，如果是则退出程序；

步骤三：Unvisited 队头网址出队列，判断 URL 是否是所需的 URL，如果是则下载 URL 指定的页面；如果不是则丢弃，返回步骤二；

步骤四：抽取出下载网页中的 URL，判断 Unvisited 和 Visited 中是否包含该 URL，如果没有，则该 URL 入队列，返回步骤二；如果有则丢弃，返回步骤二；

将爬取到的结构化文本数据映射到气象知识图谱模式，比如表对应概念，记录对应实体，记录数据对应属性值等，此种方法称之为 D2R 映射。D2R 是将关系型数据库中的数据转化为 RDF 三元组形式的语义数据。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 D2R 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，然后把这些映射得到的知识存储到气象知识图谱中。在添加实体节点时，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。

Table 3. Meteorology keywords (part)
表 3. 气象关键词(部分)

关键词分类	关键词中英文对照
气象现象	副热带高压(subtropical high)、雷暴(thunderstorm)等
气象预报	短临预报(short impending forecast)、数值预报(numerical forecast)等
气象预警	强对流天气风险预警(risk warning of severe convective weather)等
气象监测	多普勒雷达(Doppler radar)、双偏振雷达(dual polarization radar)等
气象服务	气象数据开放(open meteorological data)等

Table 4. Internet data acquisition website (part)
表 4. 互联网数据采集网站(部分)

数据来源	采集地址
维普	http://qikan.cqvip.com/
百度学术	http://xueshu.baidu.com/
百度新闻	http://news.baidu.com/

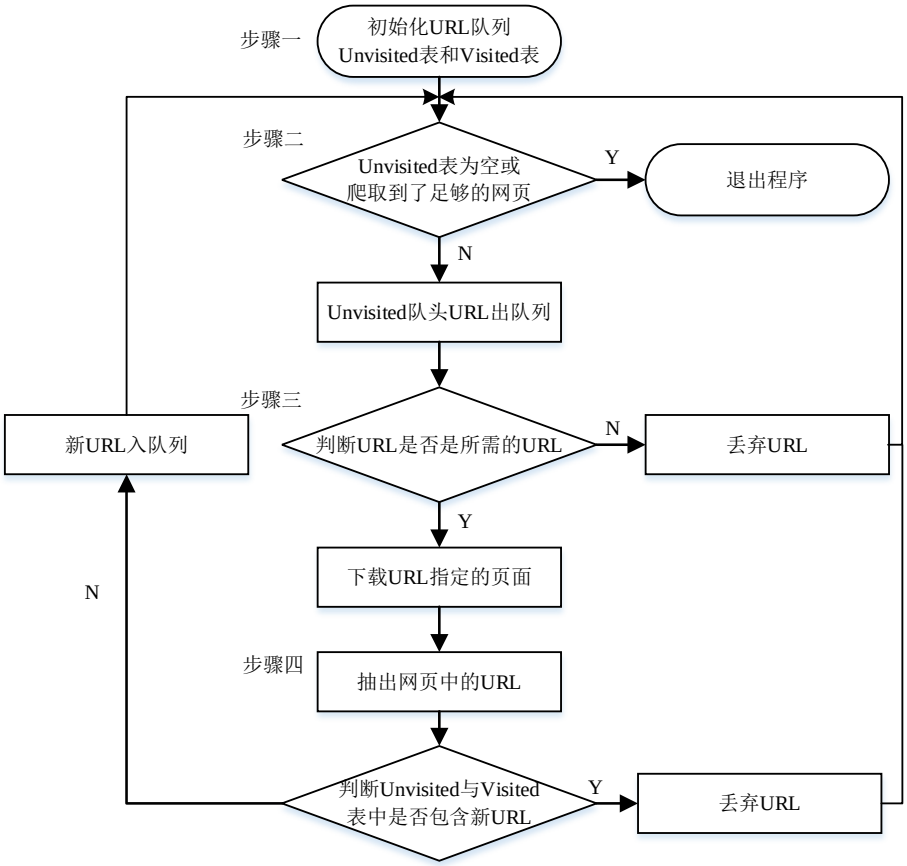

Figure 3. The workflow of crawler
图 3. 爬虫流程图

上述流程基本上完成了气象文献知识图谱的构建，当数据出现冲突时，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。

4. 气象文献知识图谱的应用

气象文献知识图谱能够很好地管理和利用海量的气象文本数据，为气象专业人员提供知识点之间的关联信息，从而抽取出新的气象知识，最终为气象的预报和预警提供服务。其主要应用场景有以下几点。

