**Computer Science and Application** 计算机科学与应用, 2018, 8(8), 1183-1194
Published Online August 2018 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2018.88130

Hans汉斯

# Acquisition and Analysis of the Research of Smart City

**Jiawang Rao[1], Yanyan Yang[2], Ronghua Ma[3]**

[1]Jiangsu Province Surveying & Mapping Engineering Institute, Nanjing Jiangsu
[2]Jinling High School, Longhu Campus, Nanjing Jiangsu
[3]Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences, Nanjing Jiangsu
Email: rjw0511230@163.com

Received: Jul. 30th, 2018; accepted: Aug. 10th, 2018; published: Aug. 17th, 2018

## Abstract

**Aiming at the shortage of the research on smart city, especially the lack of collecting a large number of research literature on the basis of overall study on smart city, based on the R language and the network construction of National Knowledge Infrastructure (CNKI), the web crawler was designed, and research papers of smart city were efficiently accessed. Literature database of smart city had been built. Automatic word segmentation model was established, based on the technology of text mining, and then the names of places, keywords and high frequency information were identified and extracted. The time and space distribution characteristics of the smart city, the source of literature and the research hot spot were studied and the research status, research hotspots and developing trends of smart city were revealed overall. Results showed that the web crawler that had been designed in this article performed high feasibility and effectiveness on collecting smart city research literature. The results can provide auxiliary support for government decision-making.**

## Keywords

**Smart City, R Language, Web Crawler, Text Mining**

# 智慧城市研究的获取与分析

**饶加旺[1]，杨颜颜[2]，马荣华[3]**

[1]江苏省测绘工程院，江苏 南京
[2]金陵中学龙湖分校，江苏 南京
[3]中国科学院南京地理与湖泊研究所，江苏 南京
Email: rjw0511230@163.com

收稿日期：2018年7月30日；录用日期：2018年8月10日；发布日期：2018年8月17日

文章引用：饶加旺，杨颜颜，马荣华. 智慧城市研究的获取与分析[J]. 计算机科学与应用, 2018, 8(8): 1183-1194.
DOI: 10.12677/csa.2018.88130

## 摘　要

**针对智慧城市研究的不足，尤其在收集大量研究文献的基础上对智慧城市进行整体研究上的欠缺，本文基于R语言设计了网络爬虫程序，高效的、稳定的获取了中国知网收录的2018年4月前以智慧城市为主题的各类研究文献，并构建了智慧城市文献数据库，在此基础上通过建立的自动分词模型，提取了地名、关键词和高频词信息。通过分析智慧城市研究的时序性、空间分布特征、文献来源和研究热点，揭示了智慧城市研究的发展历程、现状、研究热点并展望了其发展趋势。结果表明，本文设计的网络爬虫程序在获取智慧城市研究文献上具有可行性和高效性，研究成果可为政府决策提供辅助性支持。**

## 关键词

**智慧城市，R语言，网络爬虫，文本挖掘**



## 1. 引言

近年来，随着城市的发展和关注度的逐渐增加，智慧城市(Smart City)作为新兴概念逐步被越来越多的国家和社会公众所认可[1]。在我国，对智慧城市的普遍认知是在数字城市地理空间框架的基础上，通过运用物联网、云计算、大数据、空间地理信息集成等新一代信息技术手段，促进城市规划、建设、管理和服务智慧化的新理念、新模式和新形态[2] [3]。相比较而言，国内关于智慧城市的研究主要集中在物联网、云计算、大数据等方面的技术实现上[4] [5] [6]，为深入的对智慧城市进行研究，更好的促进智慧城市的建设与发展，仅仅从技术实现的层面已难以满足需要，目前关于智慧城市研究的发展历程、研究现状、热点和发展趋势上的非技术层面上的研究较少，而且研究的深度与广度都有较大的发展空间。

本文以中国知网(National Knowledge Infrastructure, CNKI) 2018 年 4 月前收录的以智慧城市为主题的各类研究文献为研究对象，基于 R 语言设计了网络爬虫程序，获取智慧城市文献信息，基于文本挖掘技术建立了面向智慧城市语料文档的自动分词模型，识别和提取地名信息、关键词信息和高频词汇。本文致力于通过分析智慧城市研究的时序性分布、空间分布特征、文献数据库来源和热点，来揭示智慧城市研究的发展历程、现状、热点，并展望智慧城市的发展趋势，更好的服务于智慧城市的建设与发展。

## 2. 建立智慧城市文献信息表

在中国知网(http://www.cnki.net/)中输入“智慧城市”查询关键词，查询结果以列表的形式展示了智慧城市文献信息，包括题名、作者、来源、发表时间、数据库、被引、下载等信息(如图 1 所示)。

根据图 1，本文设计了数字城市文献信息表，共 8 个字段，以题名为主键，(见表 1)。为确保数据正常入库，将题名、作者、来源、数据库设置为长度可变化的字符类型。MySQL 是目前较为流行的关系型数据库管理系统，具有数据交互速度快、体积小、开源和免费的特点而应用广泛[7]，因此本文选用 MySQL 作为智慧城市文献信息表的存储载体。

