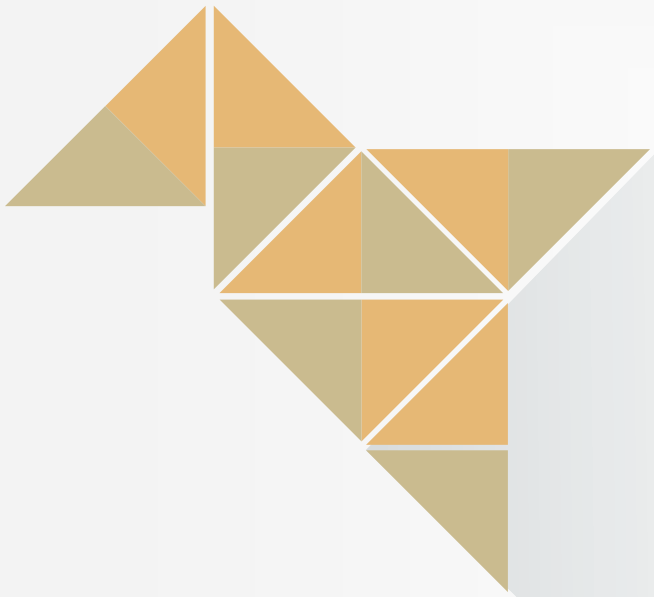
基于数据挖掘的物联网学术研究热点与趋势分析

成员：张晓宇 刘懋霖



1 项目背景

2 数据获取

3 摘要分析

4 关键词聚类分析

5 最新热点分析

6 结语

1

项目背景

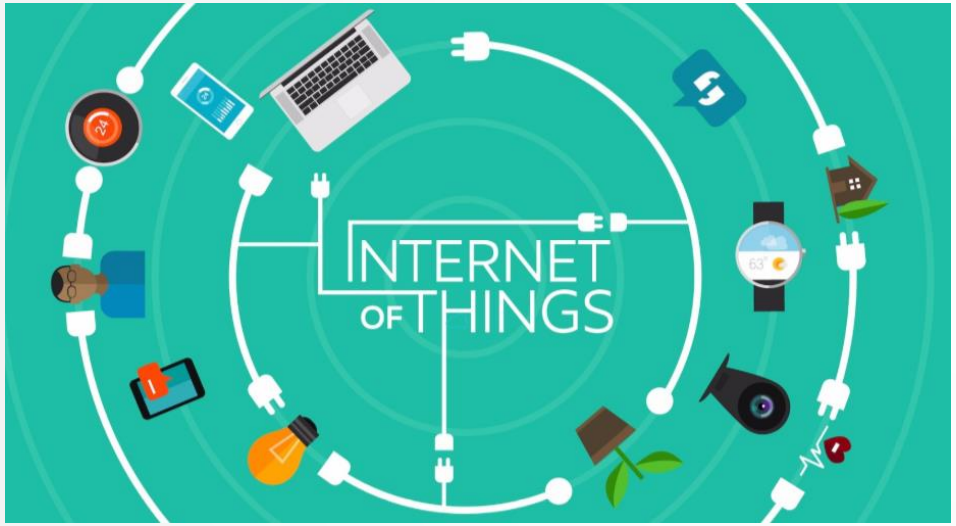
项目背景

文本数据挖掘（text data mining，TDM）是以**非结构化的语言文本**为挖掘样本，常指为了发现知识而利用数据挖掘技术**从大规模的文本集中提出隐含的、未知的、潜在有价值的信息**的过程。文本数据挖掘的对象主要包括在线新闻、科研论文、公司档案、医疗记录等电子化信息。目前，文本数据挖掘作为信息时代 的重要研究领域，逐渐成为国内外学者的重点研究方向。

物联网（The Internet of Things，简称IOT）是指通过各种信息传感器、射频识别技术、全球定位系统、红外感应器、激光扫描器等各种装置与技术，实时采集任何需要监控、连接、互动的物体或过程，采集其声、光、热、电、力学、化 学、生物、位置等各种需要的信息，通过各类可能的网络接入，实现物与物、物与人的泛在连接，实现对物品和过程的智能化感知、识别和管理。物联网是一个**基于互联网、传统电信网等的信息承载体**，它让所有能够被独立寻址的普通物理对象形成互联互通的网络。

项目背景

近年来，随着物联网技术的飞速发展，学术领域对物联网的关注和研究也越来越多，本文通过文本挖掘的方法探寻国内物联网的研究热点和趋势，以期为物联网的后续研究提供借鉴。



2

数据获取

1 数据来源

2 数据收集

3 数据预处理

数据来源

数据来源：Web of Science
检索关键词：“物联网”
时间段：2015年至2019年，2020年

数据收集

使用后羿采集器工具爬取网站，选取文献标题、年份、摘要、关键词等字段。

由于采集关键词无法在同一个页面完成，故使用循环生成page=1-30&doc=1-1498的查询，进入包含论文详细信息的页面进行爬取，获得数据1498条。

读取数据，部分数据如下所示：

```
#读取数据
iot_data<-read.csv("iot_data_new.csv",stringsAsFactors = FALSE,na.strings = "")
iot_title_keywords<-iot_data[,c(2,7,3)]
iot_title_abstract<-iot_data[,c(2,5,3)]
```

A	B	C	D	E	F	G	H	I
title_en	title_ch	year	abstract_er	abstract_ch	keywords_e	keywords_ch		
Efficient co	面向物联网	2020	Internet of	摘要:物联网	作者关键词	作者关键词:物联网通信; 盲信		
Preface of	物联网安全	2020	The concep	摘要:物联网的概念已		作者关键词:物联网; 安全技术;		
A Survey o	物联网认证	2020	The large-s	摘要:物联网	作者关键词	作者关键词:物联网; 认证协议;		
Research c	基于Augur	2020	With the ra	摘要:随着物	作者关键词	作者关键词:区块链; 身份管理;		

数据预处理

数据预处理是指在进行数据分析之前，对数据进行的一些初步处理，包括缺失值处理、噪声处理、不一致数据修正等，其目标是得到更标准、高质量的数据，纠正错误异常数据，从而提升分析的结果。

