基于空间分析的气象预报文本实时生成关键技术研究

于敏

2016年12月

中图分类号：

UDC分类号：

基于空间分析的气象预报文本实时生成关键技术研究

作 者 姓 名 于敏

学 院 名 称 计算机学院

指 导 教 师 张华平 副研究员

答辩委员会主席 陈群秀 教授

申 请 学 位 工学硕士

学 科 专 业 计算机科学与技术

学位授予单位 北京理工大学

论文答辩日期 2016年12月

**Research on Key Technologies of Meteorological Forecast Text Generating in Real Time Based on Spatial Analysis**

|  |  |
| --- | --- |
| Candidate Name： | Min Yu |
| School or Department： | School of Computer Science |
| Faculty Mentor： | Associate Prof. Huaping Zhang |
| Chair, Thesis Committee： | Prof. Qunxiu Chen |
| Degree Applied： | Master of Science |
| Major： | Computer Science and Technology |
| Degree by： | Beijing Institute of Technology |
| The Date of Defence： | December，2016 |

基于空间分析的气象预报文本实时生成关键技术研究 北京理工大学

**研究成果声明**

本人郑重声明：所提交的学位论文是我本人在指导教师的指导下进行的研究工作获得的研究成果。尽我所知，文中除特别标注和致谢的地方外，学位论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京理工大学或其它教育机构的学位或证书所使用过的材料。与我一同工作的合作者对此研究工作所做的任何贡献均已在学位论文中作了明确的说明并表示了谢意。

特此申明。

签 名： 日期：

**关于学位论文使用权的说明**

本人完全了解北京理工大学有关保管、使用学位论文的规定，其中包括：①学校有权保管、并向有关部门送交学位论文的原件与复印件；②学校可以采用影印、缩印或其它复制手段复制并保存学位论文；③学校可允许学位论文被查阅或借阅；④学校可以学术交流为目的,复制赠送和交换学位论文；⑤学校可以公布学位论文的全部或部分内容（保密学位论文在解密后遵守此规定）。

签 名： 日期：

导师签名： 日期：

**摘要**

随着社会的发展，气象业务的应用也越来越广泛，通过人工解读气象云图撰写气象文本的模式已经无法满足大众对气象服务文本的时效性和丰富性的要求。原始气象数据具有海量性和专业性的特点，而公共气象服务文本需要用通俗易懂的自然语言进行表述，二者的转换需要多种复杂技术相结合。目前国内外对该领域还缺乏深入系统的研究，无法满足当前社会对气象预报的需求。

本文立足于基于空间分析的气象预报文本实时生成关键技术研究，旨在将专业且繁杂的原始气象数据转化成通俗易懂的气象预报文本，提高气象服务的时效性、科技含量和丰富性，重点对原始气象数据进行空间计算，并建立了一套完整的空间推理规则，另外利用特征提取相关算法处理海量历史气象文本，构建气象预报文本的模板库。此基础上实现了一套面向气象大数据的气象预报文本实时生成系统，以用于及时准确地生成气象预报文本。此外，邀请了中国气象局的三位专家对本系统生成的各类气象预报文本的准确性和通顺性进行评价，准确性达到70%，通顺性达到85%；并将本系统生成的文本与人工撰写文本进行比对分析，合格率可达84%。较为理想的实验结果也证明了本系统的可行性与准确性，具有良好而广阔的应用前景。

**关键词：**自然语言处理；特征提取；空间分析；文本自动生成

**Abstract**

With the development of society, the application of meteorological service has become more and more widely. The method of manual interpreting meteorological images and composing meteorological forecast texts has been unable to meet the requirements of timeliness and richness of public. The original meteorological data is characterized by mass and specialty, while the public meteorological service text needs to be expressed in natural language which is easy to understand. The conversion of the meteorological data needs a combination of many complicated technologies. At present, there is a lack of in-depth and systematic study of this field at home and abroad, which is unable to meet the current social demand for meteorological forecast.

In this paper, the research on key technology of meteorological forecast text generating in real time based on spatial analysis, is aimed at transforming the professional and complex meteorological data into meteorological forecast text which is easy to understand, and improving timeliness, science richness of meteorological service. We focus on analyzing the original meteorological data in spatial way, and establishing a complete set of spatial reasoning rules. In addition, the feature extraction algorithm is used to process the massive historical meteorological texts and build the template library of meteorological forecast texts. On this basis, a real-time generation system of meteorological forecast text for large meteorological data is established, which can be used to generate the meteorological forecast text timely and accurately. In addition, three experts from the China Meteorological Administration were invited to evaluate the accuracy and fluency of the meteorological texts generated by the system, and the accuracy of result was 70% and the fluency of result was 85%. Besides, the texts generated by system were compared with manual texts, and the pass rate was 84%. The experimental results also prove the feasibility and accuracy of this system, and this system has good and broad application prospects.

**Key Words:** natural language processing; feature extraction; spatial analysis ; text auto-generation

**目录**

[第1章 绪论 1](#_Toc469068051)

[1.1 研究背景和意义 1](#_Toc469068052)

[1.2 国内外研究现状 3](#_Toc469068053)

[1.3 论文的主要工作 5](#_Toc469068054)

[1.4 论文的组织形式 6](#_Toc469068055)

[第2章 相关理论及方法 7](#_Toc469068056)

[2.1气象数据格式综述 7](#_Toc469068057)

[2.2气象数据处理工具综述 9](#_Toc469068058)

[2.2.1 SHIP文件 9](#_Toc469068059)

[2.2.2 GDAL库 9](#_Toc469068060)

[2.2.3 QGIS技术理论 9](#_Toc469068061)

[2.2.4 QGIS技术的特点 11](#_Toc469068062)

[2.2.5 QGIS技术与其他相关技术比较 11](#_Toc469068063)

[2.3气象文本的预处理技术 11](#_Toc469068064)

[2.3.1 中文分词 12](#_Toc469068065)

[2.3.2 句法分析 12](#_Toc469068066)

[2.3.3 文本关键特征抽取 14](#_Toc469068067)

[2.4 本章小结 18](#_Toc469068068)

[第3章 基于空间分析的气象数据模型构建 19](#_Toc469068069)

[3.1 研究概述 19](#_Toc469068070)

[3.2 原始气象数据的空间计算 19](#_Toc469068071)

[3.2.1 气象数据配置文件 20](#_Toc469068072)

[3.2.2 基于QGIS的空间计算 20](#_Toc469068073)

[3.3 原始气象数据的空间推理 23](#_Toc469068074)

[3.3.1 地理区域的划分 23](#_Toc469068075)

[3.3.2 地理区划预报顺序规则 27](#_Toc469068076)

[3.3.3 地理空间的推理分析技术 28](#_Toc469068077)

[3.4 本章小结 32](#_Toc469068078)

[第4章 气象公报文本特征提取与建模方法 33](#_Toc469068079)

[4.1 研究概述 33](#_Toc469068080)

[4.2气象预报文本结构分析 33](#_Toc469068081)

[4.3 气象报文特征提取与建模 35](#_Toc469068082)

[4.3.1 气象公报的特征提取 35](#_Toc469068083)

[4.3.2 句子模板库的构建 37](#_Toc469068084)

[4.4 本章小结 40](#_Toc469068085)

[第5章 实验系统原型设计与结果分析 41](#_Toc469068086)

[5.1 实验环境 41](#_Toc469068087)

[5.2 实验系统原型设计 41](#_Toc469068088)

[5.3 实验结果分析 42](#_Toc469068089)

[5.3.1 各气象类型的预报文本实时生成实验分析 42](#_Toc469068090)

[5.3.2 实证比对分析 50](#_Toc469068091)

[5.4 本章小结 52](#_Toc469068092)

[结论 54](#_Toc469068093)

[参考文献. 56](#_Toc469068094)

[攻读学位期间发表的论文与研究成果清单 60](#_Toc469068095)

[致谢 61](#_Toc469068096)

**第1章 绪论**

**1.1 研究背景和意义**

相关数据显示，中国是受灾害影响最严重的国家之一，气象灾害的有效防御还需要气象服务能力的进一步提高；人民生活方式的不断转变和生活质量的不断提高也需要更高层次的气象服务相匹配。但是，现在我国气象服务行业仍然存在着气象服务能力和经济社会发展要求不相适应，产品质量不高，科技含量不足等问题。

随着现代社会的进步，气象服务在我们生活乃至社会的各个领域中都显得越来越重要，同时国家对上述气象服务行业所存在的问题也给予了高度的重视，大力提倡气象科技的发展，这也为气象服务业务的发展注入了强大的生命力。近年来，GIS（Geographic Information System地理信息系统）作为一门空间信息处理技术已经大量应用到气象领域之中，在气象的可视化、图形化领域发挥了重要的作用，但是在具体的气象文本输出方面，多数气象部门仍然采用人工的方式，通过人工去解读大量的实况数据并依靠人工进行气象的描述和文本输出。很明显，这种人工的方式已经无法满足现代社会所需要的多层次、多方面、更新快的气象预报产品的要求。此外，将天气预报员从这些简单、繁琐又重复的工作中解脱出来专注于更专业的气象预报服务也是顺应时代的表现。

普通大众无法理解太过专业的气象数据和云图，日常生活中我们对于气象信息的了解大多出自于经过专家解读、预报员人工书写的气象预报文本。因此，在我们的日常生活中看到的气象产品都是图文结合的。

从气象预报产品生产部门来看，大自然天气情况千变万化，几乎不可能有两篇一模一样的天气预报文本，而目前气象局发布的气象服务文本都是由天气预报员人工编写，这就需要所有的天气预报员们都有较强的气象专业知识和丰富的工作经验，但是这其中存在了大量机械、重复的工作。通过对气象部门天气预报业务流程调研了解到，气象部门每天会获取来自全国各个站点的大量气象观测资料和预报模式文件，预报员通过对这些观测数据进行综合分析生成面向不同大众的具体天气预报自然语言描述，有时候甚至就是“看图说话”的人工文本生成模式。这种生成模式下的对文字精准的要求就造成了天气预报员机械又复杂的工作。该人工文本生成模式主要存在两方面的问题:一、效率低，时效性差，无法满足快速增长的气象服务要求；二、文本预报产品有偏差，由于观测信息是海量的，预报员素质有参差往往处理时间紧迫，难免会出现重要信息遗漏的情况，导致最后产生的气象预报产品出现偏差。

气象服务是所有气象业务产品向社会提供服务的出口，是气象工作的出发点和归宿[1]。气象部门以气象服务为需求引领的气象工作全面展开，在全面提高预报准确率的基础上，着力推动公共气象服务的发展，而气象服务产品的精细化、个性化、多样性、时效性就显得尤为重要。

“图文并茂”是气象服务产品的一般模式[2]，近几年，图形图像技术发展日趋成熟，气象服务中也应用了大量的图形自动化生成技术，如Micaps（现代化人机交互气象信息处理和天气预报制作系统）已经较好的实现了气象可视界面的现代化、精细化、专业化、智能化。但是文本自动生成系统的方法研究却一直没有得到系统深入的发展，较气象可视化处理技术的发展相比显得很滞后，很难满足现代社会对于气象服务的要求，已经成为气象部门急需改进的部分。另外，当前研究和已有产品都只停留在原始气象数据空间处理和气象预报文本自动生成分割处理的阶段，没有将二者连接成一个完整系统，无法实现气象预报文本的实时生成，难以满足当今快节奏的生活要求[3]。

通过上文可知，目前国内气象服务的文本生成的技术水平仍然处于一个较低的层次，预报员每天需要花大量的时间对气象云图进行解读和组织语言生成气象预报文本，这样不仅效率低下，无法跟上现代社会的需求，而且对专业的气象预报员来说也是一种人才的浪费。气象数据的空间分析和气象预报文本的自动生成两部分技术没能很好地连接统一，大大降低了气象预报工作的时效性。本文所研究的基于空间分析的气象预报文本实时生成关键技术，可以实现日常发布的气象公报的高效实时地自动生成，专业的气象预报员们也可以摆脱人工撰写气象文本这种技术含量低又繁琐的体力付出，转向更专业的高科技的气象技术研究，且一步到位，实现从原始气象数据到可读气象预报文本的实时生成。从长远看来，气象预报文本的实时生成技术将有拓展气象服务行业在其他领域的应用，提高气象预报文本的生成效率，提高其服务水平，满足现代社会对于气象服务的高需求。

综上所述，本研究将整合中国气象局近五年的历史气象预报文本，建立面向气象预报文本生成技术的文本特征提取、模式匹配、文本规划组织的自然语言处理模型，并利用GIS相关工具对原始气象数据进行解析和可视化处理，最后运用空间分析技术获取具体气象信息并结合文本模型生成完整的气象预报文本[4]，从而建立基于空间分析的气象预报文本实时生成系统模型与方法。

**1.2 国内外研究现状**

目前国外关于气象预报文本自动生成方法的研究较多，国内则较少，下面就相关研究内容的主要文献分析如下：

（1）国外气象领域较早的开展了天气预报文本的自动或半自动生成技术，大多数基于文字替换（模板）的方法。

国外于上世纪70年代初就开始了天气预报文本的计算机自动或半自动生成技术的研究，大多数是基于文字替换（Template-based Computer Worded Forecasting， CWF）的方法。CWF方法是首先对要生成的气象信息文本建立一个模板，模板能够根据满足用户需要的一般文本格式来建立，与气象信息有关的文字设置为模板的槽。在生成文本时，将实际的气象信息转换为合适的文字填入模板的槽中，当所有的槽都填满后，文本就生成了。代表性的有始于1970年的IFPS（Weather forecast report generator）、1986年的RAREAS、1988年的MarWords(Marine forecast report generator)、1991年的Scribe、1993年的ICWF和1999年的Siren系统[5]。CWF技术经过多年的发展，在上世纪90年代就有较多的系统投入了使用。

从国内来看，有一些气象部门使用了从简单数据到文本表格形式的预报文本，如：吉林省气象科学研究所开发的全国城市天气预报文本转换程序；北京城市气象研究所的张京江、谭晓光等人生成了表格形式的预报产品，基本上是采用了从天气代码直接到对应的文字的简单转换[6]。

CWF技术的显著优点是简单，容易实现，较多的场合用于生成表格式文档，其缺点是需要大量的模板，由此带来了管理上的困难，且不能进行有效的数据分析，同时还不能形成像自然语言那样流畅的文本，因而限制了应用范围。

（2）国内外在气象领域应用NLG技术已经取得一定的研究成果，也有一些面向特定预报领域的文本生成系统

上世纪60年代形成的自然语言处理（Natural Language Processing， NLP）技术，属于人工智能和计算语言学分支，近年来得到较好发展。在气象领域应用NLG技术，在国外已经做了一些很有价值的研究。如前所述的RAREAS以及MarWord等系统后来发展成了基于NLG的技术系统FoG（Forecast Generator），FoG系统主要用于生成海洋天气预报文本，它是在预报分析平台（FPA）基础上开发的，采用FPA的数据库进行抽取，数据经过简单的分析，然后将被选择的内容发送到语言处理器，其内容被结构化后，就形成了语言预报文本（英语和法语）[7-8]。SumTime\_Meteo是由英国阿伯丁大学自然语言生成研究小组开发形成的天气预报应用项目[9]，主要研究了数值天气预报和海洋天气预报文本的计算机自动生成技术。它采用了两年的人工海洋天气预报文本资料以及相应的天气预报数据，建立了天气预报语料库，然后通过数据分析、文档规划、宏观规划以及表现来实现预报文本的计算机自动生成。美国国家海洋和大气管理局(NOAA)于2016年宣布全球气象预报系统(GFS)升级为四维模式，采用集合混合数据同化技术，在三维空间基础上增加时间维度，为天气预报提供一种更加准确及时且不断变化的天气情况图像，并生成天气预报。

从国内来看，综合分析文献发现这方面的研究较少。由上海交通大学于1999年开始研究的多语种天气预报文本自动生成系统（MLWFA）为气象领域内主要的研究成果[10]。中国气象局于2014年研究的气象落区文本自动生成技术实现了从标准化的气象数据到篇章级气象预报文本的自动生成[11]。广西气象减灾研究所研究的农业气象情报文本自动化生成技术实现了多维农业气象数据的图表、文字的自动化生成[12]。广西师范学院的王行行[13]提出了基于拓扑原理的农业气象专业用语拆分与拓扑词库构建方法，实现了从农业气象观测数据到统计分析，再到语言编辑的全过程自动化。然而上述研究均存在一些不足：如在关注文本生成的同时缺少对原始气象数据的自动化处理；气象预报品种单一，无法满足当前多样化气象预报的需求。本论文的研究重点在于对多类型原始气象数据的空间分析，以及自然语言生成框架内的技术研究，在生成模型、文本规划、句子优化、文本连贯性表达等方面做了较为深入的研究。