1) 网络关系发现

网络关系发现是指以一个中心实体为视角，以图谱可视化的方式查看与其存在关联关系的人物，机构，资源，技术点等。在气象文献知识图谱中，以“上海海洋中心气象台”为中心实体，会出现以与此存在关联关系的其他实体，形成一个巨大的有向图。为了方便分析，选取了与中心实体相关的部分实体，如图 4 所示。在气象文献知识图谱中，每一个节点表示一个实体，实体与实体之间用有向边表示。从图 4 中可以看出，上海海洋中心气象台发表过“基于云雷达的大气 0℃层亮带识别”和“太阳光度计和微波辐射计资料在盛夏午后强对流预报中的应用”，且都发表在《干旱气象》杂志上，关键词都为强对流。当文本关系趋于复杂时，知识图谱能够更加高效地进行文本数据的分析。

2) 路径发现

路径发现是指在基于实体形成的一个网络关系图中，查询任意两个或多个实体的最短路径。如图 5 所示，查询“上海海洋中心气象台”与“浙江省舟山市气象台”之间的最短路径，可以看出，上海海洋中心

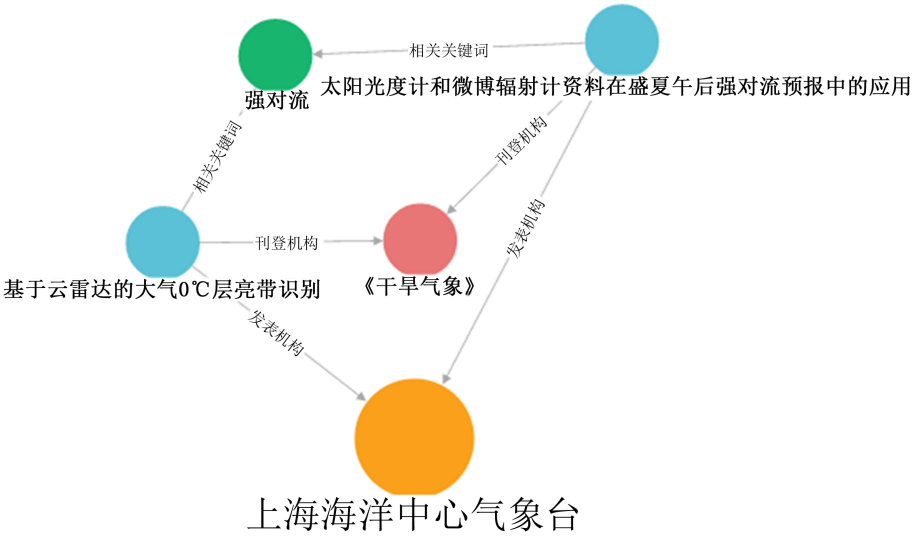

Figure 4. The diagram of network relationship discovery
图 4. 网络关系发现示意图

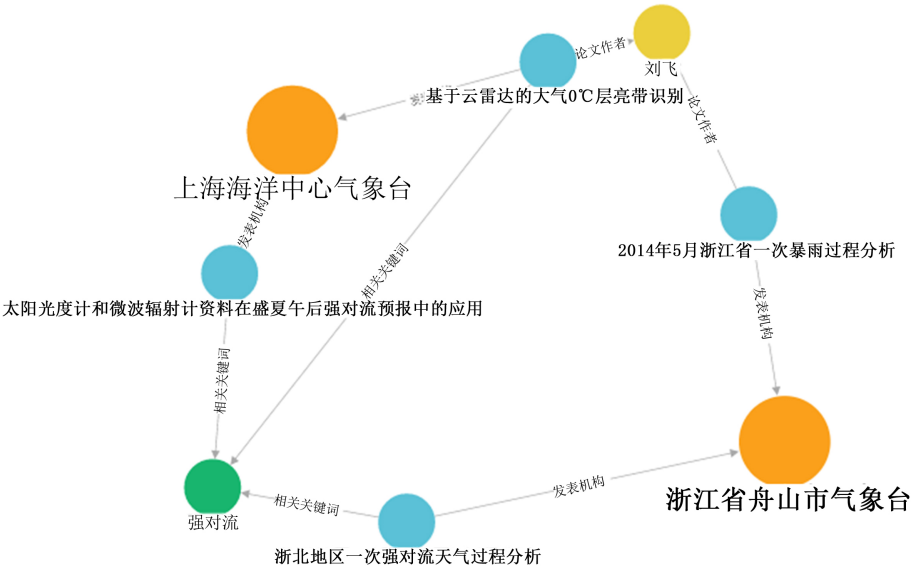

Figure 5. The diagram of path discovery between two entities
图 5. 两个实体间路径发现示意图