## 3. 爬虫程序的设计与实现

网络爬虫是依据程序，模拟访问网页、自动化提取网页信息的脚本，是快速获取网页信息的一种方

| | 题名 | 作者 | 来源 | 发表时间 | 数据库 | 被引 | 下载 |
|---|---|---|---|---|---|---|---|
| 1 | 以数据为中心的智慧城市研究综述 | 王静远; 李超; 熊璋; 单志广 | 计算机研究与发展 | 2014-02-15 | 期刊 | 182 | 14522 |
| 2 | 智慧城市视角下城市洪涝模拟研究综述 | 刘勇; 张韶月; 柳林; 王先伟; 黄华兵 | 地理科学进展 | 2015-04-29 09:13 | 期刊 | 24 | 2443 |
| 3 | 基于智慧城市的可持续城市空间发展模型总体架构 | 曹阳; 甄峰 | 地理科学进展 | 2015-04-29 09:13 | 期刊 | 30 | 3435 |
| 4 | 基于地理视角的智慧城市规划与建设的理论思考 | 甄峰; 席广亮; 秦萧 | 地理科学进展 | 2015-04-29 09:13 | 期刊 | 20 | 2251 |
| 5 | 智慧城市标准化工作进展 | 舒印彪; 范建斌 | 电网技术 | 2014-10-05 | 期刊 | 15 | 1546 |
| 6 | 智慧城市应急决策情报体系构建研究 | 李纲; 李阳 | 中国图书馆学报 | 2016-05-15 | 期刊 | 24 | 2352 |
| 7 | 当前我国智慧城市建设中的问题与对策 | 章胜阻; 杨建武; 刘江日 | 中国软科学 | 2013-01-28 | 期刊 | 220 | 16033 |
| 8 | 面向智慧城市建设的居民公共服务需求研究——以河北省石家庄市为例 | 赵勇; 张洁; 吴玉玲; 刘洋 | 地理科学进展 | 2015-04-29 09:13 | 期刊 | 22 | 2597 |
| 9 | 智慧城市是新型城镇化的动力标志 | 牛文元 | 中国科学院院刊 | 2014-01-15 | 期刊 | 30 | 1255 |
| 10 | 金融发展、科技创新与智慧城市建设——基于信息化发展视角的分析 | 湛泳; 李珊 | 财经研究 | 2016-02-03 | 期刊 | 37 | 3015 |

**Figure 1.** Partial literal information of smart city
**图 1.** 智慧城市部分文献信息展示

**Table 1.** Information of smart city literal
**表 1.** 智慧城市文献信息表

| 编号 | 字段名称 | 字段说明 | 字段类型 | 备注 |
|---|---|---|---|---|
| 1 | ID | 编号 | int | 非空 |
| 2 | Title | 题名 | varchar(100) | 主键 |
| 3 | Author | 作者 | varchar(200) | 非空 |
| 4 | Source | 来源 | varchar(100) | 非空 |
| 5 | Publish Data | 发表时间 | date | 非空 |
| 6 | Database | 数据库 | varchar(100) | 非空 |
| 7 | Cited | 被引次数 | int | |
| 8 | Downloaded | 下载次数 | int | |

式[8]，被广泛应用于信息收集与挖掘等领域，按照类型网络爬虫主要分为通用型爬虫、面向主题爬虫、分布式爬虫三种[9]，三者各有优缺点，其中面向主题爬虫应用广泛，形式较为灵活，可针对待定的网页数据进行设计高效的爬虫程序。

一个良好的网络爬虫程序是在不额外增加服务器和客户端负担的情况下，保证程序的运行稳定和数据抓取的高效性。为此本文结合智慧城市文献信息情况，基于 R 语言，选用了面向主题的爬虫方法设计了网络爬虫程序。该程序不借助第三方网络框架，直接响应服务器，提升了数据获取的效率。爬虫程序分为数据抓取、数据处理、数据入库模块，最终将抓取的文献信息批量存储到 MySQL 数据库中，如图 2 所示。