1. **定位数据中字段错位的条目，对照原网页手动校正**
2. **去除数据集中校对过后中文部分仍然缺失的条目**
3. **将摘要与关键词两个字段分别进行分词处理，存储到两张表中**。该操作将分别在摘要分析与关键词分析中进行详细介绍。

```
#数据清洗：去除标题、摘要缺失的条目
x<-is.na(iot_title_abstract$title_ch)
iot_title_abstract<-iot_title_abstract[!x,]
x<-is.na(iot_title_abstract$abstract_ch)
iot_title_abstract<-iot_title_abstract[!x,]
```

3

摘要分析

1 结巴分词

2 词频统计

3 摘要关键词提取

4 词云图展示

5 物联网领域近五年热点与趋势分析

结巴分词

“结巴”（Jieba）工具是最常用的中文文本分词和处理的工具之一，它能实现中文分词、词性标注、关键词抽取、获取词语位置等功能。

“结巴”中文分词的R语言版本，支持最大概率法，隐式马尔科夫模型，索引模型，混合模型，共四种分词模式，同时有词性标注，关键词提取，文本Simhash相似度比较等功能。

停用词是指在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言数据（或文本）之前或之后会自动过滤掉某些字或词，这些字或词即被称为Stop Words（停用词）。

包的下载与安装：

```
>install.packages('jiebaRD')

>install.packages('jiebaR')

> library(jiebaRD)
> library(jiebaR)
```

分词

```
> test <- '革命尚未成功，同志仍需努力！'
首先需要建立分词引擎
> seg<-worker()
这里“<=”表示分词运算符
> seg<=test
 [1] "革命" "尚未" "成功" "同志" "仍" "需" "努力"
与下面这句效果一样
> segment(test,seg)
 [1] "革命" "尚未" "成功" "同志" "仍" "需" "努力"
也就是有两种写法：
```

结巴分词

1. 采用R语言中的“jiebaR”包提供的混合模型
2. 使用来源于网络的IT行业词典（IT_dict.txt）以及常用停用词表（stop_word.txt）

```
library(jiebaR)
wk<-worker(user = "IT_dict.txt" ,stop_word = "stop_word.txt")
```

单段文字分词结果如下所示：

```
> wk[text]
  [1] "物联网"    "通信系统"  "活跃"      "用户数"    "低"        "数据帧"    "短"        "特性"
  [9] "信道"      "估计"      "多用户"    "识别"      "需"        "导频"      "用户"      "识别码"
 [17] "降低"      "系统"      "通信"      "效率"      "响应速度"  "提出"      "一种"      "非"
 [25] "正交"      "多址"      "接入"      "盲"        "信道"      "估计"      "多用户"    "检测"
 [33] "算法"      "利用"      "码分多址"  "系统"      "中"        "扩频"      "矩阵"      "用户"
 [41] "占用"      "载波"      "分配"      "采用"      "差分"      "编码"      "解决"      "盲"
 [49] "估计"      "引发"      "星座"      "点"        "旋转"      "用户"      "分配"      "载波"
 [57] "稀疏"      "性"        "引入"      "伯努利"    "高斯"      "分布"      "先验"      "分布"
 [65] "利用"      "变量"      "间"        "隐"        "马尔可夫"  "特性"      "因式分解"  "建模"
 [73] "用户"      "数据"      "稀疏"      "特征"      "多用户"    "识别"      "消息传递"  "算法"
```

词频统计

在一份给定的文件里，词频（term frequency，TF）指的是某一个给定的词语在该文件中出现的次数。这个数字通常会被正规化，以防止它偏向长的文件。在对文献摘要进行分词后，需要对其进行词频统计，以便于后续操作。

使用整洁数据原则是一种更容易、更有效的数据处理方式，在处理文本时也是如此。
整洁的数据集易于操作，建模和可视化，并具有特定的结构：

1. 每个变量都是一列
2. 每个观察值都是一行
3. 每种类型的观察单位是一张表格

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

Table 1: Typical presentation dataset.

	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

Table 2: The same data as in Table 1 but structured differently.

Tidy Data

person	treatment	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

词频统计

1. 将分词结果汇总统计后，每行数据表示文献标题，包含分词结果，词频，文献所属年份。
2. 在对每篇文献的摘要进行分词处理后，需统计分词结果的词频并按照整洁文本格式进行存储。
3. 整理过后的数据共包含81392个词项，部分数据如下图所示：

title	char	freq	year
面向物联网环境的高效通信接收机设计	分布	2	2020
面向物联网环境的高效通信接收机设计	高斯	1	2020
面向物联网环境的高效通信接收机设计	伯努利	1	2020
面向物联网环境的高效通信接收机设计	性	1	2020
面向物联网环境的高效通信接收机设计	稀疏	4	2020

摘要关键词提取

$$TF_w = \frac{\text{在某一类中词条}w\text{出现的次数}}{\text{该类中所有的词条数目}}$$

$$IDF = log(\frac{\text{语料库的文档总数}}{\text{包含词条}w\text{的文档数}+1}),$$

$$TF-IDF = TF * IDF$$

TF-IDF（term frequency–inverse document frequency，词频-逆向文件频率）是一种统计方法，用以评估一个字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

TF-IDF的主要思想是：如果某个单词在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。

摘要关键词提取

采用计算TF-IDF值的方式从分词结果中提取关键词：

1. 将所有目标文献的分词结果作为语料库，对每个分词结果计算其TF-IDF值。
2. 将从每篇文献中提取其摘要分词结果中TF-IDF值最高的五个词作为代表该文献方向的关键词。

```
#tf-idf值计算
library(tidytext)
title_words_tf_idf<-bind_tf_idf(tbl = title_words,term = char,document = title,n=freq)

#筛选每篇文章分词结果中tf-idf值最高的五个词作为文献摘要的关键词
library(sqldf)
titles<-sqldf("select distinct title from title_words_tf_idf")
abstract_keywords<-data.frame(title=NULL,'char'=NULL,'freq'=NULL,'year'=NULL,'tf'=NULL,'idf'=NULL,'tf_idf'=NULL)
for(i in titles$title){
  temp<-subset(title_words_tf_idf,title_words_tf_idf$title==i)
  temp<-temp[order(temp$tf_idf,decreasing = TRUE),]
  abstract_keywords<-rbind(abstract_keywords,temp[1:5,])
}
```