气象台和浙江舟山气象台都对强对流有深入的研究，两个气象台今后可在这一领域进一步合作。当路径关系越复杂时，越能体现出知识图谱管理文本数据的能力。

3) 资源统计分析

资源统计分析是对所有爬取到的气象文献资料进行可视化的统计分析。目前，一共统计分析了 22,223 篇新闻资源和 44,695 篇论文资源。从图 6 中可以看出，近年来与气象相关的文献和新闻数量呈上升趋势，说明气象是一个很有前景的研究领域。文本数据统计截止时间为 2017 年 6 月，所以在 2017 年文本数量稍有下降趋势。从图 7 中可以看出，在众多气象关键词中，暴雨和冰雹为最常见的气象灾害，也是学者研究最多的气象现象。图 8 展示了当前气象领域发文数最多的机构以及学者，通过这样的统计，可以加强不同气象机构的学者之间的交流合作，促进气象领域的进步。

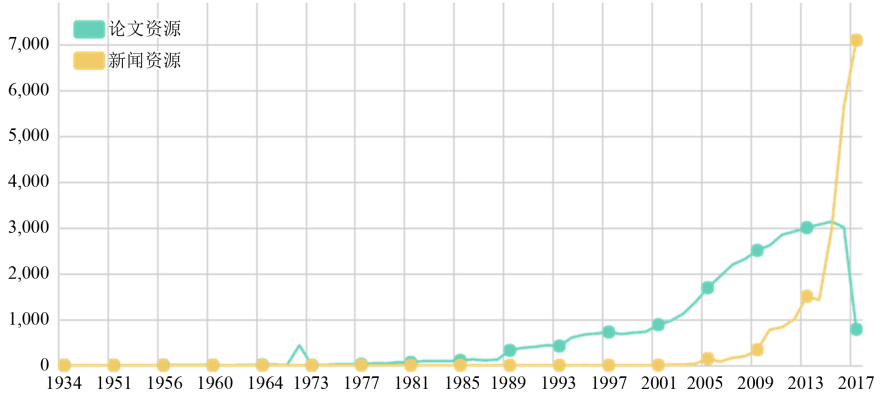

Figure 6. The proportion of resources and the trend of resource attention
图 6. 文献资源占比以及资源关注度趋势

Figure 7. Literature hot keywords
图 7. 文献热点关键词

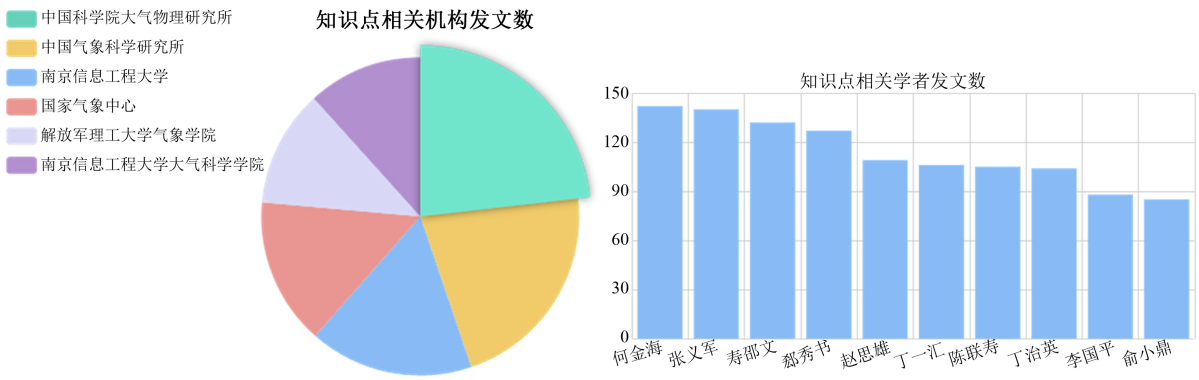

Figure 8. Comparative chart of the number of papers issued by institutions and scholars
图 8. 气象机构和学者发文数对比图