### 3.1. 数据抓取模块

数据抓取是整个爬虫程序的关键，也是构建智慧城市文献数据库的基础。在中国知网中输入“智慧

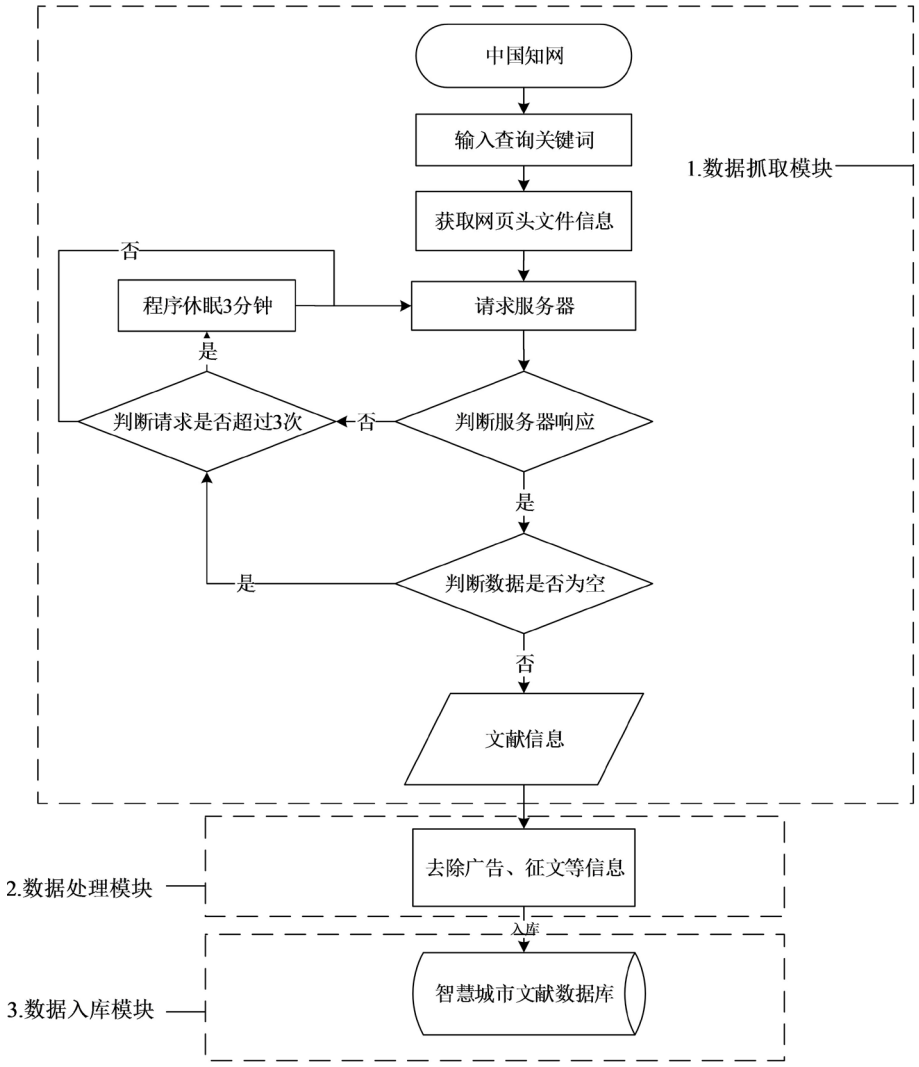


**Figure 2.** The flowchart of web crawler
**图 2.** 网络爬虫程序流程图

城市”查询关键词，获取网页请求的头文件信息，以 GET 方式向服务器发送请求，服务器接受到请求后，返回响应信息，客户端根据响应信息和返回的数据判断服务器是否正常响应，判断返回的数据是否为空，如果为空则继续响应，如果不为空则获取了文献信息(包括题名、作者、来源、发表时间、数据库、被引、下载)。

程序运行过程中要注意的问题有：第一，当遇到服务器无响应，特别是请求超过 3 次时，程序自动休眠一段再重启数据抓取，节省了客户端硬件和软件的资源开销。

第二，为提升数据的抓取效率在设计网络爬虫程序时，仅解析文献列表所在的表格，而不解析整个页面。

## 3.2. 数据处理模块

抓取模块结束后，进行数据处理，考虑到抓取文献信息的字段不同，为了便于数据的处理与入库，抓取的文献信息设置为 List 格式(List 是 R 语言常用的数据结构，常用于存储长度和属性不同的信息)。此时的文献信息包括题名、作者、来源、发表时间、数据库、被引、下载共 7 个属性，根据文献的标题与内容，

通过人工判别的方法去除广告、征文等无关信息，最终抓取了 22,936 篇智慧城市为主题的研究文献。

### 3.3. 数据入库模块

把处理好的文献按照表 1 的智慧城市文献信息表，批量存储到本地 MySQL 数据中，形成智慧城市文献数据库(共包括 ID、Title、Author、Source、PublishData、Database、Cited、Downloaded 八个字段)，作为本文数据的处理与分析的基础。

## 4. 自动分词模型的构建与实现

题名是论文内容和研究方向的总结和概括，通过挖掘题名中的高频词汇和关键词信息，能够反映研究领域的热点问题[10]。中文自动分词是文本挖掘技术的基础，也是本文地名信息、关键词信息、高频词汇提取的关键，jieba 分词是目前主流的中文分词方法，采用动态规划方法和汉字成词能力的隐马尔可夫模型把计算机不能理解的词汇按照一定的规则切分组合成计算机能理解的中文词汇序列[11]，其支持多种开发语言，具有使用方便、分词精度高等优点[12]。目前与智慧城市相关的分词模型和语料库方面还没有研究。

本文在智慧城市文献数据库的基础上，通过提取 Title 字段，得到了智慧城市语料文档，结合 jieba 分词方法构建了自动分词模型，该模型分为语料文档的预处理、分词、结果评测三个模块。模型的关键技术是先从处理各类语料文档入手，对比和整合语料文档，最终生成新词典，在实际运行中不仅避免了直接使用 jieba 分词带来的“智慧城市”、“大数据”等重要专业名词被误分的情况，而且还节省了由于反复扫描不同语料文档带来的时间上和计算机软硬件资源上的耗费，从而提高了分词的准确率和效率。通过分词处理，最终从智慧城市语料文档中提取了关键词、地名标注信息与分词信息，如图 3 所示。

### 4.1. 语料文档的预处理

语料文档和词典直接关系分词的效率和准确度。依据智慧城市语料文档中既包括了测绘科学、地理信息系统、计算机等领域的专业名词，又包括了诸如“实现”、“原理”等常用的普通名词，“的”、“和”、“在”等无实际意义的词和省市县(区)街道等地名信息，增加词典的种类和数量不仅会导致分词效率降低，往往还会出现部分词汇错分的情况。

1) 首先从语料文档和词典入手，建立基础语料文档、停用词、专业词典和地名地址语料文档。其中，基础语料文档来源于人民日报 1998 年标注语料库、微软研究院标注语料库；停用词是为了提高分词效率和检索效率，自动过滤掉某些字或词的组合[13]，从人民日报 1998 年标注语料库提取形成；专业词典来源于测绘科学、计算机科学、地理信息系统专业名词和基于知识理解的从语料库中提取的词汇；地名地址语料文档来源于中国地名录[14]、民政部全国行政区划查询平台[15]的行政区划信息(省市县乡镇)四级地名信息。

2) 语料文档预处理：基于面向词向量的文档对比算法[16]，对基础语料文档、专业词典和地名地址语料文档进行对比分析，找出相同的词汇和不同的词汇，通过整合最终得到既包括基础语料文档，又包括专业词典和地名地址语料文档的“新词典”，在分词过程中省去了重复检索不同语料文档中包含相同词汇的时间。

### 4.2. 分词模块

jieba 分词方法对智慧城市语料文档的分词过程包括基于前缀词典(新词典)的词图扫描、利用动态规划查找最大概率路径，识别出语料文档中所有可能构成词的最大切分组合、利用隐马尔可夫模型(HMM)处理智慧城市语料文档中的未登录词。具体内容如下：