摘要关键词提取

在计算过程中发现如下问题：

1. 大多数英文专有名词或简写在中文摘要的分词结果中词频极低，并且常在同一篇摘要中多次出现。大部分英文专有名词的TF-IDF值极高。然而这些英文简写并不能作为区分各领域的关键词，严重干扰摘要关键词的筛选。
2. 英文简写多是技术名称或是对于中文词汇的补充解释，去除英文条目对于各领域关键词提取的影响较小，故去除分词结果中的所有英文词项。

最终分词、计算、筛选后的部分结果如下图所示：

title	char	freq	year	tf	idf	tf_idf
面向物联网环境的高效通信接收机设计	多用户	4	2020	0.034782609	5.205379	0.18105667
面向物联网环境的高效通信接收机设计	稀疏	4	2020	0.034782609	4.886926	0.16998002
面向物联网环境的高效通信接收机设计	估计	4	2020	0.034782609	3.673903	0.12778793
面向物联网环境的高效通信接收机设计	盲	2	2020	0.017391304	5.898527	0.10258307
面向物联网环境的高效通信接收机设计	载波	2	2020	0.017391304	5.338911	0.09285062
物联网安全技术专栏序言(中英文)	阶段	3	2020	0.050847458	3.125938	0.15894599

词云图展示

词云图，是对文本中出现频率较高的“关键词”予以视觉化的展现，词云图过滤掉大量的低频低质的文本信息，使得浏览者只要一眼扫过文本就可领略文本的主旨。

选取每年TF-IDF值最高的100个词分别绘制词云图，直观展示2015-2019年每年物联网相关文献中的研究热点。

```
#绘制2015-2020词云图,每年选取tf-idf值top100的词绘制
library(wordcloud2)
for(i in 2015:2020){
  abstract_keywords_by_year<-subset(abstract_keywords,abstract_keywords$year==i)
  wordcloud_data<-sqldf("select char,sum(tf_idf) as tf_idf from abstract_keywords_by_year group by char")
  wordcloud_data_sorted<-wordcloud_data[order(wordcloud_data$tf_idf,decreasing = TRUE),]
  print(wordcloud2(wordcloud_data_sorted[1:100,],size = 0.5))
}
```