5. 总结

本文针对分散的网络气象文本数据，编写爬虫从网络上爬取相关文本数据，在此基础上构建了气象文献知识图谱。气象文献知识图谱能够更加高效地管理在线文本数据，并且实现了气象知识图谱的智能应用，如路径分析、关联分析、可视化、统计分析。这些应用能够帮助气象领域专业人员更加快捷地发现气象知识的关联，提取出新的气象知识，最终达到辅助气象预报和预警的作用。

现有的气象文献知识图谱规模相对较小，应用场景有限，因此本文下一步工作主要包括两方面：① 进一步完善气象文献知识图谱，提高气象知识图谱的覆盖面。② 扩展气象文献知识图谱的应用，比如基于

气象文献知识图谱，开发一个气象问答系统，辅助气象专业人员决策。

基金项目

本研究获得上海市科学技术项目(No. 16dz1206102)，国家自然科学基金(No. 51475334)，中央高校基本科研业务费专项资金(No. 22120170077)资助。。

参考文献

[1] Lehmann, J., Isele, R., Jakob, M., *et al.* (2015) DBpedia—A Large-Scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web*, **6**, 167-195.

[2] Bizer, C., Heath, T. and Berners-Lee, T. (2009) Linked Data—The Story So Far. *International Journal on Semantic Web and Information Systems*, **5**, 1-22. https://doi.org/10.4018/jswis.2009081901

[3] Singhal, A. (2012) Introducing the Knowledge Graph: Things, Not Strings. Official Google Blog.

[4] 阮彤, 孙程琳, 王昊奋, 等. 中医药知识图谱构建与应用[J]. 医学信息学杂志, 2016, 37(4): 8-13.

[5] 葛斌, 谭真, 张翀, 等. 军事知识图谱构建技术[J]. 指挥与控制学报, 2016, 2(4): 302-308.

[6] 邵元新. 基于 web 的工业产品知识图谱构建及应用[D]: [硕士学位论文]. 沈阳: 沈阳航空航天大学计算机系, 2017.

[7] Miller, G.A. (1995) WordNet: A Lexical Database for English. *Communications of the ACM*, **38**, 39-41. https://doi.org/10.1145/219717.219748

[8] Auer, S., Bizer, C., Kobilarov, G., *et al.* (2007) DBpedia: A Nucleus for A Web of Open Data. The Semantic Web. Springer, Berlin, Heidelberg, 722-735. https://doi.org/10.1007/978-3-540-76298-0_52

[9] Suchanek, F.M., Kasneci, G. and Weikum, G. (2007) Yago: A Core of Semantic Knowledge. *Proceedings of the* 16*th International Conference on World Wide Web*, Banff, 8-12 May 2007, 697-706. https://doi.org/10.1145/1242572.1242667

[10] Schmitz, M., Bart, R., Soderland, S., *et al.* (2012) Open Language Learning for Information Extraction. *Proceedings of the* 2012 *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, JejuIsland, 523-534.

[11] Carlson, A., Betteridge, J., Kisiel, B., *et al.* (2010) Toward an Architecture for Never-Ending Language Learning. *AAAI*, **5**, 3.

[12] Niu, X., Sun, X., Wang, H., *et al.* (2011) Zhishi.me-Weaving Chinese Linking Open Data. The Semantic Web—ISWC 2011, 205-220.

[13] Hu, F., Shao, Z. and Ruan, T. (2014) Self-Supervised Chinese Ontology Learning from Online Encyclopedias. *The Scientific World Journal*, **2014**, Article ID: 848631. https://doi.org/10.1155/2014/848631

[14] Xu, B., Xu, Y., Liang, J., *et al.* (2017) CN-DBpedia: A Never-Ending Chinese Knowledge Extraction System. *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, Springer, Cham, 428-438. https://doi.org/10.1007/978-3-319-60045-1_44

[15] Yu, T., Li, J., Yu, Q., *et al.* (2017) Knowledge Graph for TCM Health Preservation: Design, Construction, and Applications. *Artificial Intelligence in Medicine*, **77**, 48-52. https://doi.org/10.1016/j.artmed.2017.04.001

[16] 程文亮. 中文企业知识图谱构建与分析[D]: [硕士学位论文]. 上海: 华东师范大学软件工程系, 2016.

[17] 胡芳槐. 基于多种数据源的中文知识图谱构建方法研究[D]: [博士学位论文]. 上海: 华东理工大学计算机应用技术系, 2015.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2161-8801，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：csa@hanspub.org