（3）面向气象服务文本生成的多维时空气象数据的信息提取与推理方法还没有系统、深入的研究，无法形成文本生成的基础能力。

如之前所述，国内外这些项目的成功开展为气象领域的文本生成研究进行了有益的探索，然而这些研究，不管是基于CWF还是NLP技术，其研究热点在于关注预报文本语言的合理组织与表达。但由于气象信息是一种典型的多类型、多要素、多维、时态数据，并具有海量性特点，这决定了不同一般的从数据到文本的生成方法[14-15]，如何从复杂的时空信息中提取符合面向气象服务文本生成的数据也成为了关键的问题。

十年前国内开展的MLWFA研究项目，主要致力于多语种天气预报文本自动生成语言模型的研究，但由于没有很好的前端数据分析能力，导致了文本规划内容无法考虑用户模型的影响，生成的内容也略显呆板，这样大大降低了生成系统的质量；而国外的多数项目如Fog等也同样基于了NLG技术框架来研究实现，研究内容没有充分考虑到气象数据的复杂性，没能对气象数据进行深入分析；英国的SumTime\_Meteo项目尽管已经开始考虑了气象数据的空间特性和时间特性，并指出基于时空数据的信息到文本的自然语言生成将可能成为未来的研究热点[16]，但由于研究的工作刚刚开展，还没能形成完整的方法，也只是在道路气象领域中开展了前期研究与初步应用。

综上所述，国外基于自然语言生成的、面向天气预报的文本生成方法已经得到了初步的研究[17-18]；而在国内还没有引起足够的重视，除了一些业务单位采用简单的CWF技术外，最近的研究应用场景较为单一，并没有形成有效的应用，也没能得到持续深入的进一步研究；同时符合气象服务文本描述的多维复杂气象数据的推理分析还没有引起高度的关注[19-20]；此外，国内外都尚未有研究将空间分析技术与自然语言处理技术相结合，尚未实现从原始气象数据到可读气象预报文本的直接生成。

**1.3 论文的主要工作**

本文所研究的气象文本自动生成系统主要依据两大关键技术：1.基于空间分析的空间推理技术，也就是对数据进行空间上的投影分析、差值分析、叠加分析等,并建立一套完整的规则得出各行政区划与气象情况的对应关系。2.自然语言处理(Natural Language Processing，NLP)技术，主要应用的技术有文本特征提取、二元词组邻接分析、句法分析、文本库建立等，最终生成符合自然语言语法且适应大众阅读习惯的气象预报文本。

本论文的研究，与上述国内外研究相比主要有五点区别：

1、描述语言不同。目前较为成熟的国外的研究主要针对英语和法语，而且至今也没有一套完整的研究体系出现，对于语法规则完全不同的汉语来说移植性不强。

2、系统应用领域不同。国外的该领域研究主要应用于海洋预报文本的生成，应用面窄，排他性强，本研究主要针对公共气象服务，主要是陆地天气，更为复杂多变。

3、描述范围不同。国外的研究主要针对海洋或者小面积地区，天气简单单一，我国国土辽阔，跨经纬度范围大，地势复杂，从而造成天气状况较之国外也复杂多变的多。

4、关注角度不同。之前的研究多把重心放在气象预报的文本流畅上，注重其输出是否符合大众阅读习惯，但是对于海量的气象数据并没有进行深入分析，忽略了气象预报文本生成的实时性。但是本文研究的是空间分析基础上的气象文本输出，对原始气象数据进行完备的空间分析，将结果直接用于气象预报文本的输出，能够及时反映天气状况，实现了气象自动预报系统的实时性。

5.灵活性不同。传统的气象预报生成系统文字模板固定，一个模板就包括了气温、风、雨等多种气象类型，灵活性较低、可移植性不高。本文则是针对雨、雪、风等多种普通气象类型以及台风、暴雨、暴雪等多种灾害预警分别建立模板库，方便气象工作者后续的选择、拼接，灵活性高。

为了充分结合气象预报文本的语言特色，本文在国家气象局气象专家的帮助下，解析了2010-2015年四千多份历史预报文本，对其进行文本特征提取并建立相关句子模板库，利用地理信息系统组件对原始气象数据进行空间分析，最后通过文本的模式匹配及优化生成最终的气象文本，并建立了一套完整的气象预报文本自动生成系统的理论研究方法和模型。

气象数据复杂多变，监测站点繁多也造成了气象数据的海量，气象文本的生成也就不仅仅是从简单数据对应到文本的过程，目前该研究还需要解决三个主要的问题1、如何对繁杂的气象原始数据进行可视化处理，从而使我们可以提取想要的气象信息。2、如何从气象数据中提取气象预报中所需要的气象空间和时间特性信息3、如何利用自然语言处理技术生成流畅准确的气象服务文本。本文拟通过如今较为成熟的GIS技术来实现气象原始数据的分析与可视化处理，通过自然语言处理技术对历史气象预报文本进行特征提取并建立模板库，利用空间分析技术及一套完整的规则体系抽取气象信息，最后通过模式匹配、句子规划以及优化技术生成完整的气象预报文本。

**1.4 论文的组织形式**

以下为各章的内容介绍。

第1章为绪论部分。主要介绍本研究的背景，气象服务行业的发展现状、国内外研究概况以及主要研究内容。

第2章相关理论及方法。主要介绍气象服务领域常用的数据格式，空间分析所使用的工具、方法等；其次介绍相关的自然语言处理技术：中文分词、句法分析和特征抽取。

第3章介绍以QGIS为基础的气象数据的空间分析，包括气象数据解析与图形化处理、气象地理区划、预报规则。

第4章介绍文本特征提取以及建立句子模板库，为生成自然流畅的气象预报文本打好基础。

第5章实验系统原型设计与结果分析。主要介绍基于空间分析的气象预报文本实时生成模型的构建，并展示了实验结果和结果对比分析。

**第2章 相关理论及方法**

**2.1气象数据格式综述**

气象数据是一系列复杂且海量的数据，气象局每天都要从全国两千多个气象站点收集当日气象数据。当前中国气象局使用的气象数据格式为Micaps格式，一共有十四类：

（1）Micaps1：用于地面填图，主要用于雾霾预警、高温预警、寒潮预警、暴雪预警、强对流天气预警、沙尘暴预警、海上大风预警、降水预报。

（2）Micaps2：用于高空填图，在目前气象预报工作中使用较少。

（3）Micaps3：用于通用填图和离散点等值线，主要用于雾霾预警、高温预警、寒潮预警、暴雪预警、强对流天气预警、沙尘暴预警、海上大风预警、降水预报。

（4）Micaps4：用于格点等值线，主要用于雾霾预警、高温预警、寒潮预警、暴雪预警、强对流天气预警、沙尘暴预警、海上大风预警、降水预报。

（5）Micaps5：用于TLOGP和剖面图，目前使用较少。

（6）Micaps6：用于传真图，目前使用较少。

（7）Micaps7：用于台风预警。

（8）Micaps8：用于城市站点预报。

（9）Micaps9：用于底图投影变换。

（10）Micaps10：用于综合图定义。

（11）Micaps11：用于流线图。

（12）Micaps12：用于单点雷达图象。

（13）Micaps13：用于图象(卫星云图、雷达拼图、地形图等)。

（14）Micaps14：用于记录修改后的等值线，主要用于雾霾预警、高温预警、寒潮预警、暴雪预警、强对流天气预警、沙尘暴预警、海上大风预警、降水预报。

以上14类Micaps数据主要分为四个类型：点、线、面、栅格。其中第14类数据中包含点、线、面三种数据，第四类数据为栅格数据，其余数据均为点数据。不同类型的数据会生成不同的文件，我们在第三章中会具体介绍处理过程。并不是一种数据对应一种天气类型，一般来说，一种天气类型的气象信息可能会包含在几种数据中，而一种数据由于内部天气代码的不同也会包含有不同天气类型的信息。

第六类数据为传真图，没有文件说明。其余的数据均有一个文件头，定义了该数据的数据类型、该数据的监测时间、有效时间区间以及其他每个数据独有的数据参数。在数据中首先三列即为经度、纬度、海拔（由于气象预报并不考虑海拔的影响，故海拔值均初始化为0），剩下的数据由文件头中参数决定。

如图2.1所示，为中国气象局每日收集的数据一部分截图（Micaps1类）。

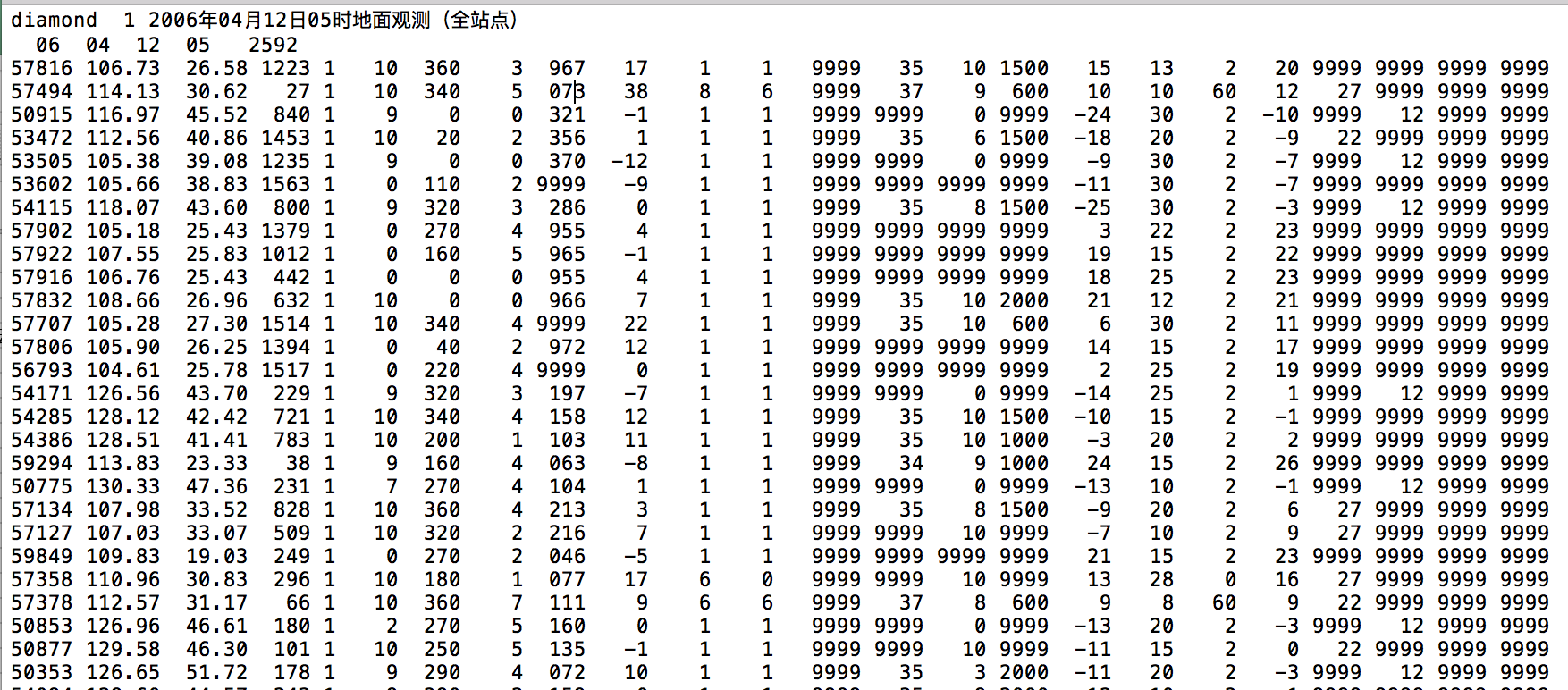
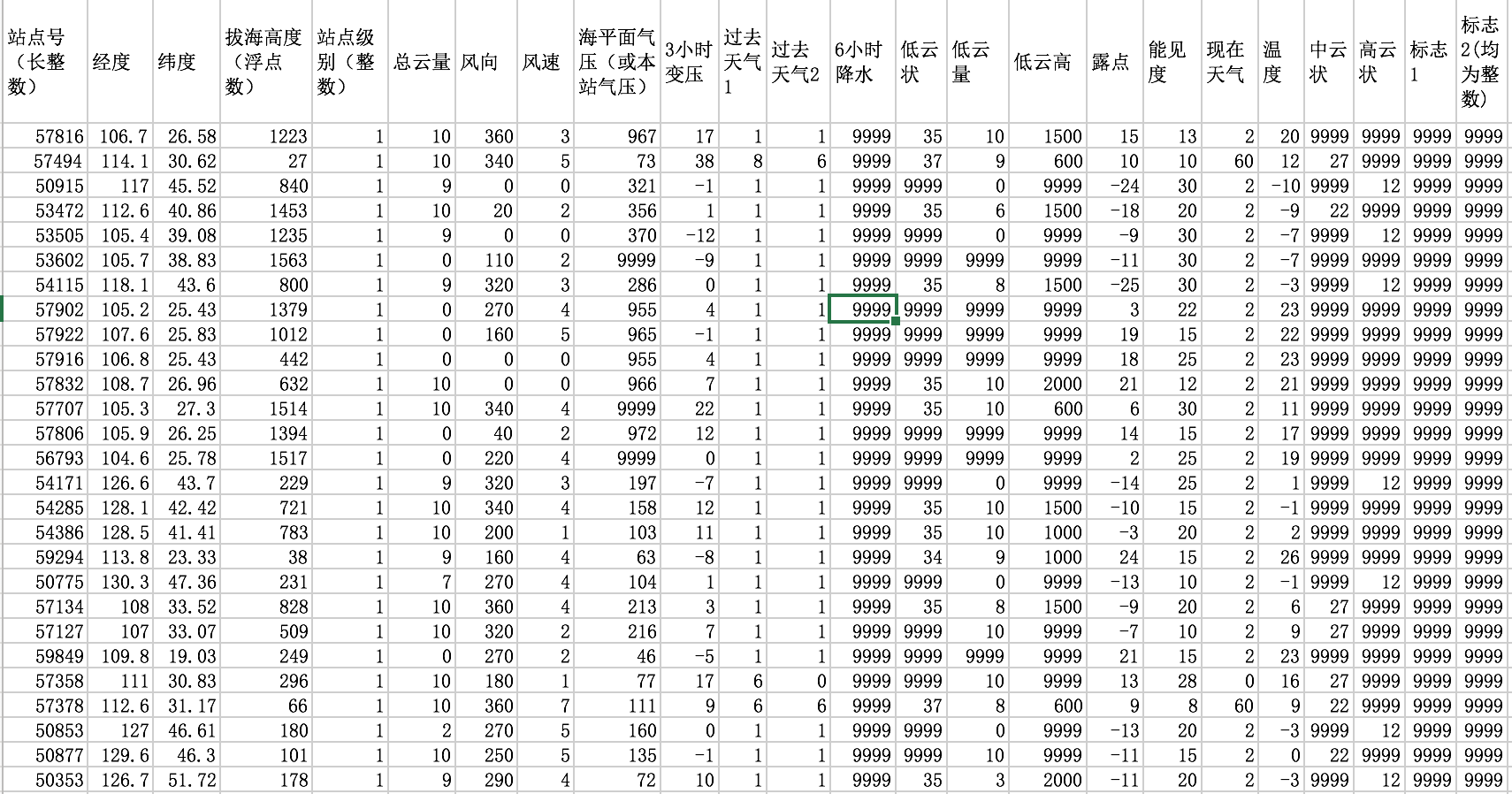


图2.1 原始气象数据部分截图

图中各列数据的含义标注如图2.2所示：

图2.2 原始气象数据含义解释

**2.2气象数据处理工具综述**

**2.2.1 SHIP文件**

SHP文件是ARCGIS支持的、包含一个主文件，一个索引文件以及一个存储矢量数据的文件，SHP文件的存储文件包含一个属性表，可以保存主文件的相关属性。

**2.2.2 GDAL库**

GDAL库全称为Geospatial Data Abstraction Library，它支持空间数据的转换和处理。本研究中可以通过该库生成SHP格式的文件。

**2.2.3 QGIS技术理论**

QGIS是基于Qt,使用C++开发的一个用户界面友好、跨平台的开源版桌面地理信息系统，可运行在[Linux](http://baike.baidu.com/subview/1634/7331705.htm)、[Unix](http://baike.baidu.com/subview/8095/8095.htm)、Mac OSX和Windows等平台之上；QGIS提供接口可用于二次开发。

QGIS项目开始于2002年 5月，是基于跨平台的图形工具Qt软件包，采用C++ 语言开发的一个[地理信息系统软件](http://baike.baidu.com/subview/4098303/4098303.htm)。目前的开发非常活跃，最新版本是QGIS 2.18 版。QGIS源码采用 GNU General Public License协议对外发布。