词云图展示

由词云图，可以直观地看到：

词的TF-IDF值越高，显示尺寸越大。

近五年来物联网应用领域十分广泛，比较热门的有认证、物流、搜索、电力、水力、农业、环境、制造、医疗、智慧城市、区块链等领域。

Figure 1 2015

Figure 2 2016

Figure 3 2017

Figure 4 2018

Figure 5 2019

Figure 6 2020

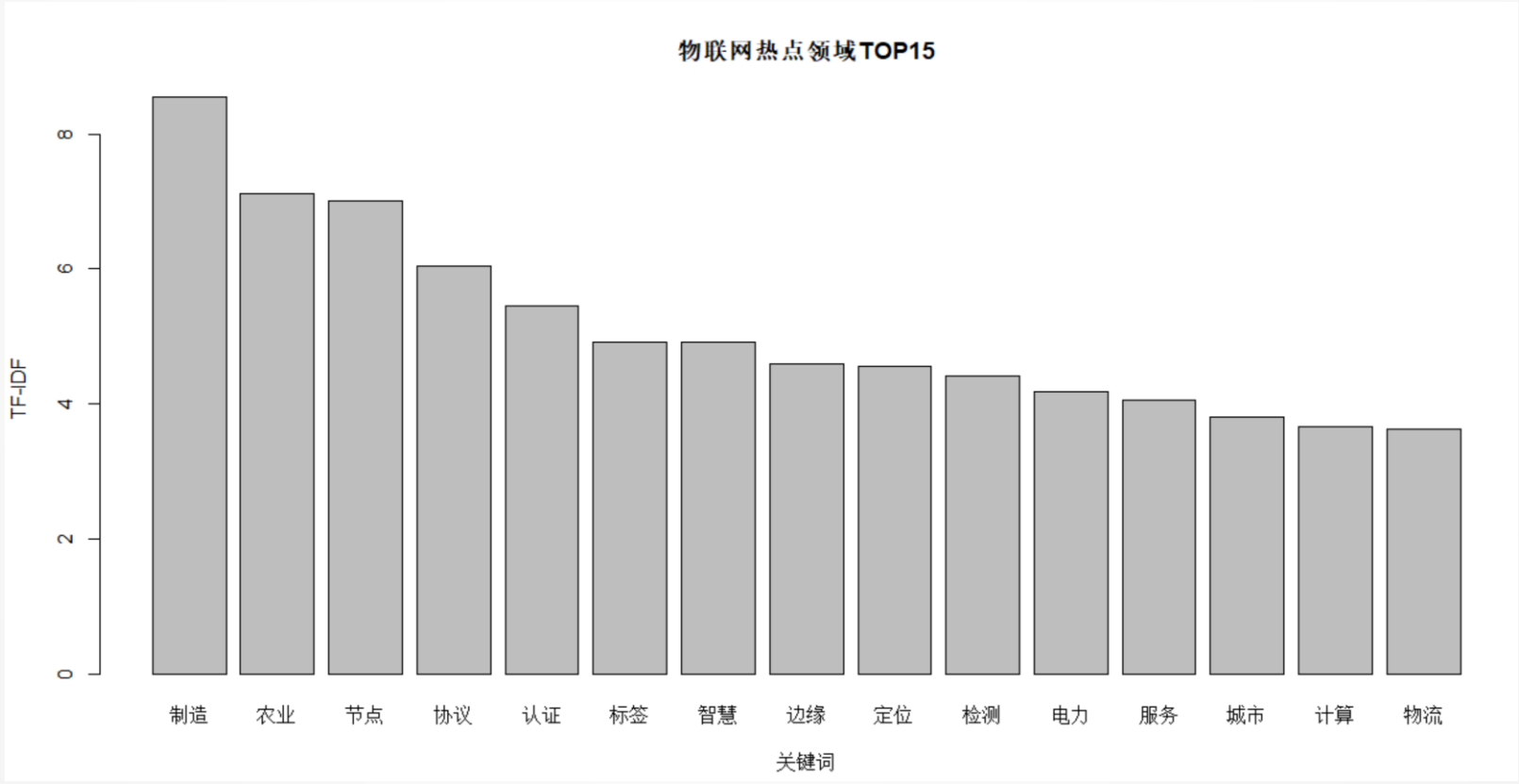
物联网领域近五年热点与趋势分析

1. 拟筛选近五年文献摘要的分词结果中TF-IDF值最高的15个词来代表15个最热门的领域。
2. 使用筛选出的15个代表词分别去匹配包含这些代表词的文献，根据匹配到的文献标题及摘要来人工标注这15个领域。
3. 标注完成后，根据每年包含这些热点词的文献数目绘制折线图，使用包含某一领域关键词的文献数目作为该领域的热度，对近五年的热点及其变化趋势简要分析。

```
#汇总分词结果及其tf-idf值，选取top15作为领域代表词
all_tf_idf<-sqldf("select char,sum(tf_idf) as tf_idf from abstract_keywords gr
field_keywords<-all_tf_idf[order(all_tf_idf$tf_idf,decreasing = TRUE),]
field_keywords<-field_keywords[1:15,]
barplot(field_keywords$tf_idf,names.arg = field_keywords$char,xlab = "关键词",y

#使用领域代表词去匹配文章，人工定义所属领域名称
field_define<-sqldf("select abstract_keywords.char,abstract_keywords.title,yea

#为五年文献标注领域：
five_year_data<-subset(field_define,field_define$year!=2020)
keyword_field<-data.frame("char"=field_keywords$char,"field"=c("智能制造","农业
keyword_field<-merge(five_year_data,keyword_field,by="char")
article_count<-sqldf("select field,count(*) as article_num,year from keyword_f

#绘制折线图
library(ggplot2)
p<-ggplot(data=article_count, aes(x=year, y=article_num, group=field,color=fie
```

物联网领域近五年热点与趋势分析

提取15个领域代表词，并绘制直方图如下所示：



物联网领域近五年热点与趋势分析

使用上述15个关键词去匹配包含这些关键词的文献，部分数据如下图所示。

char	title
制造	一种物联网环境下的制造资源配置及信息集成技术研究
制造	中国新能源装备智造化发展技术路线图研究
制造	互联网+时代下智能制造技术在我国钢铁行业的应用
制造	人工智能视角下的智能制造前世今生与未来

物联网领域近五年热点与趋势分析

查阅相关资料，根据每个关键词所包含的文献，人工标注物联网热门领域结果如下：

- 制造——智能制造
- 农业——农业
- 节点——算法设计
- 协议、认证——物联网通信
- 标签——电子标签
- 智慧——智慧型社会建设
- 边缘——边缘计算
- 定位——物联网定位技术
- 检测——信息安全
- 电力——泛在电力物联网
- 服务——智能服务系统
- 城市——智慧城市
- 计算——边缘计算/算法设计
- 物流——物流

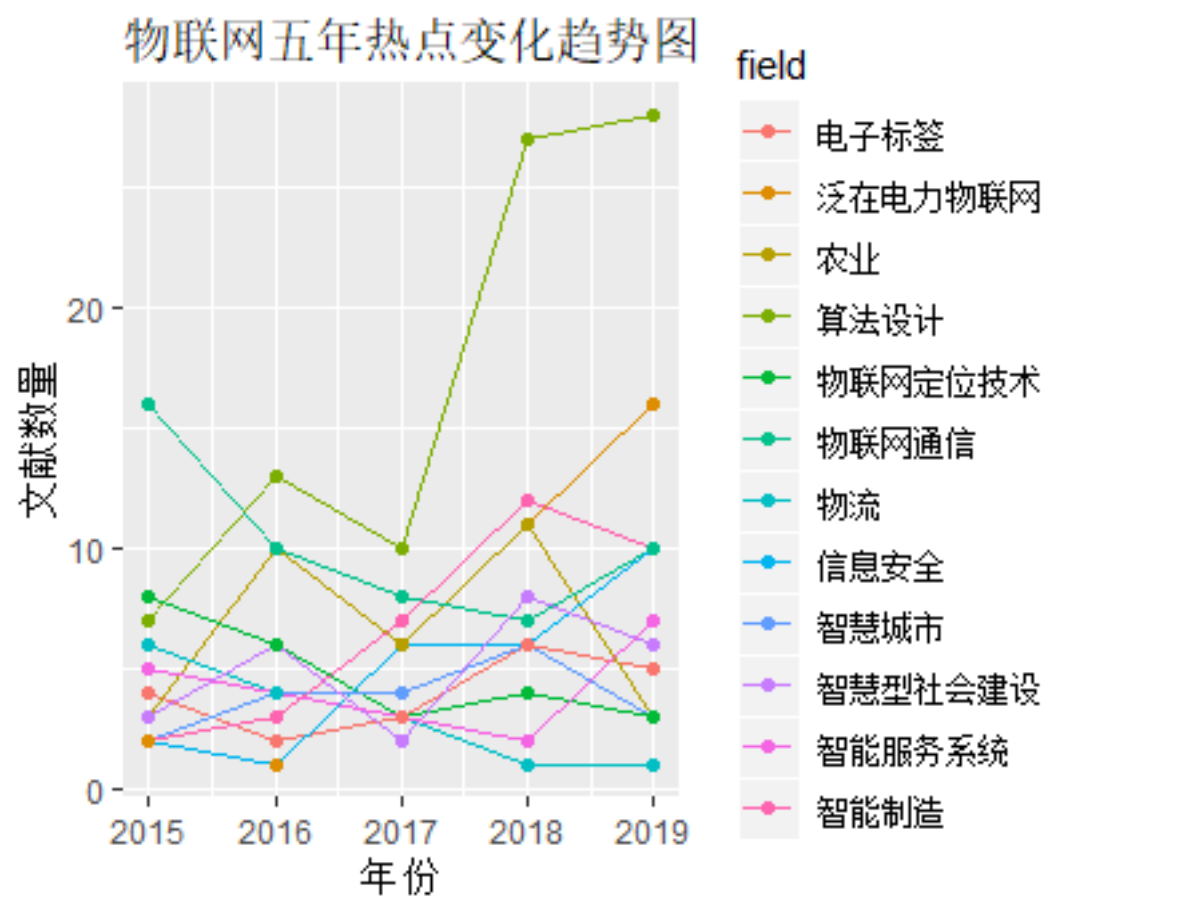
由于分词使用的词典并不是物联网行业的词典，分词结果中会出现物联网行业的专有名词被拆分的现象