1) 利用新词典对智慧城市语料文档中的语句进行词图扫描，生成所有可能的词汇构成的有向无回路

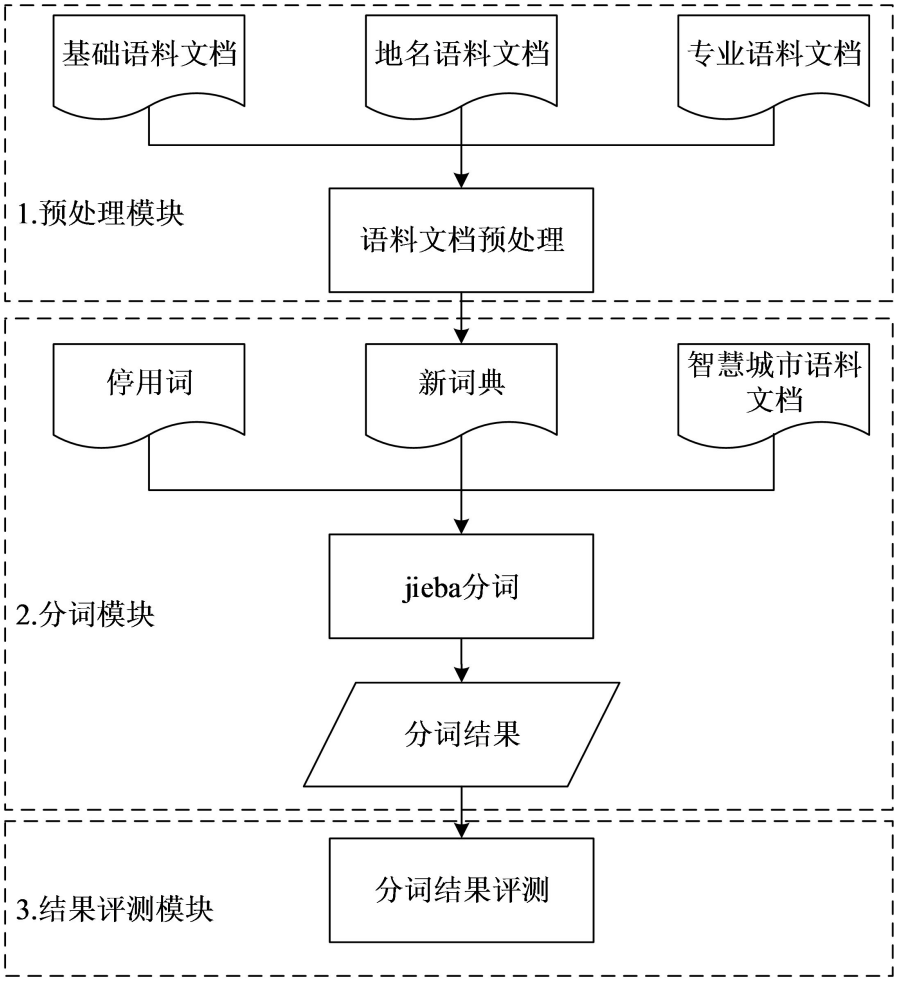


**Figure 3.** The flowchart of automatic segment model
**图 3.** 自动分词模型流程图

图(Directed Acyclic Graph, DAG)。

在智慧城市语料文档中，假设语句中包含 $k$ 个字符，对每个字符进行点的编号(0~$k$)，考虑到字符左右位置，则有 $k+1$ 个点对应的位置，根据新词典生成一个切分的词图阵列(jieba 分词的前缀词典更新速度慢，且未收录“智慧城市”、“大数据”等专业名词，因此使用新词典避免了“智慧城市”、“大数据”等专业名词被误分的情况，提高了分词的准确率)。

2) 依据第一步生成的 DAG 和在新词典中不同词组出现的次数，利用逆向最大匹配方法和动态规划查找最大概率路径法，从切分的词图阵列的右侧向左侧计算，得到最大概率的切分组合。

3) HMM 处理未登录词

使用 BEMS 四个状态对中文词汇进行描述，其中 B 表示 begin (句子开头的词汇)；E 表示 end (结尾的词汇)；M 表示 middle (句子中间的词汇)；S 表示 single (单个的词汇或字)。

对于未登录词汇，jieba 分词使用了 Viterbi 算法来求解 HMM 生成的最优状态序列。HMM 通过给定的语句序列 $P=\{P_1,P_2,\cdots,P_n\}$ 和模型参数$\gamma=(X,Y,\pi)$，找到符合特定序列的最优的隐含状态序列 $S$，其中模型参数$\gamma$经过大量语料训练得到，参数$\pi$是初始状态概率，即词语对应为 B 或 S 的开头状态的概率；$X$ 为词语的 BEMS 四种状态下位置之间在隐含状态概率的转移矩阵；$Y$ 为词语的位置状态到单个字的发射

概率。因此 HMM 是一个包含了语句序列、隐含状态序列、转移概率分布、发射概率、初始状态概率的五元组。特定序列值 $P$ 为 Viterbi 算法的输入参数，也是智慧城市语料文档中待分的未登录词，BEMS 四个状态在待分词的智慧城市语料文档句子中的位置为状态序列值，也是 Viterbi 算法的输出。

设定观察空间为 $H=\{h_1,h_2,\cdots,h_N\}$，状态空间为 $S=\{s_1,s_2,\cdots,s_K\}$，语句序列为 $P=\{P_1,P_2,\cdots,P_T\}$；$X$ 为 K × K 的转移矩阵，其中 $X_{ij}$ 为状态 $s_i$ 转移到 $s_j$ 的转移概率；$Y$ 为 K × N 的放射矩阵，其中 $Y_{ij}$ 为状态 $s_i$ 转移到 $y_j$ 的概率。路径 $Q=\{q_1,q_2,\cdots,q_T\}$ 为语句序列 $P=\{P_1,P_2,\cdots,P_T\}$ 的状态序列。