QGIS将气象数据抽象在特殊格式的文件当中，其类型主要分两种矢量文件和栅格文件。以二维表的形势存储原始的气象数据。如图2.3所示：（其中每一行我们称为Feature)

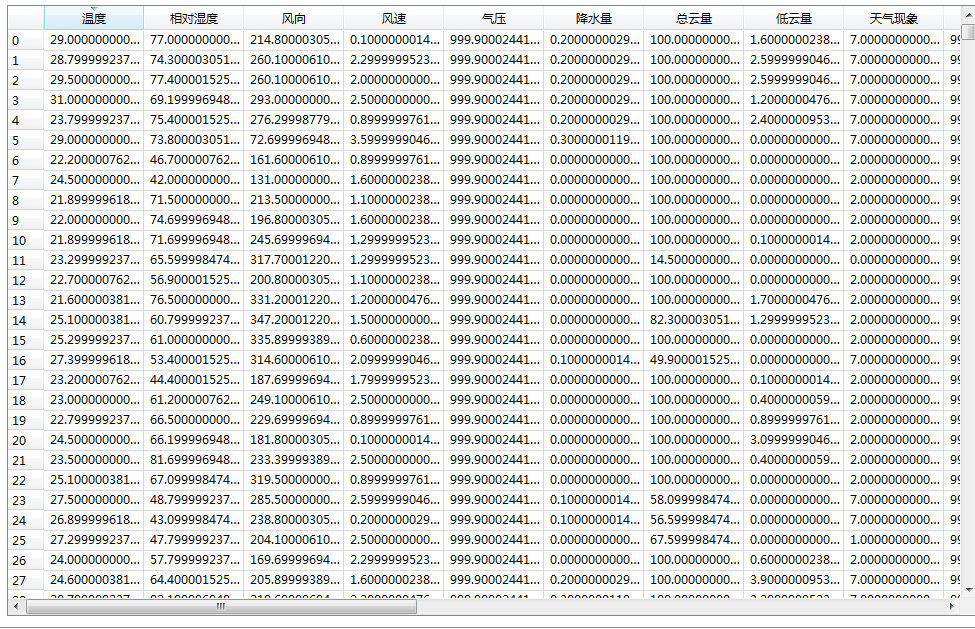


图2.3 存储气象数据的二维表

QGIS将SHP文件分成4类：点，线，面，栅格。针对这样的特点，结合Micaps数据的特性以及气象知识，我们将Micaps1, Micaps2, Micaps3，Micaps5和Micaps7归类为点数据；Micaps14数据当中包括了点，线，面三种数据。其中栅格数据是一种矩阵型的数据。QGIS技术不仅提供了相关的存储原始气象数据的技术，还提供了处理原始气象数据的一系列相关技术，比如投影，插值，叠加，剪切，面积计算等。投影的作用就是保证所有的Micaps数据都能在相同的空间坐标系下进行处理；但由于气象站点数量有限，并且空间分布是离散的，需要利用空间插值的方法获得连续有序的空间数据；叠加能够让多个SHP文件重叠在一起从而利用剪切功能找到重叠的部分，进而利用面积计算功能求得不规则图形的面积。这样我们就能精确的将原始气象数据经过QGIS技术处理后得到所表示的天气类型以及天气所影响的范围。从而获得用于生成气象预报文本的数据。

其中会经过如下几个过程：

（1）SHP文件的生成。

（2）将原始Micaps数据存储到SHP文件当中。

（3）将SHP文件利用QGIS技术进行空间处理生成最终需要的能用于生成报文的数据。

（4）其中涉及到的QGIS技术有投影，插值，叠加，剪切和面积计算等。

**2.2.4 QGIS技术的特点**

（1）将数据以二维表形式存储，利于数据的增删改查。

（2）QGIS技术是开源的，我们可以轻松得到相关的技术文档以及源码，我们能轻松开启QGIS的旅程。

（3）QGIS提供的空间处理技术足够保障原始气象数据能够进行空间分析，使得原始数据能够转化为我们需要的数据。

**2.2.5 QGIS技术与其他相关技术比较**

QGIS在国外成功的案例非常多，虽然有一些其他的GIS技术也有着相似的功能，但是经过综合分析发现，QGIS技术用于本系统是非常适合而且方便的，主要原因如下：

（1）开源降低了本系统的开发成本，且意味着无论是获取还是使用技术都非常方便，而且可以在需要的时候调用我们想要的功能。国外的技术比较偏向使用ARCGIS系统，该系统非开源，价格成本很高，代价太大。

（2）使用QGIS技术利于我们自主设计数据生成系统，不论是图层的生成或者是图层的叠加等操作，都能直接使用相关接口，方便快捷，节约修改源码的时间。

（3）QGIS利用二维表的形式存储数据，利于编程人员的理解，也利于数据操作的规范，使得学过数据库的我们理解QGIS的存储方式时比较容易。

**2.3气象文本的预处理技术**

**2.3.1 中文分词**

中文与英文不同，它是由连续的汉字和标点构成，不存在像英文文本中的空格这种天然分隔符，且词语本身也没有明显的时态、人称等标记。因此，建立气象预报文本模板库的首要步骤就是要对历史气象预报文本进行分词。所谓“分词”，就是按照相关算法将连贯的文本划分为词序，词序与词序之间通过空格或者其它标记符号分隔。

常见的分词算法主要包括以下三种[21]：基于统计模型的分词方法、基于词表的分词方法以及规则方法与统计方法相结合的分词技术。比较经典的分词方法包括：

（1）N-最短路径法。它的基本思路是利用语义词典找出字串中所有可能出现的词，并构造出有向无环图。每个词对应图中的每一条有向边，且具有相应权值。按照权值对有向无环图中的所有路径进行排列，将严格升序排列的路径集合作为相应的粗分结果集。

（2）基于词的n元语法模型的分词方法。它是典型的生成式模型，首先根据词典对句子进行简单匹配，将找出的所有词典词和所有单个字作为结点，最终构造出n元切分词图，其结点代表可能的候选词，边代表路径，边上的n元概率代表代价，再利用相关搜索算法从n元切分词图中找出代价最小的路径作为最终分词结果。

（3）由字构词的汉语分词方法。其基本思想是用字的分类任务代替分词任务，即利用简单的字重组方式实现分词任务。

基于良好的用户体验和较高的分词准确率，本文主要采用NLPIR/ICTCLAS[[1]](#footnote-2)中文分词系统进行分词处理。

**2.3.2 句法分析**

句法分析即分析一个句子的语法结构，例如：哪些是主语或宾语、哪些是动词等。在众多句法分析器中最为常用的是Stanford Parser。

Stanford Parser利用java语言编写，是基于概率的自然语言分析器[22]，基本原理是利用人工标注的语言学知识并通过基于概率的句法分析来对一个新的句子产生最有可能的句法分析结果，主要包括基于PCFG优化的和词汇化的依存句法以及二者的结合。PCFG模型分析速度较快,词汇化的概率句法分析器通过A\*算法实现了一个因式分解生成模型。

Stanford Parser不仅能够针对英文文本进行句法分析，也能够对多种语言进行句法分析，其中包括：基于宾洲树库的中文句法分析器、基于Negra语料库的德语句法分析器和基于Penn Arabic的阿拉伯语句法分析器，以及意大利语、保加利亚语和葡萄牙语等句法分析器。

本文利用Stanford Parser，并选择分析速度较快且适应于中文语句的chinesePCFG模型作为相应的语言学模型。基本配置为JDK1.8版本以上，内存100MB以上来运行PCFG句法分析器。Stanford Parser的句法分析树标注集和句法依存关系如表2-1和表2-2所示。

表2-1 句法分析树标注集



表2-2 句法依存关系表



**2.3.3 文本关键特征抽取**

本节阐述了几种关键特征抽取算法。当前流行基于统计的TF-IDF算法、基于隐含话题模型LDA的关键词抽取算法、基于图模型的TextRank算法进行关键特征抽取。

**（1）基于TF-IDF的关键特征抽取算法**

TF-IDF算法进行关键特征抽取是一种简单通用的方法[23-24]。本小节对其进行简单阐述。

基于TF-IDF算法的关键词抽取的基础工作是要对每篇文档中候选词计算tf-idf值，并对其进行打分。整个过程主要分为三步：

1)统计每篇文档中候选词的词频，并计算它们的tf值；

计算候选词的tf值时，需要注意归一化。由于会出现同一个词在长文档中词频远高于短文档词频的情况，从而发现不了候选词的重要程度，对词频数进行归一化后，可以避免它偏向长度较长的文档。归一化公式为：

(2-1)



其中是该候选词在文档中的词频，而分母部分是文档中全部字词的总和。

2）训练整个语料集的IDF(inverse document frequency)模型；

获取整个语料集中每一个候选词的idf值，这是整个过程中较为重要的一步。计算某一特定词的idf值可以由包含该词语的文件总数除语料集全体文件总数，再对计算得到的商取对数，计算公式如下：



(2-2)

其中，|D|表示研究工作需要分析的语料库中的文件总数；表示包含词语的文件数目（即的文件数目）如果该词语不在语料库中，就会导致分母为零，因此一般情况下使用1 + 。

在遍历整个语料集时，可以计算每个词的idf值。

从而计算出全部候选词的tf-idf值，排序得出关键词；

在获得候选词的tf和idf值后，计算其tf-idf值的方法：

(2-3)



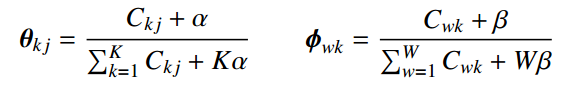
在计算完成所有候选词的tf-idf值后，对结果进行tf-idf值由大到小的排列。选取前M候选词作为其关键词，M是由用户自己指定。

**（2）基于LDA的关键特征抽取算法**

TF-IDF本身具有一定的局限性[25]，即TF-IDF这类统计方法重在考虑候选词的出现频率；这种方法经常会导致推荐的关键词集中分布在同一个很大的主题中，并没有考虑文档中的其他主题。其主要原因是由于没有考虑文档主题分布[26]。本研究工作尝试使用基于LDA主题模型的方法进行关键特征提取。

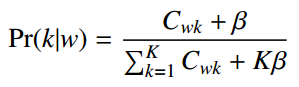
主要步骤如下：

1. 获取文档中的候选词，这一步骤在预处理阶段完成。
2. 对整个语料集进行训练，得出LDA隐含主题模型的各个参数，然后计算得到文档以及候选词的主题分布数据。在训练过程中，首先需要确定LDA中两个参数θ和ϕ：

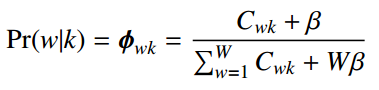


(3.4) (2-4)

其中在吉布斯采样中文档*dj*中的词被赋给主题*k*的次数，代表在训练语料库中词*w*被赋给主题*k*的次数，*W*为词的个数，*K*为隐含话题数。对于任何一个词*w*，可以根据公式(2-4)得到这个词在某个主题k中分布的概率：

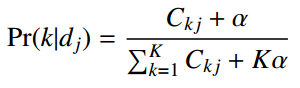
 (2-5)

同时也可以得到主题*k*中某个词*w*的概率：



(2-6)

对于文档*dj*，该文档的主题分布:



(2-7)

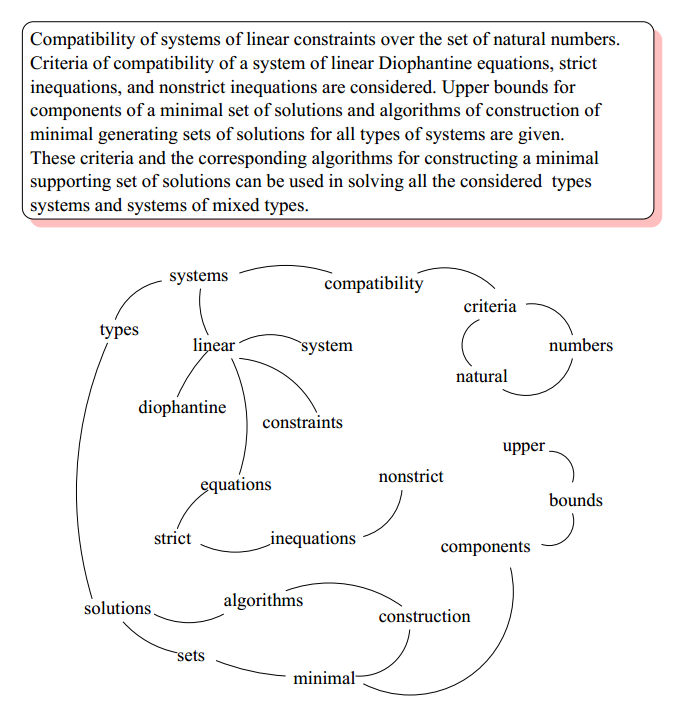
1. 在计算了文档和候选词的主题相似度之后，排序并选取最高的M结果作为关键特征。此处采用一种利用余弦来计算两个主题分布相似度的公式，从而对候选词进行排序。其中余弦相似度可以表示为：

(2-8)

**（3）基于TextRank关键特征抽取算法**

本研究采用的第3种关键特征抽取方法是基于文档结构的图模型算法TextRank，该算法是一种无需预先训练模型的关键词提取算法[27]。与TF-IDF算法不同，TextRank算法是一种更为流行的，基于图模型的候选词打分排序算法，它是将传统的网页打分算法PageRank算法在文本处理上的一种扩展。

Mihalcea[28]等人提出，可以根据文档中词与词之间的关系，构建出一个无向图或有向图。图2.4显示了如何从一个文档构建成为一个无向图。

图2.4 一篇文档构建成无向图

TextRank算法将一个文本表示为带权重的一个图模型（有向图或无向图），其中，表示整个图中的结点，表示结点之间相连的弧，且是的子集。对于有向图，两个节点有一条从指向的有向弧，其权重为；对于任意给定的结点，表示指向结点的所有点的集合，则表示结点从出发的有向弧指向的所有点的集合。因此，结点的得分被定义为：



(2-9)

其中，为阻尼系数，取值范围为0到1之间，通常取值0.85，代表的意义是从图中某一结点随机跳转到所指向的任意结点的概率。

在进行基于TextRank的关键特征抽取算法时，通常情况由下列两个步骤构成：

1. 基于图模型的词汇图构建。

进行TextRank关键特征抽取的第一步就是将原始文本构建成为一个词汇图，即如何确定词与词之间的相关弧。通常情况下，词与词之间的任意关系都可以作为构建相关弧的依据。文献[28]中提到，通过使用指定大小的滑动窗口，窗口大小一般是2-10个词，在该滑动窗口之内共现的任意两词之间可以构建一条弧。文献[28]发现，当窗口取2的时候，关键特征抽取能够取得最好的结果。

1. 结点迭代打分

完成图模型构建后，需要利用公式(2-9)逐一对每个结点进行打分计算，并记录所有分数。在对整个图模型进行迭代打分运算时，若所有结点的相邻两次迭代之间得分之差的绝对值小于预定的阈值（一般为0.0001），可以看作整个模型达到收敛状态，即可停止进行。迭代打分公式如下：

(2-10)



其中，表示第迭次迭代，为结点得分，为第个结点，

当结点的迭代打分达到收敛状态，即满足公式(2-10)的情况下，可得到所有候选词的重要性得分。对候选词得分按照从大到小的顺序排序，取其中前M 候选词作为关键特征。

**2.4 本章小结**

本章重点表述了论文所使用算法的理论基础以及研究中所涉及的关键技术及工具。首先对14类Micaps格式的气象数据进行介绍。其次，本章还详细介绍了气象数据处理所使用的工具QGIS。最后介绍气象文本预处理过程中用到的技术：中文分词、句法分析、文本特征抽取。本章为随后的章节打下理论基础。

**第3章 基于空间分析的气象数据模型构建**

**3.1 研究概述**

当今社会的生产劳动对气象预报产品的准确性和时效性需求迅速增长，气象站点每日观测到的降水、温度、风力、雾霾等数据，对于农林牧渔业、航空航天、灾害预防、环境治理等领域都有相当重要的参考意义。气象数据具有极强的地域特征，空间分析技术对于气象数据的分析以及预报文本的生成有指导性意义。

本研究将构造一个包含所有气象类型的模板数据库，通过对比实际气象预报文本需求选择合适的句子模板构成气象公报的前期框架，其中包含空间变量和气象变量。通过QGIS技术对原始气象数据进行空间计算生成储存中间结果的SHP文件，并制定一套规则将SHP文件中的气象信息转化成具体的天气情况、地理区域的文字描述，再将文字描述填入第四章生成的模板中，一份完整的气象文本就诞生了，本章主要介绍基于空间分析的气象数据模型构建。