↓

将存在交叉的关键词简单整理

物联网领域近五年热点与趋势分析

根据上述领域标注，绘制2015-2019物联网各领域文献数量变化折线图如下所示：



1. 目前物联网行业中最热门的研究领域是物联网相关的算法设计领域。
2. 泛在电力物联网领域正逐年发展，在2019年成为物联网行业第二大热门领域。
3. 电子标签、物联网定位技术、物联网通信、信息安全、智慧城市、智慧型社会建设、智能服务系统、智能制造等行业随热度曲线有所波动，但总体发展平稳。
4. 农业、物联网定位技术、物流变化趋势较为平坦，在后两年热度下降，有趋于饱和的迹象。
5. 物联网在各个行业的应用已经较为成熟且普遍，但物联网相关的新兴技术以及顶尖应用尚处于快速发展的阶段。

4

关键词聚类分析

层次聚类

层次聚类算法又称为树聚类算法，它根据数据之间的距离，透过一种层次架构方式，反复将数据进行聚合，创建一个层次以分解给定的数据集。在层次聚类中，每一个观测值自成一类，这些类每次两两合并，直到所有的类被聚成一类为止。

采用数值0表示该文献不包含该关键词，采用数值1表示该文献包含该关键词。基于此数据特征，采用平均联动的方法来进行层次聚类。

```
#聚类实现
library(NbClust)
devAskNewPage(ask = TRUE)
rownames(data_wide2)<-data_wide2$number
data_wide3<-data_wide2[,-1]
data_wide3.scaled<-scale(data_wide3)
nc<-NbClust(data_wide3.scaled[1:10,1:5],distance = "euclidean",min.nc = 2,max.nc = 15
table(nc$Best.n[1,])
barplot(table(nc$Best.n[1,]),xlab = "Number of Clusters",ylab = "Number of Criteria",

d<-dist(data_wide2,method="euclidean") #计算矩阵距离
fit <- hclust(d, method="average")  #层次聚类算法
plot(fit)
```

变量聚类分析

- 源表的每个关键词字段包含了3-8个数量不等的关键词，
- 以分号为分隔符拆分每个关键词字段后再将单个关键词与文献标题匹配，构成标题-词项二元组的整洁数据格式。
- 由于每个文献与关键词之间只存在包含关系，故将所有的关键词词频设为1。
- 整理后部分数据如下图所示。

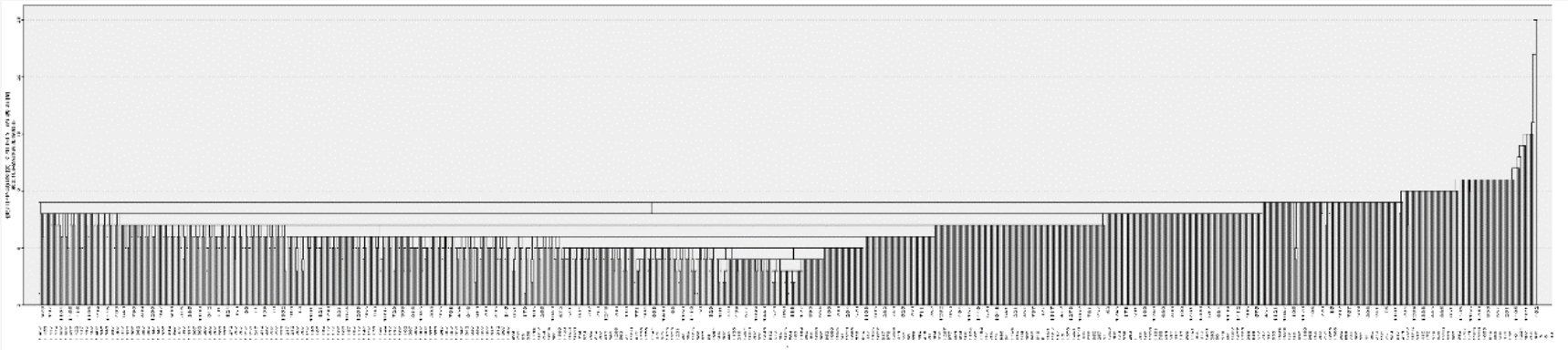
title	keywords	year	freq
面向物联网环境的高效通信接收机设计	物联网通信	2020	1
面向物联网环境的高效通信接收机设计	盲信道估计	2020	1
面向物联网环境的高效通信接收机设计	概率图模型	2020	1
面向物联网环境的高效通信接收机设计	迭代接收机设计	2020	1
面向物联网环境的高效通信接收机设计	非正交多址接入	2020	1

全变量聚类分析

计算关键词的TF-IDF值并构建文献-关键词的TF-IDF矩阵。
计算整合后部分数据如下图所示:

number	year	迭代接收机设计	非正交多址接入	概率图模型	盲信道估计	物联网通信	安全技术	物联网	专栏
1	2020	1.456964	1.456964	1.456964	1.456964	1.318335	*NA*	*NA*	*NA*
2	2020	*NA*	*NA*	*NA*	*NA*	*NA*	2.428274	0.2687704	2.428274

按文献其关键词进行聚类，最终生成聚类系谱图如下所示：



全变量聚类分析

从分析方法来考虑，文献与作者关键词的关系仅为二分类，关键词的区分度将仅仅取决于关键词在语料库中的词频，而与文献本身无关，导致TF-IDF值无法代表某一关键词对于文献的重要程度，从而使文献之间的相似度计算出现较大误差。