此时构建了 Viterbi 算法的两个大小为 K × T 的二维表 $T_1$，$T_2$。其中 $T_1$ 的每个元素 $T_1[i,j]$保存了生成 $P=\{P_1,P_2,\cdots,P_j\}$ 时最有可能的路径 $\hat{X}=\{\hat{x}_1,\hat{x}_2,\cdots,\hat{x}_j\}$，$\hat{x}_j=s_i$ 的概率；$T_2$ 的每个元素 $T_2[i,j]$保存了最有可能的路径 $\hat{X}=\{\hat{x}_1,\hat{x}_2,\cdots,\hat{x}_{j-1},\hat{x}_j\}$。

分词功能通过使用 jiebaR 功能包[17]，编写 R 语言脚本实现，该包是 jieba 分词方法运行在 R 语言环境中的功能包，结合停用词、新词典对智慧城市语料文档进行分词，同时 jiebaR 集成了词频－逆向文本频率算法[18]和 ICTCLAS 词性标注法[19]，在分词过程中实现了对关键词和标注词性词汇的提取。

### 4.3. 分词结果评测

分词结果与原文档内容进行比对分析，采用人工判读的方法对分词结果进行评测，本文构建的自动模型在分词速度和准确度上比未处理语料库和词典的方法分别提高了 5%和 6.7%。对于未成功分词的，通过人工判读识别，直至最终完成分词。分词结果中除了中文分词词汇以外，还包含了关键词和标注词性的词汇信息，三者分别存储在计算机内存中。

## 5. 结果与分析

### 5.1. 智慧城市研究的时序性分布

研究智慧城市的时序性分布在于分析智慧城市研究的发展历程、现状与发展趋势，探究其随时间分布的主要决定因素。关于智慧城市的时序性分布特征，有学者虽然做了研究[6] [20]，但研究文献的数量很少，且采用的智慧城市方面的文献数量、类型以及时间序列上均存在不足。本文通过提取智慧城市文献数据库的 Publish Data (发表时间)字段，通过信息的整合，得到如图 4 所示智慧城市研究的时间分布。

中国知网收录的智慧城市研究文献最早始于 2005 年，到 2008 年之前年均不超过 5 篇，智慧城市的研究较少；2008~2017 年呈现增长趋势，其中 2008~2014 年呈现快速增长，从 2008 年的 4 篇增加到 2014 年的 3925 篇(6 年期间增长了约 980 倍)，2014~2016 年稳定增长，其中 2016 年达到最大值的 4042 篇。2008 年 IBM 提出了“智慧地球”的概念，并先后与中国多个城市合作共建“智慧城市”[21]，“智慧城市”在中国兴起；智慧城市的研究与国家相关政策紧密相关，2011~2013 年随着“智慧城市指标体系 1.0”、《中国工程科技中长期发展战略研究报告》、《智慧城市时空信息云平台建设技术指南》等一系列相关政策和技术指南先后出台，和先后确定了 290 个国家级智慧城市试点城市，近 400 个城市开展了智慧城市的建设[22]，大大促进了智慧城市的研究，其中国家新型城镇化规划(2014~2020 年)中明确指出“推进智慧城市建设”的发展方针，智慧城市建设已上升为国家战略，使得智慧城市的研究文献呈现爆发式的增长。2017 年原国家测绘地理信息局颁布了《智慧城市时空大数据与云平台建设技术大纲》为智慧城市的建设提出实施标准，以及部分省份已开始智慧城市的建设，预计今后几年智慧城市的研究将持续较热。

### 5.2. 空间分布

分词结果中的标注为地名信息的词汇共 7929 条，通过去除国外的地名、将街道(乡镇)、县(区)名和市名统计到所属省或直辖市中、对于同一个文献题名中既包括了省名和市名、县(区)名的按省名出现 1

次统计、对于同一地名的情况如鼓楼区(南京市、徐州市和福州市都存在)，借助原始语料库和研究文献的具体内容识别，最终得到 7448 条地名词汇，其中包含“中国”的有 1497 条，省、自治区、直辖市的共 5951 条。结合省份、自治区、直辖市出现的次数，得到智慧城市研究的空间分布图，如图 5 所示。关于