QGIS技术结合地理空间认知、信息抽取、统计学等方法，可以对多元气象信息的空间分布进行计算[29-30]。如何对气象情况发生所在的地理区域进行合理划分，并在地理空间识别以及空间特征描述的基础上，建立气象信息的空间特征推理与分析使对于某一气象现象的地理区域描述不至于过大也不会有遗漏是空间分析技术在本文中需要解决的重要问题。例如，在气象云图上某一天气存在不同的量级的无规则分布，怎样通过空间推理技术获取该天气准确合理的地理区域描述。

本章将详细介绍具体的空间分析方法。

**3.2 原始气象数据的空间计算**

气象数据中包含了地理方位、气象代码、距离、方向、等级等多个特征，从这些特征中综合提取出气象空间特征[31]。空间特征的提取还需要充分考虑气象数据的不同气象要素（降水、温度、风力、湿度等）、不同表现形式（单站点数据如观测点数据、格点数据如降水数值），对于特定类型的气象数据在空间特征提取的方法上也存在差异（如观测站点需要先进行插值处理形成空间分布场然后再表达，或者直接提取单站点极值来表达）[32]。

**3.2.1 气象数据配置文件**

中国气象局目前所有的原始气象数据主要为14类Micaps数据，包括降水、气温、台风、雾霾等各类天气要素的气象信息。主要通过气象代码区分各气象类型，不同的气象代码代表了不同的天气类型和天气等级，所以在原始数据解析之前首先需要依据气象局提供的数据说明来配置相关文件，本模型中配置文件为XML形式。

具体配置内容如图3.1所示：

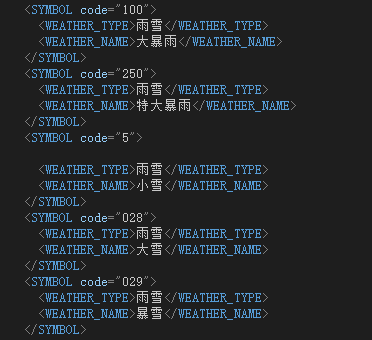


图3.1 原始气象数据配置文件截图

如图，<SYMBOL code>表示天气代码，<WEATHER\_TYPE>表示天气类型，<WEATHER\_NAME>表示天气级别，即具体天气，<SYMBOL code="250">中的250即为雨雪天气类型，级别为特大暴雨。

**3.2.2 基于QGIS的空间计算**

配置文件完成以后，就可以对原始数据进行初步的读取处理。上述原始气象数据分为点、线、面、栅格四类，根据数据类型的不同调用不同的GDAL库接口从而创建不同格式的文件；点、线、面数据对应生成SHP格式文件，栅格数据生成TIF格式文件[33]。以Micaps7数据为例，它区别于其余13种数据，既不与其他数据一起说明同一种天气，也不通过改变天气代码来参与几种天气的说明，它单独决定台风情况。图3.2为Micaps7数据的数据说明，图3.3为处理后得到的SHP文件，图3.4为添加数据之后生成的SHP文件属性表，可以看到属性表中含有台风的时间、位置、速度等属性。

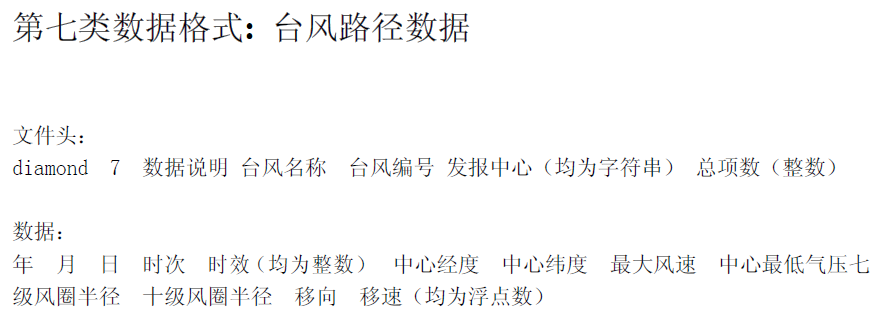


图3.2 Micaps7数据说明

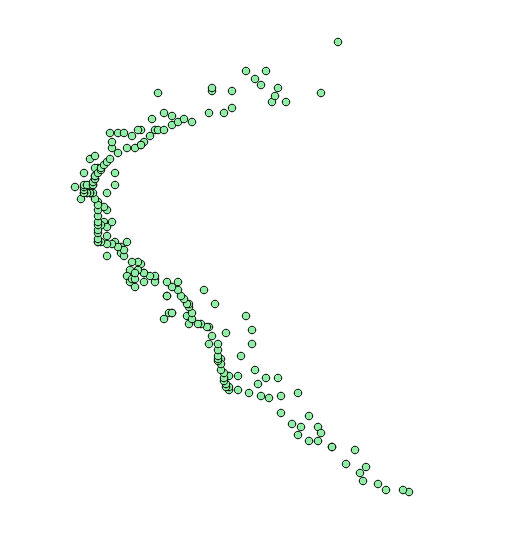


图3.3 Micaps7数据的SHP文件



图3.4 Micaps7数据属性表

由于无论何种数据最后都需要输出叠加之后的面SHP文件，这之前需要对点、线、面数据生成的SHP文件，栅格数据形成的TIF文件进行不同的处理[34]。具体的每种数据处理流程如图3.5所示。可以看到在进行空间叠加之前，点数据会生成点SHP文件，经过插值处理生成TIF栅格数据，再进一步生成面SHP文件；线数据也会生成点SHP文件；面数据直接生成面SHP文件；栅格数据会转换成TIF栅格文件再进一步生成面SHP文件。

数据处理 - 副本

图3.5 各类型数据处理流程图

原始气象数据经过上述步骤的可视化处理，最终生成的SHP文件中包含气象类型的分布图以及对应的属性，为下一步的同中国地图叠加获取相应天气要素的地理区域分布做准备。

接下来将气象类型的SHP文件与中国行政区划的SHP文件进行叠加分析，将两个图层不重合的部分裁减，留下重叠部分，并计算重叠部分的图形面积[35-37]。最终我们将获得地理名称-气象类型-气象等级-区域范围的对应关系表。

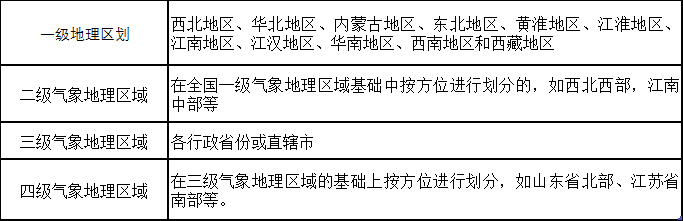
**3.3 原始气象数据的空间推理**

位置特征是空间特征提取的重要组成部分，为了提取有效的位置特征，在进行气象数据空间特征提取时需要建立面向气象预报文本的地理区划，主要面对的问题是如何建立精确且合理的地理区划（如某一灾害天气地理位置，是通过如东北地区即区域级还是通过几个省份一起来描述），地理区域的划分将直接影响生成的气象预报文本的流畅性与可阅读性[38]。另外，空间区域的预报顺序等规则也需要建立[39]。

**3.3.1 地理区域的划分**

本课题对2013-2015年间两千多份各类气象预报文本做了信息提取，同时也抽取了其中的常用地理描述词组。综合数据分析发现，气象预报文本中出现的地理名词描述形式固定，数量非常有限。本课题对近两年出现的地理名词进行统计，针对不同范围的天气情况描述，生成了四级地理区划名词，如表3-1所示。

表3-1 气象地理名词归总表



如上表所示，一级气象地理区划共将中国地理版图划分成为了11个大的地区，二级气象地理区划是在一级地理区划上进行的方位划分，三级气象地理区划是我国的省级行政区划，四级气象地理区划是三级气象地理区划在方位上的再划分，一二级地理区划的关系同三四级地理区划的关系是一样的。可以发现四级气象地理区划之间存在包含关系，具体如下：









就是说，二级区划与三级区划都直接包含于一级区划，又相互独立分别构成了一级区划，三级区划直接包含四级区划。四级区划具体如图3.6、图3.7、图3.8、图3.9所示：



图3.6 一级地理区划

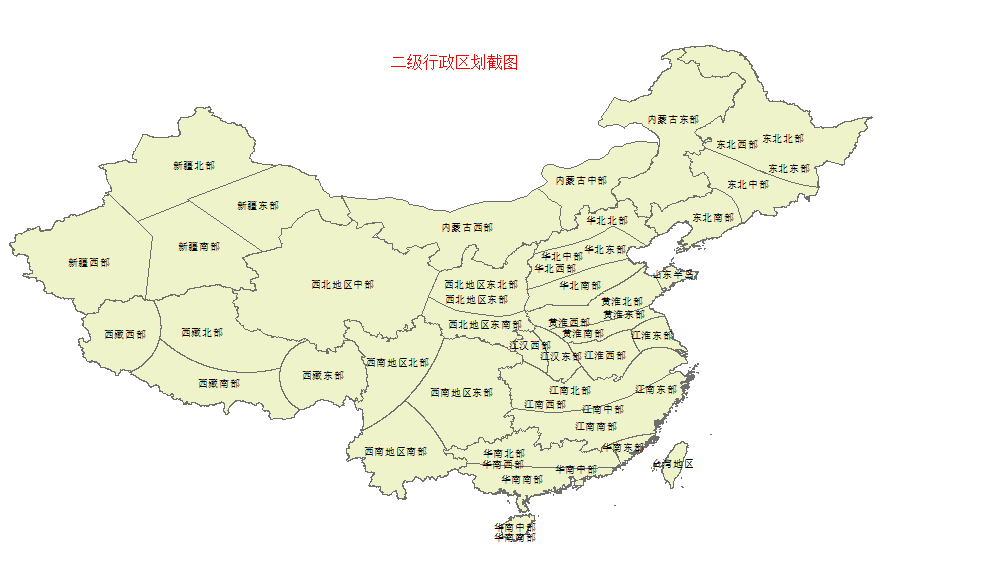


图3.7 二级地理区划



图3.8 三级地理区划

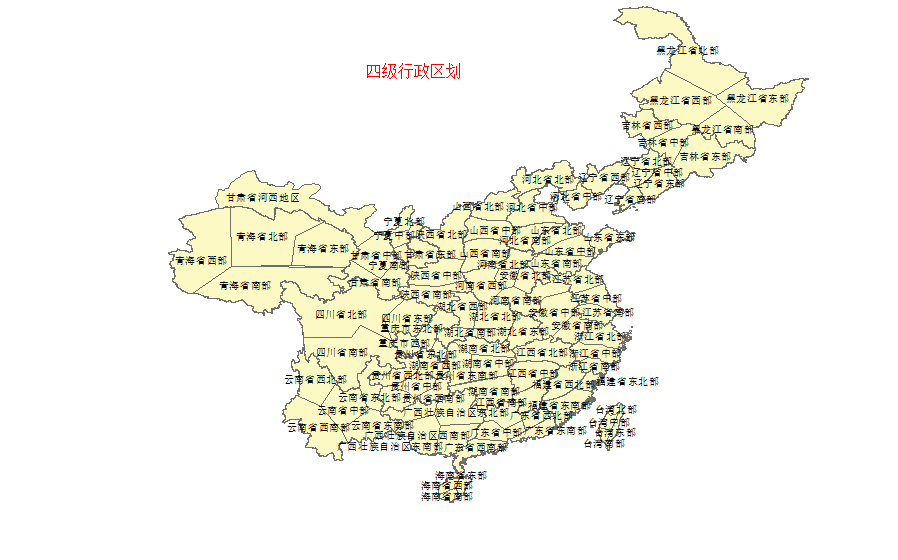


图3.9 四级地理区划

虽然四级气象地理区划可以准确地表述出天气要素的地理区域，但是并不完全符合气象预报员多年来形成的预报习惯，对气象预报文本中出现的地理名词统计发现，除了四级地理区划名词之外，还出现了使用地名、河流、高原、山脉等地理变量名词来描述天气的情况，如青藏高原。因此，要建立合理的气象地理区划，还需要通过气象要素的空间推理技术结合对气象预报文本中地名的命名实体识别来对四级标准区划进行合理的调整，使其更符合气象预报员的预报习惯[40-41]。除此之外，还有中国海域部分，主要包括渤海海域、黄海海域、东海海域、台湾海峡、南海海域、北部湾海域。海域部分不再分二级区域。

**3.3.2 地理区划预报顺序规则**

传统的气象预报文本生成都是依靠气象预报员去解读大量的气象观测数据然后依据过去的气象资料来生成，很大程度上要依靠气象预报员的经验，这样也造成了不同的气象预报员撰写的气象预报也会有所差别。通过对之前的气象服务文本进行分析并对中国气象局专家进行调研发现预报文本的顺序存在一定的规则[42]：

(1)例如天气要素为降水时，描述其所在的全国范围内地理区域时的大体顺序为先从西向东进行预报，再由北向南预报，此规则同样适用于别的天气要素，但是不一定会严格遵守。

根据之前表中列出的一级地理区划名称和上述规则，可以得到一级气象地理区划的预报顺序如图3.10所示。

西北地区常年气候干燥，很少会有降水的情况出现，因此在气象预报中针对此区域的降水天气预报都是从二级地理区划甚至是省份开始的。三级气象地理区划的预报顺序也遵守以上规则，如图3.11所示。



图3.10 一级地理区划预报顺序



图3.11 气象地理区划预报顺序图

(2)预报的区域范围由大到小，一般气象预报中天气要素的预报范围都是从一级地理区划开始，然后是二级地理区划，以此类推……在预报中一个地区不能被预报两次，具体来说就是预报二级区域时不能预报包含在已经预报过的一级中[43]。比如江南地区是一级地理区域，江南南部是第二地理区域，如果已经预报过江南地区，那么后文的预报中则不能再预报“江南南部”之类的地理区域。

(3)预报某地区时一般都不会预报该地区的全称，如在预报广西壮族自治区有小雨时预报文本为“广西有小雨”，这样的好处是更符合自然语言习惯。

(4)在天气预报中预报区域范围的优先级高于预报区域顺序。

如青海省属于全国三级气象地理区域，同时又在全国一级气象地理区域西北地区的范围内，内蒙古自治区属于全国一级气象地理区域，如果是逐级按顺序预报的话，即预报区域顺序优先级高于预报区域范围，那么预报顺序应是这样的：先一级区域，接着二级区域，再到三级区域，即是先预报内蒙古再预报青海，但由于预报区域范围的优先级高于预报区域顺序，所以实际情况却是先预报青海再预报内蒙古。

**3.3.3 地理空间的推理分析技术**

每种天气要素的发生地理位置在气象原始数据中是以经纬度的形式出现的，我们并不能通过气象数据提取到有效的信息，必须要利用QGIS技术对其进行可视化处理得到具体的天气信息，再结合上文中构造的四级气象地理区划，就能获取到每个级别的地理区划对应的天气信息。

利用QGIS模块并结合空间叠加原理，将每个级别的气象地理区划的空间文件分别同预报的天气要素地理位置文件进行叠加，得到重合的部分，利用现有的GIS技术软件就可以读取该重合部分所属的每级地理区划（如叠加部分属于江南地区、江南地区东部、江西省、江西省北部）、天气信息（如天气的类别、级别等）、重合部分面积以及此面积占整个所属地理区划面积的比例[44-45]。确定了如上信息就可以进行空间推理分析了。

以降水为例，具体推理分析过程如下：

假设江南地区降水区域与江南地区面积之比为P，与江南地区中的三级区划面积之比为Q。

if(P≤0.05) 所占比例太小忽略不计

else if（P≥0.9） 所占比例很大，输出江南地区

else if（0.7≤P＜0.9） 输出江南大部地区

else if（P＜0.7） ∥如果不够0.7，则需要分析二三级区划

{

∥判断输出二级区划还是三级

if（符合二级区划） 输出江南西部地区

else if（符合三级区划）

{

for（江南地区每一个省份）

{

∥逐个分析江南地区每一个省份，以湖北省为例

if（Q≤0.05） 比例太低，忽略输出

else if（Q≥0.9） 直接输出湖北省

else if（0.7≤Q＜0.9） 输出湖北省大部

else if（Q＜0.7） ∥进入四级区划输出湖北省西北部或湖北省局部

}

}

}

二级区划与三级区划属于并列关系，描述时只能选择一个，选择规则如下：

假设某个一级区划中，有降水的二级区划个数为m,有降水的三级区划个数为n，该规则制定的主要依据为有降水的地区的数量，如果大多数为二级区划，则分析二级区划；反之，则分析三级区划；若分布均匀，考虑到大众阅读习惯，则优先考虑三级区划。

else if（P＜0.7） ∥如果不够0.7，则需要分析二三级区划

{

∥判断输出二级区划还是三级

if（m==1||m==2）&& m<n 分析二级区划

else if （n==1||n==2）&& m>n 分析三级区划

else if (m > 2 && n > 2) 分析三级区划

}

在进入四级区划后，若仍然满足P<0.7，则依据如下规则进行推理：

以湖北省为例，将湖北省存在降水天气的四级区划按照其降水的面积进行排序，求得降水的最大面积和降水的最小面积，如果最大的面积与最小的面积比值大于阈值（经过多次的实验并与专家商讨，确定阈值的值为10），则将最小的降水面积去掉，然后从剩下的区域中继续选择最小的降水面积，仍然判断是否需要进行删除，直到最小的降水面积不再符合删除的条件。输出规则为：