从物联网的行业特性来考虑，物联网是基于IT行业，却不局限于某一特定领域。物联网是新一代信息技术的重要组成部分，IT行业又叫：泛互联，意指物物相连，万物万联。因此，物联网涉及的领域太过广泛，并不仅仅包含计算机行业，任何行业都有可能涉及。然而不同行业间的差异较大，其关键词甚至可能完全不同，导致这些文献虽同属物联网领域，但实际上文本相似度并不高。

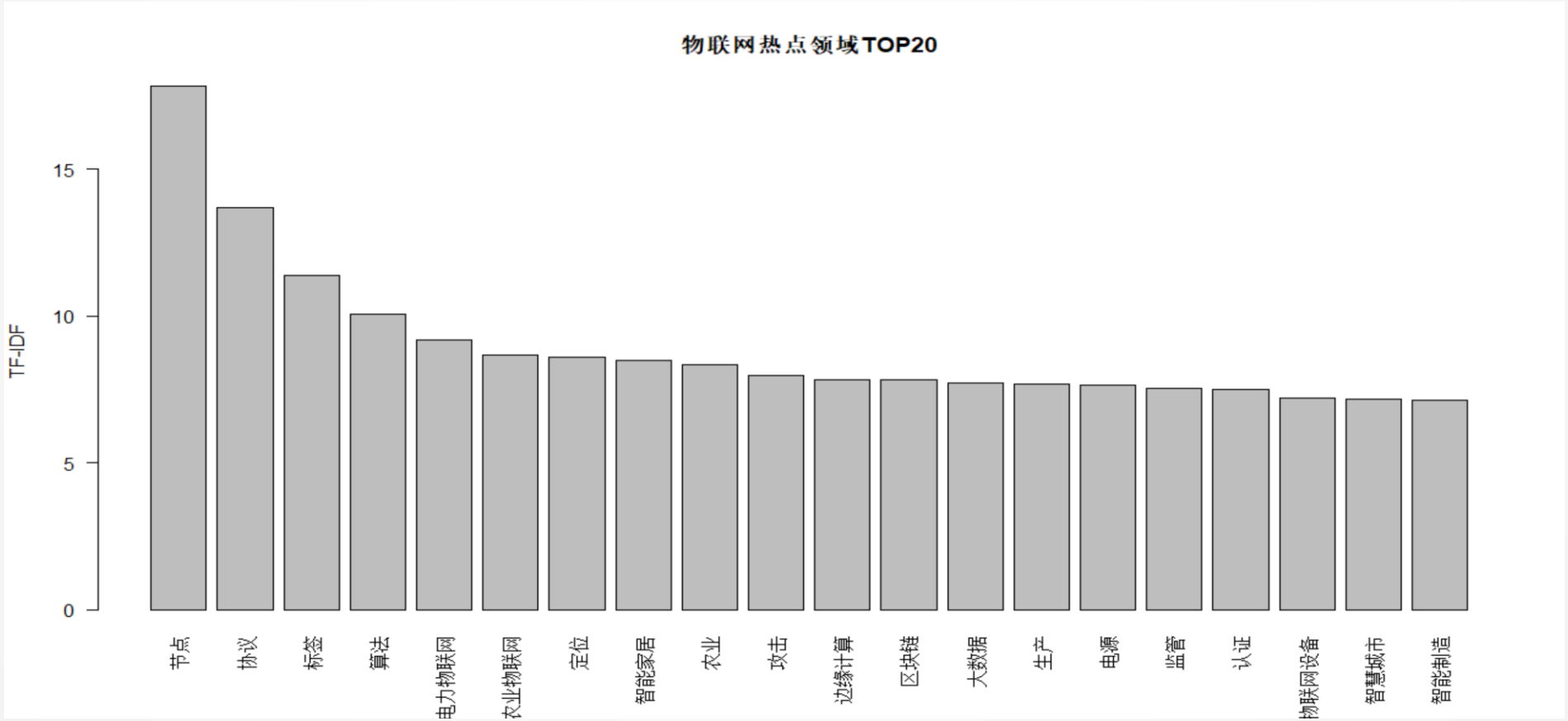
关键词标注分析

- 分析上述造成聚类效果不佳的因素，其中文本相似度不高的问题属于物联网领域本身的特性，无法改变。因而着重解决关键词的区分度不足的问题。
- 为使得不同关键词在同一文献中的重要程度有所区分，考虑使用关键词在文献摘要中的出现频次作为新的词频，改变文献-词项的二分类关系后再进行TF-IDF值的计算。
- 主要方法是将作者关键词的分词结果追加到原有词典上重新进行分词，保证关键词不会被拆分，同时计算新的TF-IDF来代表关键词的重要程度后再进行分析。

```
#筛选关键词TOP20
#先从每篇文章中筛选TF-IDF值最高的3个词作为关键词
titles<-sqldf("select distinct title from new_title_words_tf_idf")
new_3_keywords<-data.frame(title=NULL,'char'=NULL,'freq'=NULL,'year'=NULL,'tf'=NULL,'idf'=NULL,'tf_idf'=NULL)
for(i in titles$title){
  temp<-subset(new_title_words_tf_idf,new_title_words_tf_idf$title==i)
  temp<-temp[order(temp$tf_idf,decreasing = TRUE),]
  new_3_keywords<-rbind(new_3_keywords,temp[1:3,])
}

new_sum_tf_idf<-sqldf("select keywords,sum(tf_idf) as tf_idf from new_3_keywords group by keywords")
top_keywords<-new_sum_tf_idf[order(new_sum_tf_idf$tf_idf,decreasing = T),][1:20,]
barplot(top_keywords$tf_idf,names.arg = top_keywords$keywords,ylab = "TF-IDF",main = "物联网热点领域TOP20",las=2)

#使用筛选出的关键词去匹配文档
cluster_article<-sqldf("select title,top_keywords.keywords,new_cluster_keywords.tf_idf,year from top_keywords,new_cluster_keywords,connec
#cluster_article<-sqldf("select number,top_keywords.keywords,new_cluster_keywords.tf_idf from top_keywords,new_cluster_keywords where top

#为五年文献标注领域：
five_year_data<-subset(cluster_article,cluster_article$year!=2020)
keyword_field<-data.frame("char"=top_keywords$keywords,"field"=c("算法设计","认证协议","电子标签","算法设计","泛在电力物联网","农业物联网","物
keyword_field<-merge(five_year_data,keyword_field,by.y = "char",by.x = "keywords")
article_count<-sqldf("select field,count(*) as article_num,year from keyword_field group by field,year")

#绘制折线图
library(ggplot2)
```