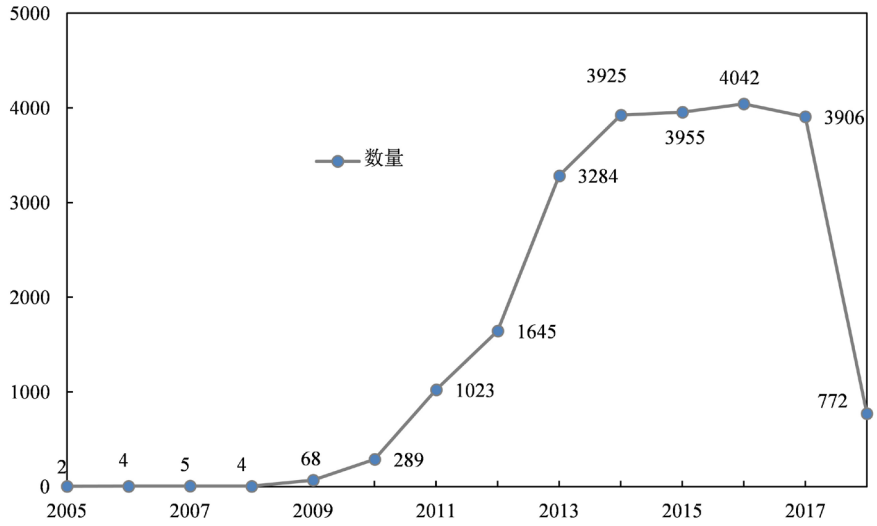


**Figure 4.** Time distribution of smart city literal
**图 4.** 智慧城市研究文献的时间分布

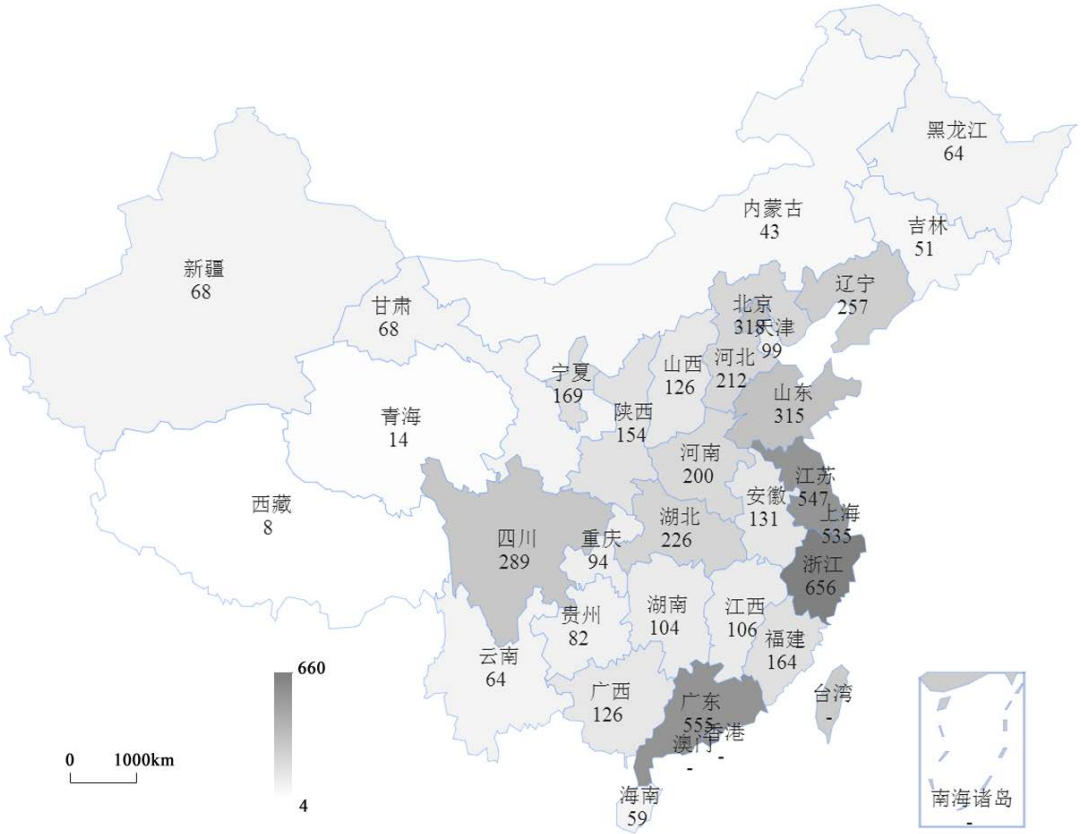


**Figure 5.** Spatial distribution of smart city literal
**图 5.** 智慧城市研究的空间分布

智慧城市研究的空间分布方面，国内的研究者更多关注智慧城市开展的区域[23] [24]和领域[25]，但在智慧城市研究的空间分布特征，尤其是在全国范围内空间差异性分布上还没有研究。

因受资料的限制，香港、澳门和台湾不在本文分析的范围。按照出现次数在 300 以上、100~300 次和 100 次以下划分三个阶梯。

处于第一阶梯的为位于长三角、珠三角、京津冀的浙江、广东、江苏、上海、北京、山东等省市，为中国经济最发达的地区，处于智慧城市研究的最前沿。IBM 公司最早与上海市，江苏省、广东省和浙江省的部分城市合作共建智慧城市，江苏省开发了全国首个省级智慧城市群综合接入平台“智慧江苏”、智慧城市指标体系 1.0 最早在上海发布、北京市和山东省先后制定了智慧城市发展规划、浙江省率先启动了 20 个智慧城市示范项目，我国先后公布了三批国家智慧城市试点名单和扩大试点名单[26] [27]，山东省和江苏省是国家级智慧城市试点最多的两个省份。

处于第二阶梯的大都位于中部地区，经济实力较强，国家级试点城市较多。

处于第三阶梯的大部分是西部省份和边疆地区、经济欠发达地区，国家级智慧城市试点城市较少。

因此从全国范围来看，智慧城市研究的现状是分布不平衡，主要表现在：① 东部发达省份出现次数最多，其次是中部省份，西部和边疆省份出现的最少；② 东部和西部省份差距较大，其中浙江省(656 次)是西藏自治区(8 次)的 82 倍；③ 区域内的差异性，东北的辽宁省出现的次数大于黑龙江和吉林两省出现之和，同为东部沿海省份的福建，出现次数仅为同地区浙江省的 1/4。

智慧城市是建设数字中国的重要组成部分[28]，作为决策层、领导层、学者层面都应重视智慧城市研究的地区间、区域内的差异。在未来的智慧城市发展中，国家应加大向中西部和边疆地区的支持力度，各级地方政府应重视智慧城市的建设对于经济社会发展带来的机遇，更好的促进智慧城市的均衡发展。