设剩下m个区划数目，

if (n==1||n==2) 输出：湖北省西北部

else 输出：湖北省局部

在分析二级区划时我们发现，由于某些一级区划形状和位置特殊，简单地将其按照东西南北四个方位来划分并不准确，比如，东北地区可以分为东北东部和东北西部，但是某些情况下划分为东北南部、东北中部、东北北部三个二级区划更能准确描述降水位置。所以我们在原来二级区划的基础上对一级区划的划分进行了修改，最终建立了两套二级地理区划，具体如表3-2所示。图3.13和图3.14为图样示例。

表3-2 二级区划（2）划分表





图3.13 二级地理区划（1）示意图 图3.14 二级地理区划（2）示意图

当降水地理位置图3.15所示时（浅绿色位置），通过空间推理我们最终会输出“东北中南部”，很明显，此描述更准确。



图3.15 某日降水气象云图截图

实现代码如下：

if (符合二级区划)

{

if (Division\_1 is preferred)

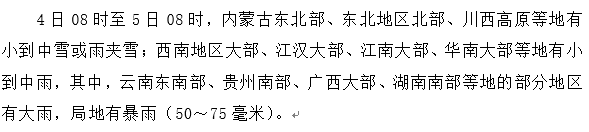
输出：东北东部和西部

if (Division\_2 is preferred)

输出：东北中南部

}

降水模块经过空间分析之后，可以得到一句一句关于降水区域地理描述的句子，经过整合就可以得到全国范围内降水的预报文本，例如：



**3.4 本章小结**

本章提出了基于空间分析的气象数据模型构建方法。详细阐述了基于QGIS处理原始气象数据的方法步骤，包括投影、插值、叠加、裁剪、面积计算等，进而建立气象预报的空间规则，包括全国各级行政区划的划分规则、气象预报的地理空间顺序规则、气象范围与预报行政等级的对应规则。

**第4章 气象预报文本特征提取与建模方法**

**4.1 研究概述**

气象预报文本特征提取，是一种典型的文本信息抽取，为了得到合理、准确的抽取内容，在领域专家的指导下，运用相应的知识库方法，拟先建立面向服务文本的领域本体库，形成领域实体概念、领域属性概念、领域属性值及相互关系，以及该领域所具有的特性和规律的一种形式化描述。具体将从时间描述（早、中、晚、昨天、明天）、空间描述（区域，省级，市级等）、各类天气现象描述（如降水量超过多少是暴雨级别）、服务手段与对策建议等方面建立若干实体以及相互关系，研究形成面向气象服务文本生成的本体库。

气象服务本体库提供了领域内的明确的语义信息。本质上是一种知识库，并具有一定的规则与推理能力。在其基础上，运用信息抽取方法，即对于现有气象服务领域面向不同主题的大量结构化、半结构化或者自由的文本材料进行抽取，经过词法、句法、语义分析等处理来实现，抽取不同命名实体（分别从时间、地点方位、天气条件、天气现象等方面）、以及实体之间的关系等任务，从而形成从时间、地点、天气原因、天气现象事件以及服务手段等多方面的形式化描述特征和推理规则，最后分类建立参数化的知识框架模型[46]。

**4.2气象预报文本结构分析**

在现有大量的、多种类型的历史气象服务自然语言文本材料基础上，如中央气象台日常发布的国内外“天气预报”、“海洋天气预报”、“灾害性天气预警”、“两半刊物信息”、“重大信息专报”以及公共气象服务中心发布的“天气服务公报”等多类产品，研究在合理构建与组织知识库的基础上，如何从这些自然语言表示的自由文本中快速的提取文本特征方法，就是我们本章要讨论的，即段落是如何组织的，句子结构如何，句子中又包括了哪些元素，尤其是如何从描述气象服务内容的时间（when）、地点（where）、事件（Event，即成因）、天气（weather）、服务（services，即如何服务）等多方面内容来进行特征提取，并按不同的服务文本类型进行分类建模，转换为结构化模型元组即形式化描述，形成框架模型[47]。

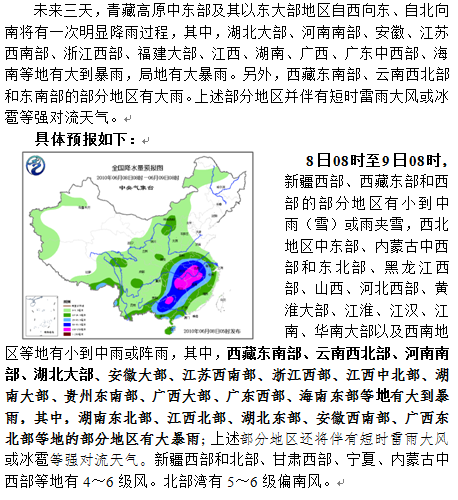
如图4.1所示，为某日发布的气象预报实例，它包含两个部分：气象云图和预报文本。本论文的一个研究中心就是探究如何将气象云图中所包含的气象信息通过计算机而非人工去解读出来并进行文本整合，最终生成符合大众阅读习惯且书写规范的气象预报文本。现对图4.1所示的气象预报文本进行结构组成分析，分析结果如图4.2。

图4.1 气象台发布的每日气象预报

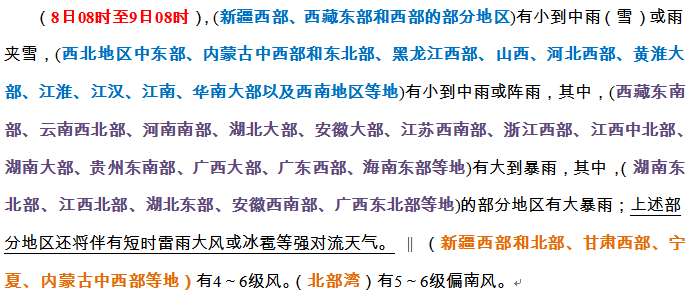


图4.2 天气预报的文本结构分析

结合上图的可以看到，该气象文本的预报内容主要由时间、地理区域、雨雪天气、强对流天气、风力几个部分组成。

时间描述在第一个括号内，用红字标出，可以看到本段是一天的天气情况，所以在时间描述里是一天的跨度，如果时效有变化，则该处的时间描述也会发生相应的变化，所以日期数字可以在文本的自动生成过程中作为一个变量。

气象描述在本预报文本中即是对降水（降雨、降雪）描述、大风天气的描述，天气描述的结构可以用简单的[地理区域]有[气象类型]来进行表述。地理区域描述为除第1、第3个括号外，其余所有括号中的内容，是一连串的地理位置描述的罗列，其中地理区域的划分、描述顺序以及描述规则已在第三章详细介绍，此处不做重复的说明。[气象类型]是对该[地理区域]的天气情况一个描述，在该预报文本中如“中雨”、“大到暴雨”，可以看出一般天气状况是要分等级、分顺序描述的。

**4.3 气象预报文本的特征提取与建模**

如图4.3所示，本文结合气象预报文本的语言特点、大众对于气象信息的需求以及自然语言处理技术建立了一套气象预报文本实时生成系统模型，详细展示了本文如何从气象原始数据结合历史预报文本经过短语、句子组织生成一篇完整的气象预报文本。

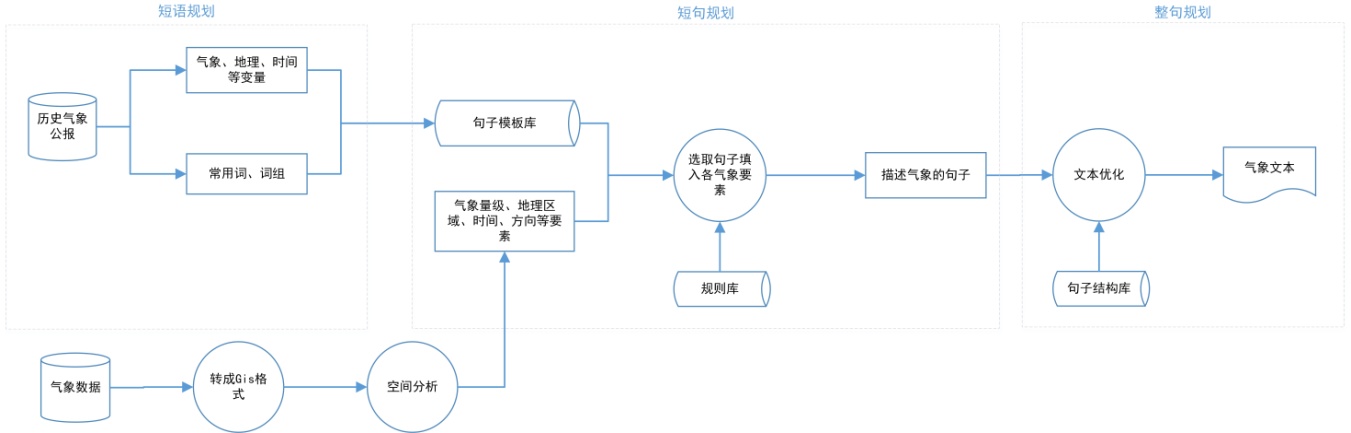
可以看到本模型主要包括三个部分：特征提取、短句规划、整句规划。

图4.3 气象预报文本实时生成系统流程图

**4.3.1 气象预报文本的特征提取**

实验过程中引入中国气象局2013-2015年间各类型的历史气象文本资料共2017份，作为本模型的训练数据，利用张华平老师开发的NLPIR-ICTCLAS（自然语言文本分析工具），对气象文本数据进行信息抽取[48-49]。主要抽取以下两部分信息：1.天气（天气类型和天气预警）和地理区域变量描述词组；2.描述天气的短句模板。这两部分信息结合就可以组成一句完整的天气描述的句子。对上述2017份文本中出现的词语进行词频统计，部分结果如表4-1所示：

表4-1 词频统计部分结果

|  |  |  |
| --- | --- | --- |
| 总词数为2679，平均词频为：498 | | |
| 词语 | 词频 | 概率 |
| 小雨 | 4684 | 0.003897 |
| 中雨 | 4398 | 0.003660 |
| 大雨 | 4899 | 0.004076 |
| 暴雨 | 3468 | 0.002885 |
| 风 | 2355 | 0.001959 |
| 东北 | 5381 | 0.004477 |
| 华北 | 4624 | 0.003877 |
| 西南 | 4293 | 0.003572 |
| 西北 | 3713 | 0.003089 |
| 华南 | 3385 | 0.002816 |
| 内蒙古 | 1738 | 0.001446 |

通过对出现过的所有词语进行词频统计可知，在这一年的所有气象预报文本中，一共出现了2679个词语，可以看到有些词全年只出现了1-2次，属于分词中生僻词。对预报文本进行词频统计信息抽取的目的是得到气象文本编写的通用规律，所以在研究过程中只需要关注重点词汇就可以了，低频率的词汇可以忽略。在出现的词语中，筛去出现频率低的词语，剩下的常用词语其实不多。

二元词组邻接分析，是用于分析文本库中的相邻两个词的关系。部分结果如表4-2所示，所有的二元词对总数只有16251。某两个词在两年中的文本历史库所一起出现的次数，称为共现频次。在前一个词出现的情况下，某词会出现在它后面相邻的概率，称为二元词组的转移频率。这种文本分析对研究预报员的文本写作习惯是非常有效的。

表4-2 二元词组邻接分析部分结果

|  |  |  |  |
| --- | --- | --- | --- |
| 二元词对总数为:16251 | | | |
| 前一个词 | 后一个词 | 共现频次 | 转移概率 |
| 内蒙古 | 东部 | 550 | 0.059276 |
| 黑龙江 | 西北部 | 539 | 0.160343 |
| 四川盆地 | 、 | 538 | 0.216317 |
| 黄淮 | 西部 | 536 | 0.107262 |
| 东北地区 | 、 | 533 | 0.273378 |
| 至 | 明天 | 531 | 0.044718 |
| 西部 | 部分地区 | 525 | 0.120748 |
| 大风 | 。 | 524 | 0.175411 |
| 黄海 | 大部 | 524 | 0.0278 |
| 度 | ， | 521 | 0.35771 |
| 黄淮 | 大部 | 519 | 0.072965 |
| : | 今天 | 512 | 0.595564 |
| 地 | 的 | 508 | 0.175411 |
| 云南 | 大部 | 506 | 0.107269 |
| 福建 | 东部 | 505 | 0.11582 |

可以看出，“内蒙古”和“东部”两个词是一对二元词组，文本历史库中“内蒙古东部”这个词一共出现了550次，其中，当“内蒙古”这个词出现的情况下，后面出现“东部”这个词的概率大约有5.9%，也就是平均每出现20次的“内蒙古”，后面至少有一次会出现“东部”这个词。

结合上述词频统计结果和二元词组邻接分析结果，归纳出气象预报文本必须包含的信息：模板规则和变量词汇，其中变量词汇包括气象变量和地理区域变量，两类变量通常是同时出现的，气象变量主要是指描述各类气象要素的专业气象词汇，比如“雾霾”、“雷暴”、“强对流”、“高温”、“冰雹”等词汇，地理变量信息是描述天气所在位置的地理区域，可以将全国分为四个等级来逐步缩小天气的描述范围，如“华北地区”、“华北东部”、“山东省”、“山东省北部”来使天气范围更为精准。这类变量名词均是可数且有限，指代（天气、地理区域）明确的。

**4.3.2 句子模板库的构建**

经过上述对文本的变量词汇和模板规则的提取，可以建立相应的短句模板和气象变量词汇文本库，描述天气情况的句子主要是由这两部分组成。

表4-3为建立的各类气象预报模板库，其中中括号[]内部的信息表示必填变量，如时间、地点、气象等级等；大括号{}内部信息为选填项，根据气象等级的不同选择是否出现。

表4-3 句子模板库

|  |  |  |
| --- | --- | --- |
| 接口名称 | 句法模板 | 所需数据 |
| 1.雾霾预警 | 中央气象台[xx]月[xx]日[xx]时{继续}发布霾 {黄色，橙色} 预警：  预计[xx]月[xx]日[xx]时，[地点]等地出现{中度}雾和霾，能见度不足[xx]米，{其中[地点]等地{的局部}有重度霾，[地点]等地的局部地区有严重霾，能见度不足[xx]米。} | 雾霾数据  能见度数据  M1，M3，M4，M14 |
| 2.高温 | 中央气象台[xx]月[xx]日[xx]时发布高温{黄色，橙色，红色}预警：  预计[xx]日白天，[地点]等地有[数字范围]的高温天气。{其中，[地点]{大部，部分，局部}温度可达[数字范围]；[地点]{大部，部分，局部}温度可达[数字范围]}}。 | 温度数据  M1，M3，M4，M14 |
| 3.台风预警 | 中央气象台[xx]月[xx]日[xx]时{继续} 发布台风 {蓝色，黄色，橙色，红色} 预警：  今年第[xx]号台风[台风名称]于[xx]月[xx]日[xx]时{早晨[x]时，上午[x]时，下午[x]时，晚上 [x]时}， 位于{[地点]} 北纬[xx]度，东经[xx]度，中心附近最大风力有[数字]级（[该级别下台风风速]），{中心最低气压[xx百帕]}，[数字]级风圈半径[数字]公里，目前正以[台风速度，公里/小时]的速度向[方位]方向移动。 | 台风数据  M7 |
| 4.寒潮预警 | 中央气象台发布寒潮[蓝、黄、橙、红]预警：  预计[时效]，[地点]降温[数字范围]°C，风力可达[级数]，{阵风[级数]}其中[地点]降温可达[数字范围]°C，风力可达[级数]。 | 温度数据  风力数据  M1，M3，M4，M14 |
| 5.暴雪预警 | 中央气象台[xx]月[xx]日[xx]时发布暴雪{蓝色，黄色，橙色，红色}预警：  预计，[xx]日[xx]时至[xx]日[xx]时，[地点]等地部分地区有大雪，{{局部地区有暴雪}，{其中，[地点]有暴雪}}，降雪量[数字范围]mm,{新增积雪深度[数字范围]mm,[地点(突出地点说明)]可达[数字范围]mm}. | 降雪量  M1，M3，M4，M14 |
| 6.降雨 | 中央气象台[xx]月[xx]日[xx]时发布暴雨{蓝色，黄色，橙色，红色}预警：  预计，[xx]日[xx]时至[xx]日[xx]时，[地点]等地部分地区有暴雨，{{局部地区有大暴雨（降水量[数字]mm）}，{其中，[地点]有大暴雨（降水量[数字]mm）}}，{[地点]有特大暴雨，降水量[数字范围]mm }。 | 降水量  M1，M3，M4，M14 |
| 7.强对流天气预警 | 中央气象台发布强对流{蓝色，黄色，橙色，红色}预警：  [xx]月[xx]日[xx]时，[地点]{部分，大部分}可能出现{[风级]雷暴大风，{短时}强降雨，大暴雨，冰雹}等强对流天气。[地点]累计雨量可达[数字]mm,{其中，[地点]可达到[数字]mm}，[地点]等{局部地区}出现直径约[数字范围]的冰雹。 | 冰雹直径，风级，降水量  M1，M3，M4，M14 |
| 8.沙尘暴预警 | 中央气象台[xx]月[XX]日[xx]时发布沙尘暴{蓝色，黄色，橙色，红色}预警：  [xx]月[xx]日[xx]时，[地点]等地{部分，大部分}出现{强沙尘暴，沙尘暴，扬沙，浮沉}天气，其中，[地点]出现{强沙尘暴，沙尘暴}天气。 | 沙尘暴数据  M1，M3，M4，M14 |
| 9.海上大风预警 | 中央气象台发布海上大风{蓝色，黄色，橙色，红色}预警:  预计, [xx]日[xx]时[地点]将有[风级]、阵风[风级]{增加到[风级]，阵风[风级]}的[方位]风，[地点]将有[风级]、阵风[风级]的[方向]。 | 沙尘暴数据  M1，M3，M4，M14 |
| 10.暴雨预警 | 中央气象台[xx]月[XX]日[xx]时发布暴雨{蓝色，黄色，橙色，红色}预警：  预计，[xx]日[xx]时至[xx]日[xx]时，[地点]等地部分地区有大雨，{{局部地区有暴雨}，{其中，[地点]有暴雨}}，降雨量[数字范围]mm,[地点(突出地点说明)]可达[数字范围]mm}. | 沙尘暴数据  M1，M3，M4，M14 |