关键词标注分析

筛选出TF-IDF较高的20个词汇，将其作为热点关键词进行分析，筛选结果如下所示：



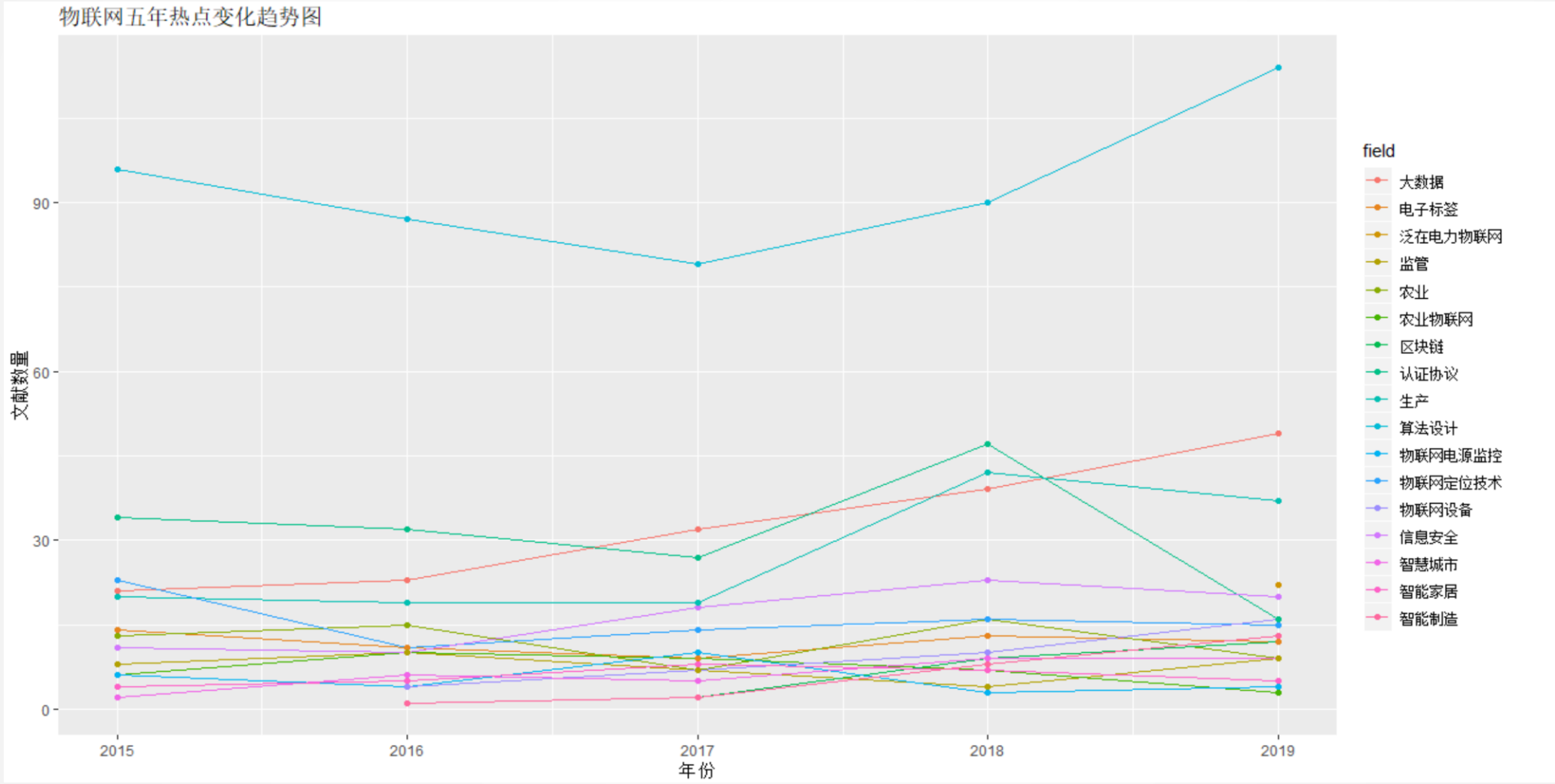
关键词标注分析

对上述20个关键词进行人工标注，得到关键词——代表领域关系如下：

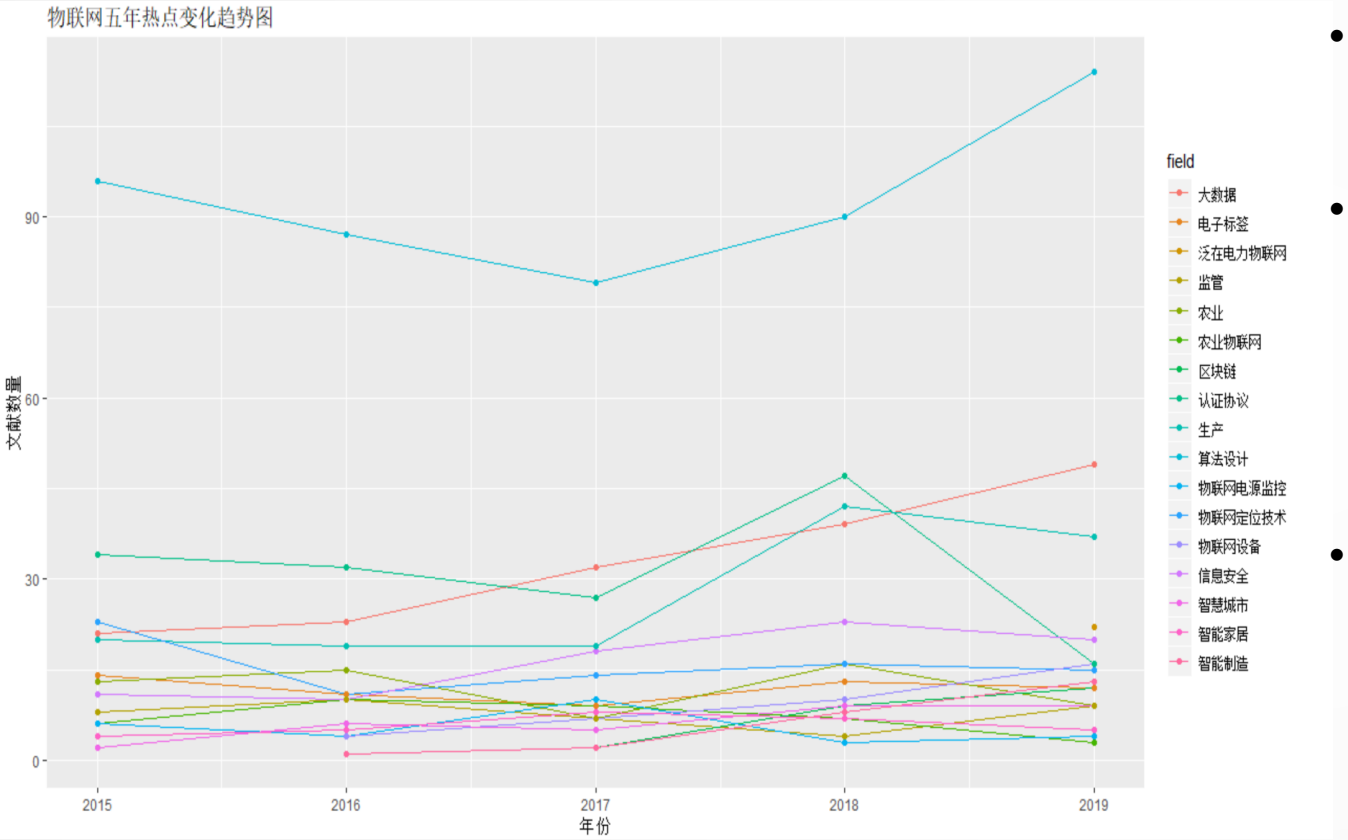
- 节点——算法设计
- 协议——认证协议
- 标签——电子标签
- 算法——算法设计
- 泛在电力物联网
- 农业物联网
- 定位——物联网定位技术
- 智能家居
- 农业
- 攻击——信息安全
- 边缘计算——算法设计
- 区块链
- 大数据
- 生产
- 电源——物联网电源监控
- 监管
- 认证——认证协议
- 物联网设备
- 智慧城市
- 智能制造

关键词标注分析

绘制2015-2019各领域文献量变化曲线如下：



关键词标注分析



- 算法设计领域是物联网行业中最热门的领域。
- 大数据、区块链、生产业以及信息安全领域的热门程度明显高于其他领域。
- 其中，物联网中的大数据领域呈现增长趋势，区块链、生产业以及信息安全领域的发展较为平稳。其余行业涉及面较为广泛，且发展趋势同样较为平稳，没有显著的特征与波动。