### 5.3. 研究文献的类别情况

提取智慧城市文献数据库的 Database 字段信息，按照统计分析的方法对期刊、报纸、硕士论文、博士论文、学术辑刊五个类别下的文献进行合并，得到了如图 6 所示的智慧城市文献来源分布。文献来源位列第一至第五的有：期刊(12,617 篇)、报纸(9070 篇)、硕士论文(758 篇)、会议论文(428 篇)、博士论文(39 篇)。

期刊和报纸的比例分别为 55%和 39.5%，是智慧城市研究文献的主要来源。提取智慧城市数据库的 Source 字段信息，得到收录智慧城市文献在 200 篇以上的期刊，如表 2 所示。通过对每个期刊的特点、介绍和设置栏目内容进行关键词分析，得出主要期刊收录的智慧城市方面的文献更加偏重于信息化、城市规划与建设、智能交通等领域。

### 5.4. 智慧城市研究的热点

高频词汇和关键词常被研究为研究领域的热点问题，为了便于分析，提取了分词结果中前 30 个高频词汇和关键词，如表 3 所示。“智慧城市”、“智慧”、“建设”、“城市”、“研究”位列高频词汇和关键词前列，表明智慧城市的建设，离不开“城市”这个载体、“智慧”的前提，以及“智慧城市”的最终目标；高频词汇和关键词中同时出现了“大数据”、“物联网”、“信息化”、“互联网+”等智慧城市建设的技术手段和研究内容以及“打造”、“设计”、“构建”等研究方法，与国内主流观点相一致[6]，表明用高频词汇和关键词分析智慧城市研究热点的方法是可行的。同时，排名靠前的高频词汇与关键词中没有出现与智慧城市建设主体—人相关的信息，没有突出人在智慧城市中的核心地位，在这种导向下，追求技术实现而忽视智慧城市的个性化需求，容易造成信息孤岛和单一的智慧城市建设模式，为此建设好、发展好智慧城市应以人为基础，把人的因素和信息技术结合起来[29]。

## 6. 结论

本文基于 R 语言设计的网络爬虫程序，高效的获取了中国知网收录的以智慧城市为主题的各类研究文献，构建了智慧城市文献数据库，在此基础之上构建了智慧城市自动分词模型，提取了地名信息、关键词和高频词汇，分析了智慧城市研究的时序性、空间分布特征、文献类别和研究热点，结论如下：

1) 近年来关于智慧城市的文献发表数量较大，数量与国家相关政策关联紧密，在未来的几年智慧城市将一直成为研究热点；

2) 当前智慧城市的研究在空间分布上存在差异性。经济发达的省份和地区处于智慧城市研究的前列，西部省份开展智慧城市的研究较少，同时还存在区域内的差距，展望未来智慧城市的研究与发展在空间分布上将会趋于合理；

3) 期刊和报纸是智慧城市研究文献的主要来源，主流期刊收录的智慧城市方向的研究文献更加偏向于信息化、城市规划与建设、智能交通等领域，智慧城市研究的热点主要集中在技术要素，因此在未来

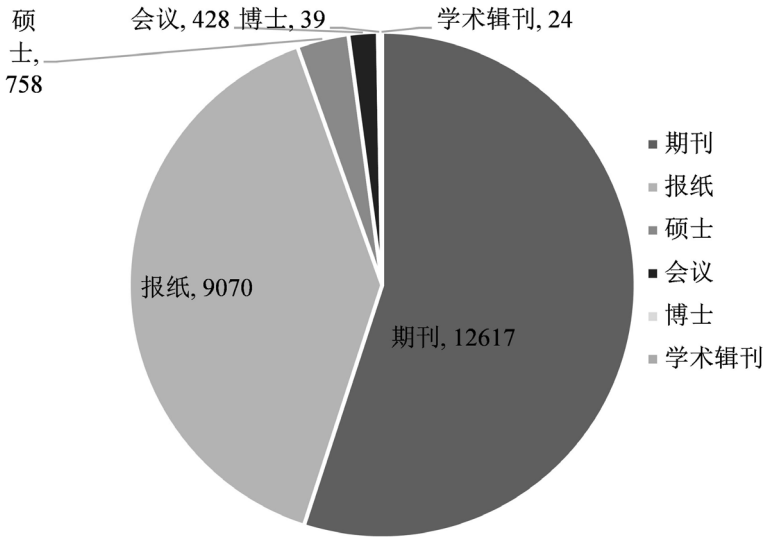


**Figure 6.** Source distribution of smart city literal
**图 6.** 智慧城市文献来源分布

**Table 2.** Journal of collection more than 200 papers
**表 2.** 收录 200 篇以上的期刊情况

| 排名 | 期刊名称 | 收录论文数 |
|---|---|---|
| 1 | 中国公众安全 | 479 |
| 2 | 中国信息界 | 466 |
| 3 | 中国建设信息化 | 313 |
| 4 | 中国建设信息 | 260 |
| 5 | 信息化建设 | 250 |
| 6 | 智能建筑与智慧城市 | 229 |
| 7 | 智能建筑 | 222 |
| 8 | 智能建筑与城市信息 | 217 |
| 9 | 中国安防 | 206 |

**Table 3.** High frequency and keywords of smart city literal
**表 3.** 智慧城市研究的高频词汇和关键词

| 排名 | 高频词汇 | 关键词 | 排名 | 高频词汇 | 关键词 |
|---|---|---|---|---|---|
| 1 | 智慧城市 | 智慧城市 | 16 | 产业 | 中国 |
| 2 | 建设 | 智慧 | 17 | 推进 | 推进 |
| 3 | 智慧 | 建设 | 18 | 智能 | 助力 |
| 4 | 发展 | 城市 | 19 | 信息化 | 新型 |
| 5 | 城市 | 大数据 | 20 | 服务 | 产业 |
| 6 | 研究 | 研究 | 21 | 新型 | 技术 |
| 7 | 中国 | 发展 | 22 | 设计 | 平台 |
| 8 | 应用 | 物联网 | 23 | 管理 | 服务 |
| 9 | 创新 | 应用 | 24 | 平台 | 设计 |
| 10 | 大数据 | 打造 | 25 | 模式 | 构建 |
| 11 | 打造 | 创新 | 26 | 时代 | 转型 |
| 12 | 技术 | 互联网+ | 27 | 生活 | 为例 |
| 13 | 信息 | 信息化 | 28 | 战略 | 社区 |
| 14 | 基于 | 智能 | 29 | 助力 | 模式 |
| 15 | 物联网 | 信息 | 30 | 互联网+ | 思考 |