QGIS模块会对气象数据进行处理，将处理后得到的气象变量信息保存在一张附带属性表的shp文件中，通过空间分析技术确定每一气象类型对应的具体地理区域，最后系统会从不同类别的子库中选择合适的句子进行描述[50]。将合适的气象信息填入句子模板的过程就是一个简单的“填槽”过程，首先根据气象类型从句子模板中选取合适的句子，再将空间分析得到的具体气象信息转化成合适的文字描述填入句子模板的合适位置上，这就输出了一段完整的气象预报文本，至于如何利用空间分析技术得到具体的气象信息本文已在第三章做出详细介绍。

例如，现在有一个模板句子为：

“[时间],[地点]气温将{上升/下降}{度数}℃，其中{地区}{局部地区}{上升/下降}温度可达{度数}℃”

利用QGIS模块处理原始气象数据，我们可以获得温度变化较大的地区的地理名称以及具体的温差，即“山东省北部、河北省大部、江南大部分地区、西北北部地区”，将其填入[地点]处地点对应的位置可得

“20日8时，山东省北部、河北省大部、江南大部分地区、西北北部地区气温将上升4~8℃，其中，河北省大部、江南大部分地区局部地区上升温度可达10~12℃”

**4.4 本章小结**

本章重点阐述了气象预报文本模板库的构建方法，首先分析了气象文本的篇章结构，另外对两千多份历史气象文本进行关键词抽取与二元词组邻接分析，抽取气象预报文本的表达规律、天气名词、地理名词等，最终建立包含所有气象类型的预报模板库。

**第5章 实验系统原型设计与结果分析**

**5.1 实验环境**

硬件配置：CPU 3.20GHz，内存8G，硬盘500G；软件配置：NLPIR2015，Stanford Parser；系统环境：Windows 8.1。

**5.2 实验系统原型设计**

本研究根据前文所述算法，将历史气象文本和气象数据分别进行处理。首先对于历史气象文本，采用关键词抽取、二元词组邻接分析等自然语言处理相关方法提取特征，构建针对不同天气类型的气象预报模板库。另外利用QGIS对气象数据进行空间计算，并且提出一套完整的空间推理规则，最终能够生成符合气象规则且人们可读的实时气象预报文本。实验系统架构图如图5.1所示：

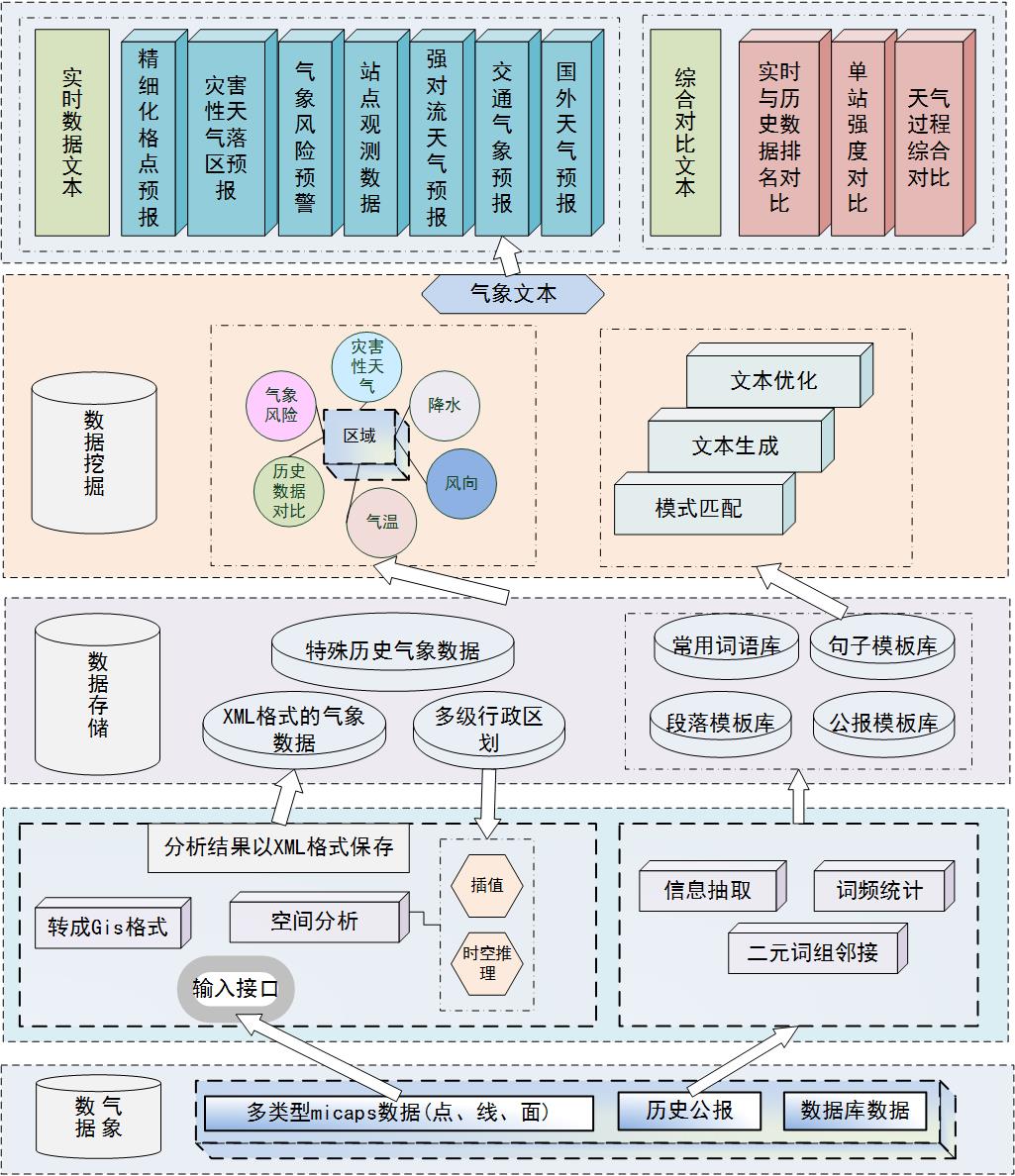


图5.1 实验系统架构图

**5.3 实验结果分析**

本实验采用的气象数据来自中国气象局，该气象数据包括气温、降雨、降雪等普通气象类型以及霾预警、台风预警、暴雨预警等各种预警类型，类别范围广且具有代表性。实验结果分为以下两种评价方式，邀请三位中国气象局专业人员进行评价：

1.对系统生成的各气象类型预报文本分别进行人工评级，从准确性和通顺性两个角度，评价等级分为五级：很好、好、一般、不好、差，其中准确性是指文本中的气象类型、气象数值、地理变量等是否与原始气象数据所表达的一致；通顺性是指文本用词、语句是否符合气象预报的规范。

2.对系统生成的200篇气象预报文本与对应的历史文本（由人工生成）进行相似性比对，分为：很好、好、一般、不好、差五个等级，并统计合格率。

本论文完成的气象预报实时生成系统主要应用于中国气象局的从原始气象数据到可读气象文本实时生成的工作场景中。

**5.3.1 各气象类型的预报文本实时生成实验分析**

本实验以降水预报、暴雨预警、海上台风预警为例，介绍整个模型的实验过程并对实验结果进行展示。原始气象数据的处理在第三章已经有所展示，本章中就不进行多余的描述，主要展示气象空间叠加结果以及预报文本输出结果。

**（1）降水预报结果展示**

将降水圈SHP文件同中国地图SHP文件进行叠加，结果如图5.2所示，其对应的属性表如图5.3所示，

可以看到虽然在叠加SHP文件中没有经过更复杂的着色处理，但是在属性表中可以看到每个降水圈的强度、所在的各级地理区划的名称、降水的面积等属性，下一步只需读取该属性表中相应信息，结合文本生成库以及空间推理规则就能生成完整的降水预报文本。

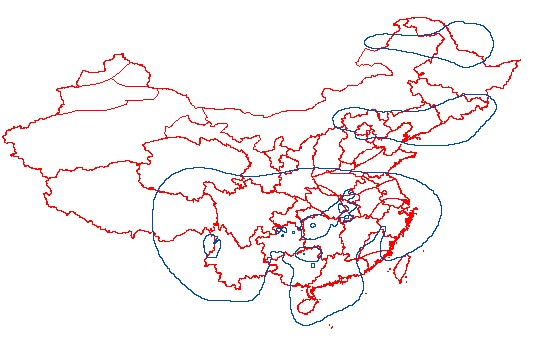


图5.2 降水空间叠加结果

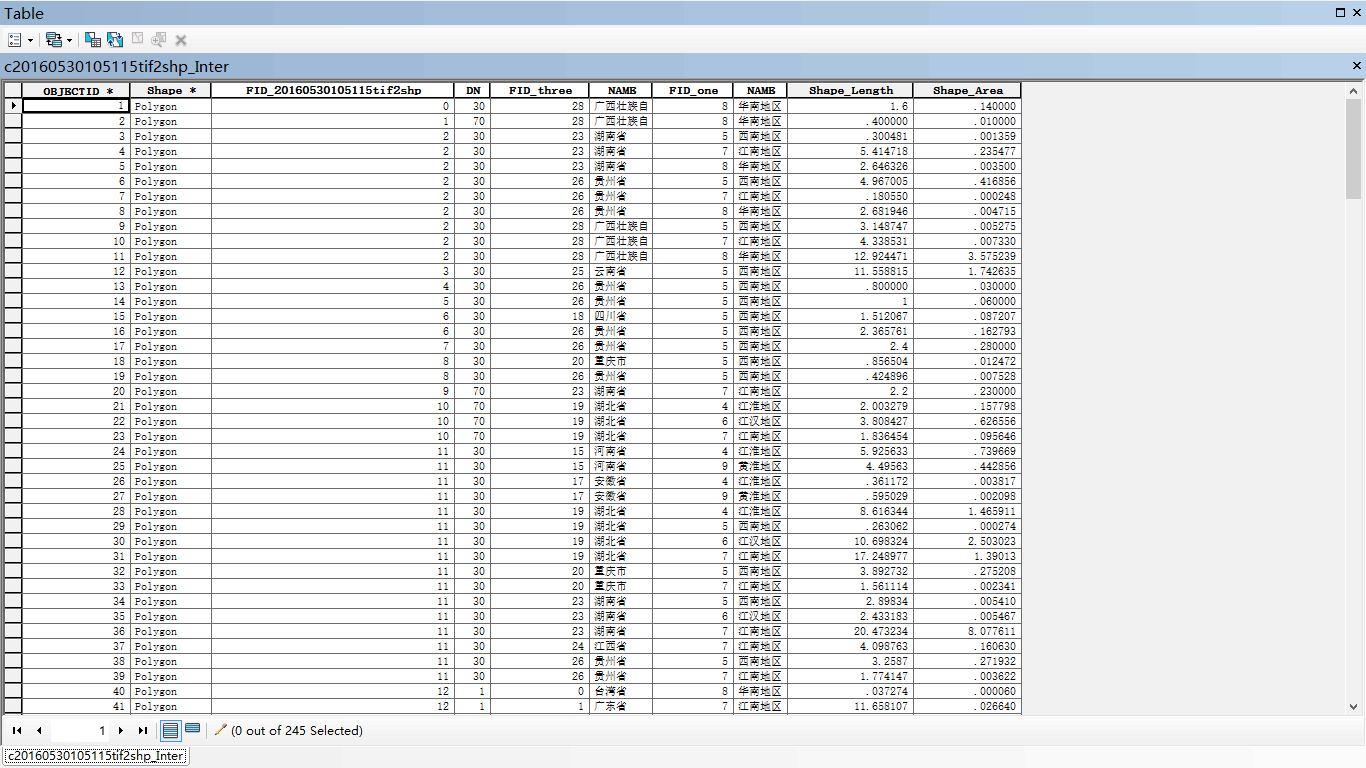


图5.3 降水叠加结果属性表

降水预报的输出结果如图5.4所示：



图5.4 降水预报文本输出

气象人员对该结果进行打分如表5-1所示：

表5-1 气象人员对本次降水预报文本的打分结果

|  |  |  |
| --- | --- | --- |
| 气象人员 | 准确性 | 通顺性 |
| 吕终亮 | 好 | 很好 |
| 张文东 | 好 | 很好 |
| 吴焕萍 | 好 | 很好 |

可以看出，本次系统生成降水预报文本的准确性好，通顺性很好。

此外，随机抽取100份原始降水数据进行处理，由系统自动生成100篇降水预报文本，邀请中国气象局三位气象专业人员分别针对文本的准确性和通顺性进行评级，准确性、通顺性结果如表5-2、表5-3所示：

表5-2 气象人员对降水预报文本的准确性打分评价表

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 气象人员 |  | 很好 | 好 |  | 一般 | 不好 |  | 差 | 合格率 |
| 吕终亮 |  | 5 | 63 |  | 19 | 13 |  | 0 | 87.0% |
| 张文东 |  | 3 | 55 |  | 23 | 17 |  | 2 | 81.0% |
| 吴焕萍 |  | 1 | 48 |  | 31 | 15 |  | 5 | 80.0% |

表5-3 气象人员对降水预报文本的通顺性打分评价表

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 气象人员 |  | 很好 | 好 |  | 一般 | 不好 |  | 差 | 合格率 |
| 吕终亮 |  | 33 | 46 |  | 15 | 5 |  | 1 | 94.0% |
| 张文东 |  | 33 | 45 |  | 15 | 6 |  | 1 | 93.0% |
| 吴焕萍 |  | 27 | 51 |  | 12 | 15 |  | 5 | 90.0% |

可以看出，对于降水数据，系统自动生成的预报文本准确性在80%以上，通顺性可达90%。

**（2）暴雨预警结果展示**

将暴雨SHP文件同中国地图SHP文件进行叠加，其实暴雨预警的SHP文件就是在降水文件的基础上删除未达到暴雨级别的降水圈，只留下暴雨的地理分布区域，其结果如图5.5所示，其属性表如图5.6所示：



图5.5 暴雨空间叠加结果

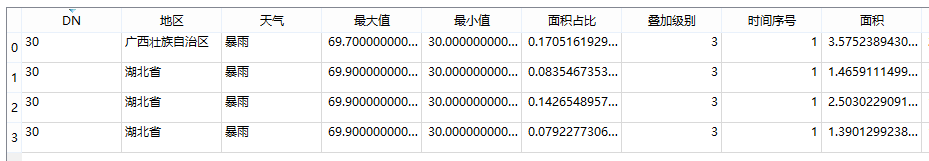


图5.6 暴雨叠加结果属性表

暴雨输出结果如图5.7所示：



图5.7 暴雨预警文本输出

气象人员对该结果打分如表5-4所示：

表5-4 气象人员对本次暴雨预警文本的打分结果

|  |  |  |
| --- | --- | --- |
| 气象专业人员 | 准确性 | 通顺性 |
| 吕终亮 | 好 | 很好 |
| 张文东 | 好 | 好 |
| 吴焕萍 | 一般 | 很好 |

可以看出，本次系统生成暴雨预警文本的准确性合格，通顺性好。

此外，随机抽取100份原始暴雨数据进行处理，由系统自动生成100篇暴雨预警文本，邀请中国气象局三位气象专业人员分别针对文本的准确性和通顺性进行评级，准确性、通顺性结果如表5-5、表5-6所示：

表5-5 气象人员对暴雨预警文本的准确性打分评价表

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 气象人员 |  | 很好 | 好 |  | 一般 | 不好 |  | 差 | 合格率 |
| 吕终亮 |  | 2 | 48 |  | 28 | 20 |  | 2 | 78.0% |
| 张文东 |  | 0 | 47 |  | 27 | 24 |  | 2 | 74.0% |
| 吴焕萍 |  | 0 | 40 |  | 31 | 26 |  | 3 | 71.0% |

表5-6 气象人员对暴雨预警文本的通顺性打分评价表

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 气象人员 |  | 很好 | 好 |  | 一般 | 不好 |  | 差 | 合格率 |
| 吕终亮 |  | 30 | 45 |  | 13 | 12 |  | 0 | 88.0% |
| 张文东 |  | 28 | 43 |  | 14 | 14 |  | 1 | 85.0% |
| 吴焕萍 |  | 28 | 45 |  | 16 | 10 |  | 1 | 89.0% |

可以看出，对于暴雨数据，系统自动生成的预报文本准确性在70%以上，通顺性达到85%。

**（3）寒潮降温预警结果展示**

将寒潮降温SHP文件同中国地图SHP文件进行叠加，结果如图5.8所示，其对应的属性表如图5.9所示：

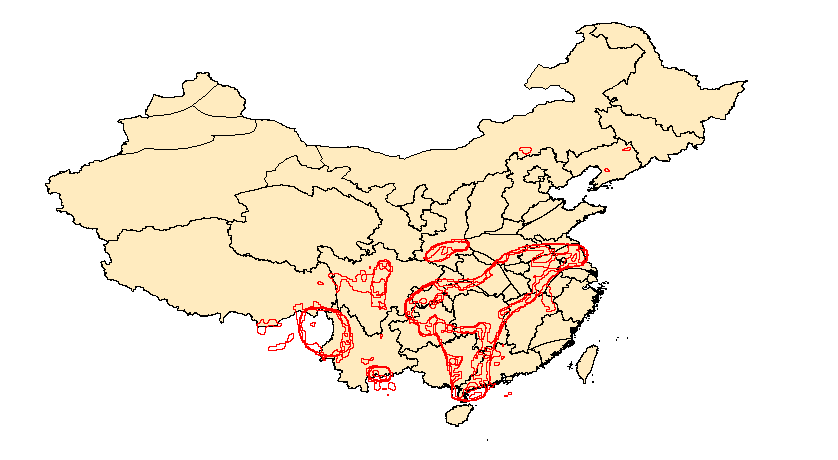


图5.8 寒潮降温空间叠加结果

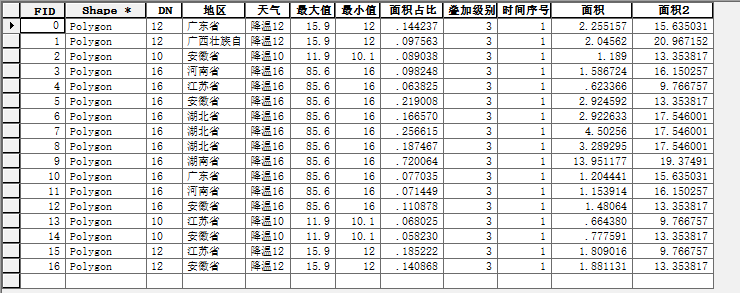


图5.9 寒潮降温叠加结果属性表

寒潮预警文本输出结果如图5.10所示：



图5.10 寒潮预警报文输出

气象人员对本次结果进行打分如表5-7所示：

表5-7 气象人员对本次寒潮预警文本的打分结果

|  |  |  |
| --- | --- | --- |
| 气象专业人员 | 准确性 | 通顺性 |
| 吕终亮 | 好 | 好 |
| 张文东 | 一般 | 好 |
| 吴焕萍 | 好 | 好 |

可以看出，系统生成寒潮预警文本的准确性合格，通顺性好。

此外，我们随机抽取100份原始寒潮数据进行处理，由系统自动生成100篇寒潮预警文本，邀请中国气象局三位气象专业人员分别针对文本的准确性和通顺性进行评级，准确性、通顺性结果如表5-8、表5-9所示：

表5-8 气象人员对寒潮预警文本的准确性打分评价表

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 气象人员 |  | 很好 | 好 |  | 一般 | 不好 |  | 差 | 合格率 |
| 吕终亮 |  | 4 | 48 |  | 28 | 20 |  | 0 | 80.0% |
| 张文东 |  | 2 | 44 |  | 30 | 23 |  | 1 | 76.0% |
| 吴焕萍 |  | 2 | 46 |  | 31 | 21 |  | 0 | 79.0% |

表5-9 气象人员对寒潮预警文本的通顺性打分评价表

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 气象人员 |  | 很好 | 好 |  | 一般 | 不好 |  | 差 | 合格率 |
| 吕终亮 |  | 32 | 45 |  | 13 | 12 |  | 0 | 88.0% |
| 张文东 |  | 30 | 42 |  | 14 | 14 |  | 0 | 86.0% |
| 吴焕萍 |  | 27 | 45 |  | 16 | 12 |  | 0 | 88.0% |

可以看出，对于寒潮数据，系统自动生成的预报文本准确性在76%左右，通顺性在86%左右。

**（4）台风预警结果展示**

将台风SHP文件同中国地图SHP文件进行叠加，结果如图5.11所示，其对应的属性表如图5.12所示：

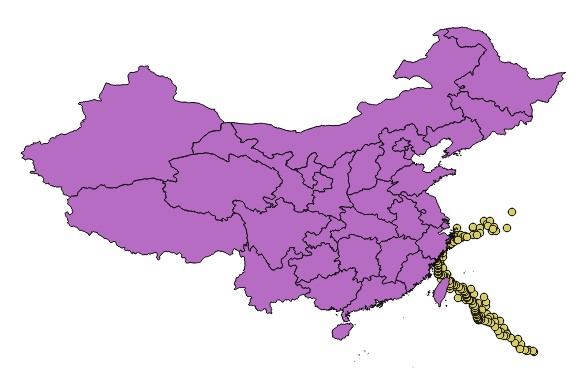


图5.11 台风空间叠加结果

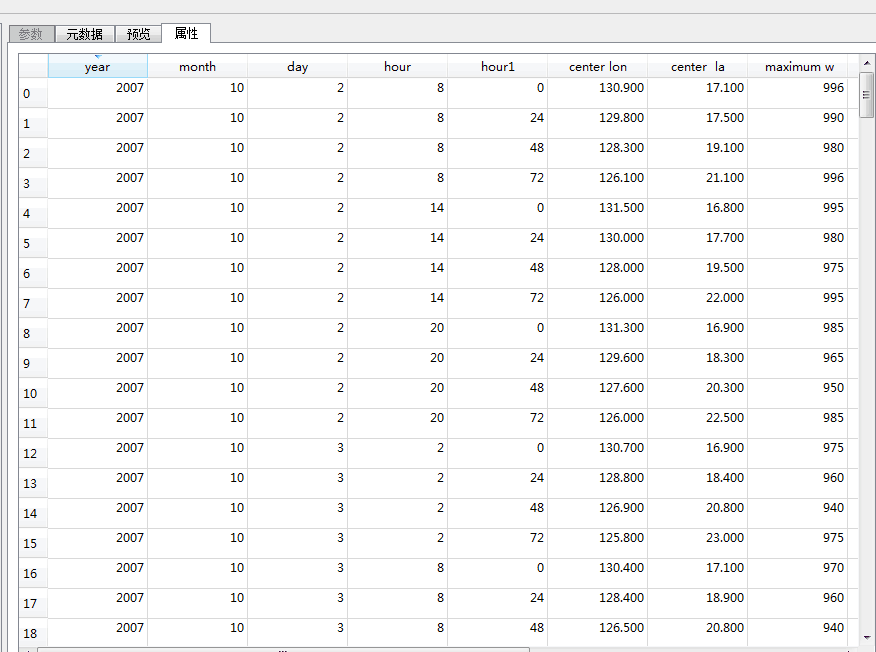


图5.12 台风叠加结果属性表

台风预警文本输出结果如图5.13所示：

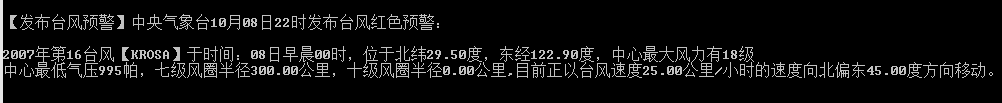


图5.13 台风预警报文输出

气象人员对本次结果进行打分如表5-10所示：

表5-10 气象人员对台风预警文本的打分结果

|  |  |  |
| --- | --- | --- |
| 气象人员 | 准确性 | 通顺性 |
| 吕终亮 | 很好 | 很好 |
| 张文东 | 很好 | 很好 |
| 吴焕萍 | 很好 | 很好 |

可以看出，本次系统生成台风预警文本的准确性和通顺性都很好。

此外，本文随机抽取100份原始台风数据进行处理，由系统自动生成100篇台风预警文本，邀请中国气象局三位气象专业人员分别针对文本的准确性和通顺性进行评级，准确性、通顺性结果如表5-11、表5-12所示：

表5-11 气象人员对台风预警文本的准确性打分评价表

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 气象人员 |  | 很好 | 好 |  | 一般 | 不好 |  | 差 | 合格率 |
| 吕终亮 |  | 47 | 30 |  | 14 | 9 |  | 0 | 91.0% |
| 张文东 |  | 45 | 35 |  | 12 | 8 |  | 0 | 92.0% |
| 吴焕萍 |  | 44 | 37 |  | 10 | 9 |  | 0 | 91.0% |

表5-12 气象人员对台风预警文本的通顺性打分评价表

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 气象人员 |  | 很好 | 好 |  | 一般 | 不好 |  | 差 | 合格率 |
| 吕终亮 |  | 53 | 27 |  | 13 | 7 |  | 0 | 93.0% |
| 张文东 |  | 55 | 25 |  | 15 | 5 |  | 0 | 95.0% |
| 吴焕萍 |  | 49 | 33 |  | 10 | 8 |  | 0 | 92.0% |

可以看出，对于台风数据，系统自动生成的预报文本准确性在91%左右，通顺性可达92%。

综上可以看出，本系统生成的气象预报文本的准确性都在70%以上，通顺性都在85%以上。

**5.3.2 实证比对分析**

在国家气象局的指导和配合下，完成了面向气象大数据的文本自动生成系统，该系统可以直接对大量的原始气象数据进行分析，通过GIS技术和空间推理分析技术获取天气的地理分布区域，集合模式匹配和句子文本库，最终生成流畅通顺的气象预报文本。

目前系统支持各类天气的实况预报和灾害天气预警，为了验证其输出的准确性，随机选取2016年5月28日发布的全国降水预报文本并配以气象云图作为对比，如图5.15 所示：

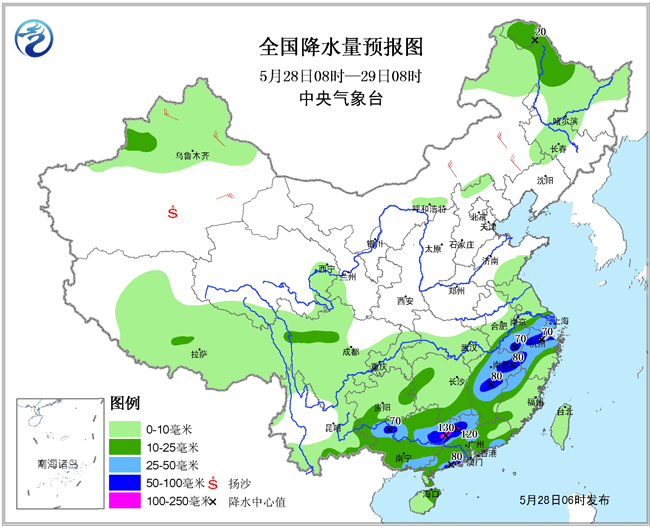


图5.15 5月28日全国降水云图

中央气象台发布的气象预报文本如图5.16：

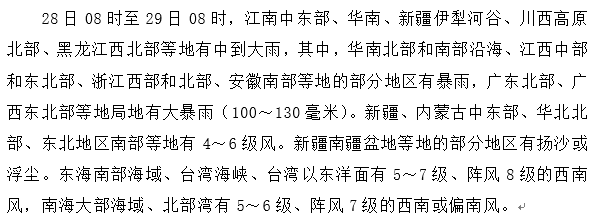


图5.16 气象台具体气象预报文本

原始数据经过系统处理后输出的气象预报文本如图5.17：

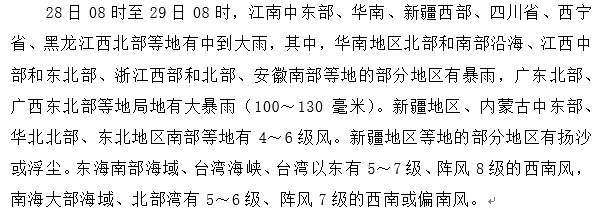


图5.17 文本生成系统输出的气象预报文本

本研究从2013-2015年的气象预报文本数据库中，随即抽取了200天的天气预报（这些文本是预报员手工写的）以及其相对应的原始气象数据（矩阵格式），同时用此系统调用这200天的原始气象数据进行分析，生成相应的天气预报文本。

这两套来自于同一原始数据的天气预报文本结果，由中国气象局的三位气象专业人员亲自查阅对比，并制定了一套5级的打分方法来评判这两种预报文本的相似性，分数越高表明相似度越高，分数越低表明相似度越低。本研究把专家打分的结果经过统计，得到如表5-13的结果，证明了系统自动生成的气象预报文本易读、可用性高。

表5-13 气象人员打分评价表（对比200天的人工预报与系统预报）

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 气象人员 |  | 很好 | 好 |  | 一般 | 不好 |  | 差 | 合格率 |
| 吕终亮 |  | 66 | 93 |  | 19 | 19 |  | 3 | 89.0% |
| 张文东 |  | 57 | 88 |  | 26 | 23 |  | 6 | 85.5% |
| 吴焕萍 |  | 52 | 84 |  | 32 | 24 |  | 8 | 84.0% |

气象专业人员的打分原则是系统生成的文本能否准确并简练的描述当日气象信息，采取五分制原则，分为5个级别，由高到低分别是很好、好、一般、不好、差。在这里规定文本达到一般或者一般以上为合格，最后三人打分的合格率分别为89.0%、85.5%和84.0%，并且达到“好”以上级别的文本能够达到73.3%，由此可见，系统预报的结果能够比较好的描述气象信息，本文所提出的基于空间分析的方法也有足够的可行性。

**5.4 本章小结**

本章对各类气象预报文本的实时生成进行实验分析。利用QGIS处理原始气象数据，包括投影、插值、叠加等，获得叠加图层及叠加结果属性表，再根据空间推理规则结合对应类型的气象预报模板生成最终气象预报文本。还邀请了三位中国气象局的气象专业人员对结果进行打分：首先针对各类气象类型的预报文本，分别从准确性和通顺性两个角度进行评级，准确性结果均在70%以上，通顺性结果均在85%以上；其次抽取2013-2015年内的200份原始气象数据，通过本文提出的技术自动生成气象预报文本，并与对应的由人工撰写的历史气象预报文本进行比对，合格率在84%以上。实验结果表明，本文提出的基于空间分析的气象预报文本实时生成系统能够取得较好的效果。