的研究工作中技术要素依然占据主要，对人等非技术要素的关注将趋于增多。

本文设计的网络爬虫程序在获取智慧城市研究文献上具有可行性和有效性。通过分析智慧城市的时序性分布特征、空间分布特征、文献类别、研究热点揭示了智慧城市研究的发展历程、现状、研究热点，并展望了发展趋势，研究成果可为政府部门在智慧城市研究、建设方面提供决策支持。

## 基金项目

国家自然科学基金(No. 41771366)。

## 参考文献

[1] Robert, W. and Siegfried, R. (2018) The Governance of Smart Cities: A Systematic Literature Review. Cities.

[2] 刘晓丽, 李成名, 印洁. 智慧城市时空信息云平台评价指标体系研究[J]. 测绘通报, 2017(3): 38-41.

[3] 李德仁, 柳来星. 上下文感知的智慧城市空间信息服务组合[J]. 武汉大学学报(信息科学版), 2016, 41(7): 853-860.

[4] 王朝晖, 郑新奇. 基于共词分析的智慧城市研究现状与展望[J]. 地域研究与开发, 2014, 33(4): 59-63.

[5] Hashem, I.A.T., Chang, V., Anuar N.B., *et al.* (2016) The Role of Big Data in Smart City. *International Journal of Information Management*, **36**, 748-758. https://doi.org/10.1016/j.ijinfomgt.2016.05.002

[6] 楚金华. 我国智慧城市建设研究述评[J]. 现代城市研究, 2017(8): 115-120.

[7] Widenius, M. (2002) Mysql Reference Manual. O'Reilly & Associates, Inc.

[8] Munzert, S., Rubba, C., Meissner, P., *et al.* (2015) Automated Data Collection with R. A Practical Guide to Web Scraping and Text Mining.

[9] Broucke, S.V. and Baesens, B. (2018) From Web Scraping to Web Crawling. Practical Web Scraping for Data Science.

[10] 韩毅, 伍玉, 申东阳, 等. 中文科研论文未被引探索 II: 基于关键词的内容因素影响研究——以图书馆情报与文

献学为例[J]. 图书情报工作, 2018(4): 14-20.

[11] Sun Junyi. Jieba 中文分词[EB/OL]. https://github.com/fxsjy/jieba/tree/master, 2018.

[12] 于重重, 操镭, 尹蔚彬, 等. 吕苏语口语标注语料的自动分词方法研究[J]. 计算机应用研究, 2017, 34(5): 1325-1328.

[13] Ted Kwartler. (2017) Text Mining in Practice with R. John Wiley & Sons, Ltd., Chichester, UK. https://doi.org/10.1002/9781119282105

[14] 国家测绘局地名研究所. 中国地名录[M]. 北京: 中国地图出版社, 1997.

[15] 民政部全国行政区划查询平台[EB/OL]. http://xzqh.mca.gov.cn/map

[16] Joung, J. and Kim, K. (2017) Monitoring Emerging Technologies for Technology Planning Using Technical Keyword Based Analysis from Patent Data. *Technological Forecasting & Social Change*, **114**, 281-292. https://doi.org/10.1016/j.techfore.2016.08.020

[17] Qin, W. (2015) jiebaR: CRAN Version 0.4.

[18] Qin, P., Xu, W. and Guo, J. (2016) A Novel Negative Sampling Based on TFIDF for Learning Word Representation. Elsevier Science Publishers B. V., New York.

[19] Li, X. and Zhang, C. (2013) Research on Enhancing the Effectiveness of the Chinese Text Automatic Categorization Based on ICTCLAS Segmentation Method. *IEEE International Conference on Software Engineering and Service Science*, Beijing, 23-25 May 2013, 267-270.

[20] Liu, Z. (2014) Origin and Development of Smart Cities. China Standardization, 44-45.

[21] 牛文元. 智慧城市是新型城镇化的动力标志[J]. 中国科学院院刊, 2014, 9(1): 34-41.

[22] 宋资勤, 刘影, 张涛. 构建绿色智慧城市的关键技术探讨[J]. 可持续发展, 2016, 6(3): 208-215.

[23] 安爽. 智慧城市空间发展趋势与城市规划应对研究[C]//2015 中国城市规划年会. 2015.

[24] 张建伟, 李贝歌, 毕东方, 等. 中国智慧城市发展水平空间差异研究[J]. 世界地理研究, 2017, 26(2): 82-90.

[25] 蓝荣钦, 王家耀. 智慧城市空间信息基础设施支撑力评价体系研究[J]. 测绘科学技术学报, 2015(1): 78-81.

[26] 住房城乡建设部公布首批国家智慧城市试点名单[EB/OL]. http://www.mohurd.gov.cn/wjfb/201308/t20130805_214634.html, 2013.

[27] 住建部公布 2013 年度国家智慧城市试点名单[EB/OL]. http://www.mgov.cn/complexity/info1308.htm, 2013.

[28] 于慧. 新型智慧城市成数字中国重要内容[N]. 人民邮电, 2018.

[29] 成思危. 广义智慧城市导论[M]. 北京: 人民出版社, 2016.