**结论**

气象是一个很复杂的自然现象，它随时间变化而变化，同时对于天气的描述又需要极为准确。与从一般数据直接生成文本的过程不同，气象数据兼具海量性和时态变化性特征，气象预报的生成需要能准确、快速对天气状态和变化趋势进行描述，也就是说怎样从复杂繁琐的原始气象数据中提取出关于气象要素的空间信息以及怎样建立面向气象预报文本生成的模板库成了本研究的关键技术问题。在本文中，我们利用自然语言处理技术分析海量历史气象文本，抽取其中的语法、用词规律，针对每一种天气类型建立对应的预报文本模板；另外利用QGIS技术对原始气象数据进行解析，提取空间信息，并建立了一套完备的空间推理规律，最后通过模式匹配、文本生成并优化来生成气象预报文本。目前，在中国气象局相关领域专家的支持和帮助下，我们已经证实了该系统的准确性、实时性和实用性，下一步会在中国气象局首先实现和应用此系统。但是，中国气象局发布的气象预报种类繁多，也就是说，本文所研究的系统仅仅是一个开始，在气象领域的文本自动生成技术还有很多可发展空间，届时会需要更多的空间推理方法作为支撑，这也是本研究的后续研究。

本论文主要有两点创新之处：1、较之之前的研究，本研究最大的特点就是引入了空间分析技术。之前大多都是从已经处理好的气象数据到气象文本的生成，重点研究文本生成的流畅性而忽视了气象数据本身的复杂性，但是本文加入了空间分析技术，使系统能够实时生成气象预报文本，更能实时反映天气状况且更灵活流畅。2、已有研究所针对的气象类型较为单一，主要是海洋或者农业气象预报，而本研究能够生成包括雨、雪、气温以及各类灾害预警在内的十种天气类型的预报文本，适用范围广，因此本研究的开展为气象领域的科技支撑注入了一股新的力量，使国内气象预报的实时生成技术提高到了一个新的层次。

虽然实验结果取得了较好的效果，但本文尚有不足之处：（1）仅仅对原始气象数据进行插值，投影等空间分析，没有考虑到实际地理因素比如海拔、海洋、经纬度等对气象数值的影响；（2）气象预报中常常会出现当前气象情况与历史值的横向、纵向对比，然而本文还没有对这一方面进行研究；（3）由于气象预报文本的语料规模相对较大，且打分信息因人而异，会出现一些偏差，针对不同人的不同打分标准如何进行归一化，有待改进。

针对以上不足，作者计划对未来的研究工作做出如下安排：（1）收集更多的语料数据，希望在气象预报文本生成领域实现语料库的构建；（2）更深层次的调研原始气象数据与地理要素的关系，提出更合理的方法能对气象数据进行空间分析；（3）可以在评测阶段对评测者的打分信息进行分布模拟，并根据拟合的分布提出归一化的方法。

**参考文献****.**

[1] 中国气象局业务技术体制改革总体方案[R]．中国气象局办公室文件[EB/OL]．[2006-01-12]．http://www.cma.gov.cn/qxzt/yjtzgg/wjhb/zt/t20060406\_170222.phtml．

[2] Kelly Potts. Word’s first forecast technology aimed at reducing winter road accidents, the Communications Team, Office of External Affairs, University of Aberdeen, King’s College, Aberdeen. [EB/OL]. http://www.abdn.ac.uk/news/archive-details-4153.php.

[3] 傅爱平．计算语言学和自然语言信息处理研究和应用综述．[EB/OL]．[2007-10-01]．http://ling.cass.cn/yingyong/courses/nlpbase.htm．

[4] 刘大有，胡鹤，王生生，谢琦．时空推理研究进展[J]．软件学报，2004，(08)．

[5] 吴焕萍，罗兵，王维国，段延鹤．GIS支持下的决策气象服务系统建设研究[J]．应用气象学报，2008.19(3):380-383．

[6] 张京江，谭晓光．气象服务文本产品的自动生成[J]．气象科技，2004，(S1)．

[7] Goldberg, E．FoG：Synthesizing Forecast Text Directly From Weather Maps[R]. Proceedings of the Ninth Conference on Artificial Intelligence for Applications, p 156-62, 1993.

[8] Eli Goldberg, Norbert Driedger Richard I．Kittredge. Using Natural-Language Processing to Produce Weather Forecasts[J]. IEEE Expert, v 9, n 2, p 45-53, April 1994.

[9] KiRredge Richard, d Lavoie. Benoit Meteocogent: A knowledge-based Tool for Generating Weather Forecast Texts. In Proceedings of the American Meteorological Society AI Conference(AMS-98)[C], 1998, Phoenix, Arizona.

[10] 曾庆辉，姚天畴，顾建峰．天气预报文本生成系统中的天气图处理器[J]．计算机工程，2000，26(3)：28—30．

[11] 吴焕萍，吕终亮，张华平，罗兵，高健，等．气象落区文本自动生成研究[J]．计算机工程与应用，2014，50(13)：247-251．

[12] [李嘉洁](http://xueshu.baidu.com/s?wd=author%3A%28%E6%9D%8E%E5%98%89%E6%B4%81%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson)，[孙涵](http://xueshu.baidu.com/s?wd=author%3A%28%E5%AD%99%E6%B6%B5%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson)，[华璀](http://xueshu.baidu.com/s?wd=author%3A%28%E5%8D%8E%E7%92%80%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson)，[邓树林](http://xueshu.baidu.com/s?wd=author%3A%28%E9%82%93%E6%A0%91%E6%9E%97%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson)，[王行行](http://xueshu.baidu.com/s?wd=author%3A%28%E7%8E%8B%E8%A1%8C%E8%A1%8C%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson)，等．基于气象GIS的农业气象情报文本自动华生成技术研究[J]．气象与环境科学，2016，39(2)：137-143．

[13] 王行行．基于拓扑原理的农业气象情报文档编辑自动化技术研究与实现[J]．广西师范学院，2014．

[14] San-min Wang, Zhao Bin, A logical framework for fuzzy quantifiers Part I：Basic properties, Advances in Soft Computing. 42, 602-611, 2007.

[15] Galanis D, Androutsopoulos I. An extractive supervised two-stage method for sentence compression[C]. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 885-893.

[16] 冯祥胜．模糊量词理论的研究及其在气象预报中的应用[D]．南昌大学，2009．

[17] Pani AK, Bhattacharjee GP. Temporal representation and reasoning in artificial intelligence: A review. Mathematical and Computer Modeling, 2001, 34(1-2):55~80.

[18] Cohn AG, Hazarika SM. Qualitative spatial representation and reasoning: An overview. Fundamental Informatics, 2001, 46(1-2): 1~29.

[19] Zhang XF, Cui WH. Spatio-Temporal analysis and modeling based on the integration of GIS and CA model. Journal of Image and Graphics, 2000, 5A(12):1012~1018 (in Chinese with English abstract).

[20] Zheng KG, Tan SY, Pan YH. A unified spatio-temporal data model based on state and change[J]. Journal of Software, 2001, 12(9): 1360~1365 (in Chinese with English abstract).

[21] 宗成庆．统计自然语言处理[M]．清华大学出版社，2008．

[22] [http：//nlp.stanford.edu/software/lex-parser.shtml](http://nlp.stanford.edu/software/lex-parser.shtml).

[23] 欧健文，董守斌，蔡斌．模板化网页主题信息的提取方法[J]．清华大学学报：自然科学版．2008 (S1)．

[24] 张彦超，刘云，李勇，等．基于自动生成模板的 Web 信息抽取技术[J]．北京交通大学学报：自然科学版，2009，33(5): 40-45．

[25] Ashraf F, Ozyer T, Alhajj R. Employing clustering techniques for automatic information extraction from HTML documents[J]. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 2008, 38(5): 660-673.

[26] Hong J C P, Qian-li C J M. Automatic Text Summarization Based on Thematic Word Weight and Sentence Features[J]. Journal of South China University of Technology (Natural Science Edition), 2010, 7: 012.

[27] 康绍鹏．增强现实关键技术研究[D]．沈阳航空工业学院，2009．

[28] Mihalcea R, Tarau P. TextRank: Bringing Order into Texts [C]. Proceedings of EMNLP, 2004. 404–411.

[29] 黄昌宁，统计语言模型能做什么？[J]．语言文字应用，2002(1)：77- 84．

[30] 蔡自兴，徐光佑．人工智能及其运用[M]．北京：清华大学出版社，1986．

[31] 龚健雅，李小龙，吴华意．[实时GIS时空数据模型](http://search.cnki.net/kcms/detail/detail.aspx?filename=CHXB201403003&dbcode=CJFQ&dbname=CJFD2014&v=MDAwNDFyQ1VSTHllWmVkdEZpcmdXN3pPSmlYVGJMRzRIOVhNckk5Rlo0UitDMzg0emg0WG5EMExUZzJYMmhzeEY=)[J]．测绘学报，2014(03)．

[32] Somayajulu G Sfipada, Ehud Reiter, Ian Davy, Kristian Nilssen．Lessons from deploying NLG technology for marine weather forecast text generation. Proceedings of PAIS[C]. 2004, 2004.

[33] 袁林旺，俞肇元，罗文，易琳，闾国年．[多维统一GIS构建与实现的几何代数方法](http://search.cnki.net/kcms/detail/detail.aspx?filename=DLYJ201305020&dbcode=CJFQ&dbname=CJFD2013&v=MTU3MjRaTEc0SDlMTXFvOUhaSVIrQzM4NHpoNFhuRDBMVGcyWDJoc3hGckNVUkx5ZVplZHRGaXJnVzd6T0lTSFM=)[J]．地理研究，2013(05)．

[34] Ross Turner, Somayajulu Sripada, Ehud Reiter, Ian P Davy. Building a Parallel Spatio-Temporal Data-Text Corpus for Summary Generation. In Applications and Innovations in Intelligent Systems(In Proceedings of the LRE2008 Workshop on Methodologies and Resources for Processing Spatial Language, Marrakech, Morocco, 31 May 2008) [C]. XV: pages 75-78.

[35] Turner R.. Sripada S., Reiter E. and Davy I. Using Spatial Reference Frames to Generate Grounded Textual Summaries of Georeferenced Data. In Proceedings of INLG08[C]. Salt Fork, Ohio, USA, 12-14th June 2008.

[36] Estivill-Castro V, Lee, I. Multi-level clustering and its visualization for exploratory spatial analysis [J]. GeoInformatica, v 6, n 2, p 123-52, June 2002.

[37] Ross Turner, Yaji Sripada, Ehud Reiter. Generating Approximate Geographic Descriptions[C]. Proceeding of the 12th European Workshop on Natural Language Generation. Athens, Greece, 2009: 42-49.

[38] 王洪庆, 张焱, 郑永光等．气象信息科学视算环境及其若干问题[J]．应用气象学报，2004，10:908-719．

[39] 王玉棚, 周勇, 梁丰等．2008年北京奥运会气象服务中的信息资源整合[J]．气象，2009(6)： 109-117．

[40] Harry. R.Glahn. Computer-Produced Worded forecasts[J]. Bulletin of the American Meteorological Society, 1970. 51(12): 1126-1131.

[41] Hua-Ping Zhang, Huan-Ping Wu, Jian Gao. Meteorological Bulletin Automatic Generation based on Spatio-Temporal Reasoning. Proceedings of the 2011 International Conference on Machine Learning and Cybernetics (ICMLC 2011)[C]. p 1927-31, 2011.

[42] 段文广，周晓军，石永炜．[数据挖掘技术在精细化温度预报中的应用](http://search.cnki.net/kcms/detail/detail.aspx?filename=GSQX201201022&dbcode=CJFQ&dbname=CJFD2012&v=MDYzMzQyWDJoc3hGckNVUkx5ZVplZHRGaTNrVzcvUElqN2Fkckc0SDlQTXJvOUhab1IrQzM4NHpoNFhuRDBMVGc=)[J]．干旱气象，2012(01)．

[43] 王红霞，李松．[关联规则挖掘在干旱预测中的研究与应用](http://search.cnki.net/kcms/detail/detail.aspx?filename=WJSJ201012011&dbcode=CJFQ&dbname=CJFD2010&v=MTE0NzhSK0MzODR6aDRYbkQwTFRnMlgyaHN4RnJDVVJMeWVaZWR0Rmkza1c3L1BNaWZZWkxHNEg5SE5yWTlFWlk=)[J]．微计算机信息，2010(12).

[44] 徐薇, 黄厚宽, 秦勇．时空本体研究及在地理信息系统中的应用[J]．铁道学报，2005，27( 4)：119- 124．

[45] 胡文东，丁建军，陈晓光，刘建军．[宁夏精细化预报产品显示与评价业务系统](http://search.cnki.net/kcms/detail/detail.aspx?filename=QXKJ200405013&dbcode=CJFQ&dbname=CJFD2004&v=MTgxMjNBWkxHNEh0WE1xbzlFWjRSK0MzODR6aDRYbkQwTFRnMlgyaHN4RnJDVVJMeWVaZWR0RmkzbVVyM01ORFg=)[J]．气象科技，2004(05)．

[46] 吴建生，刘丽萍，金龙．[粒子群-神经网络集成学习算法气象预报建模研究](http://search.cnki.net/kcms/detail/detail.aspx?filename=RDQX200806013&dbcode=CJFQ&dbname=CJFD2008&v=MjQ0NjRSTHllWmVkdEZpM2tXNy9QTnluYWRyRzRIdG5NcVk5RVo0UitDMzg0emg0WG5EMExUZzJYMmhzeEZyQ1U=)[J]．热带气象学报，2008(06)．

[47] 滕卫平，胡波，滕舟，钟元．[SVM回归法在西太平洋热带气旋路径预报中的应用研究](http://search.cnki.net/kcms/detail/detail.aspx?filename=KJTB201211013&dbcode=CJFQ&dbname=CJFD2012&v=MDAwNDFyQ1VSTHllWmVkdEZpM2tXNy9QTGlmZmJMRzRIOVBOcm85RVo0UitDMzg0emg0WG5EMExUZzJYMmhzeEY=)[J]．科技通报，2012(11)．

[48] 马清，龚建东，李莉，李应林．[超级集合预报的误差订正与集成研究](http://search.cnki.net/kcms/detail/detail.aspx?filename=QXXX200803009&dbcode=CJFQ&dbname=CJFD2008&v=MjY1MDBGckNVUkx5ZVplZHRGaTNuVzczSk5EWFRkckc0SHRuTXJJOUZiWVIrQzM4NHpoNFhuRDBMVGcyWDJoc3g=)[J]．气象，2008(03)．

[49] Berg-Kirkpatrick T, Gillick D, Klein D. Jointly learning to extract and compress[C]. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 481-490.

[50] 钟洛加．[湖北省地质灾害气象预警预报电子政务信息系统的设计与实现](http://search.cnki.net/kcms/detail/detail.aspx?filename=HBDK200503018&dbcode=CJFQ&dbname=CJFD2005&v=MTU3MjRaYkc0SHRUTXJJOUViSVIrQzM4NHpoNFhuRDBMVGcyWDJoc3hGckNVUkx5ZVplZHRGaTNnVnIvTkxTL1A=)[J]．资源环境与工程，2005(03)．

**攻读学位期间发表的论文与研究成果清单**

[1] 第1作者，Yu M, Zhang H, Zhang Y, et al. News Abridgement Algorithm Based on Word Alignment and Syntactic Parsing[M]// Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. 2016.

[2] 第2作者，Zhang Y, Yu M, He Y, et al. Events Influence Computation on Social Media[C]// International Conference on Electronics, Network and Computer Engineering. 2016.

[3] 第3作者，Qiao Y, Zhang H, Yu M, et al. Sina-Weibo Spammer Detection with GBDT[M]// Social Media Processing. Springer Singapore, 2016.

**致谢**

归程，总是比迷途长，长于一生；重逢，总是比告别少，只少一次。

转眼间，六年学习生涯即将结束。从本科到硕士，从十八岁到二十四岁，在北京理工大学度过的这些年充满了感动。不仅学习了各门专业课程，培养了扎实的专业技术，同时也树立了正确的人生观，价值观，“德以明理，学以精工”这八字校训将铭记一生。美丽的校园，丰富的社团活动，让我们的校园生活充实而快乐；严谨的教学，多种学科的竞赛，激励着我们不断汲取知识，为祖国的现代化建设储备力量。

十分幸运成为张华平老师的硕士研究生。不仅获得了在自然语言处理领域的专业指导，更学习到了一位成功导师的为人处世之道。研究生两年半，跟着张老师做项目、搞科研，不仅学到了各种算法知识、增强了项目能力，更学会了如何在困难中学习，如何将所学知识应用于实践，如何培养创新思维。同时，也要感谢张老师在我找工作期间给予我的指导与帮助。衷心感谢实验室张老师和商老师对我学习和生活上的帮助与鼓励，此生难忘！

即将分别，最不舍的是实验室的同学：张雨、乔杨。一起上课、一起复习考试、一起玩的画面还历历在目，写论文、找工作时相互鼓励的场景是最珍贵的回忆。还要感谢新宇、松泽、程程、亚男对我的帮助与鼓励！我将永远铭记一起并肩奋斗，相互陪伴的时光。

最后要感谢我的家人，感谢家人对我的鼓励与鞭策，感谢家人对我每一次选择的尊重与理解。

1. http://ictclas.nlpir.org [↑](#footnote-ref-2)