- 就热点变迁而言，有关物联网的算法设计领域热度居高不下。除此之外，区块链领域前四年逐年上升，在第五年热度降低，低于生产业。大数据行业热度逐年上升，在第五年超越区块链与生产业，位居热度第二。信息安全领域稳居热点第五，逐年发展趋势平稳。

5

最新热点分析

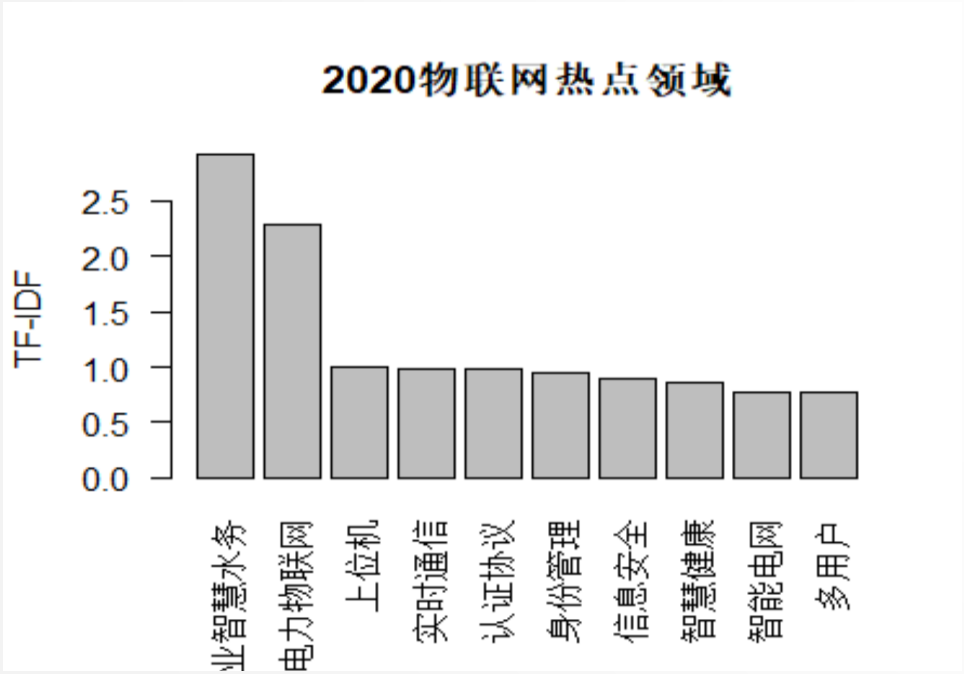
最新热点分析

抽取2020年的文献及其分词、关键词数据，进行最新热点的分析。

```
#最新热点分析
latest_data<-subset(new_title_words_tf_idf,new_title_words_tf_idf$year==2020)
titles<-sqldf("select distinct title from latest_data")
latest_5_keywords<-data.frame(title=NULL,'char'=NULL,'freq'=NULL,'year'=NULL,'tf'=NULL,'idf'=NULL,'tf_idf'=NULL)
for(i in titles$title){
  temp<-subset(latest_data,latest_data$title==i)
  temp<-temp[order(temp$tf_idf,decreasing = TRUE),]
  latest_5_keywords<-rbind(latest_5_keywords,temp[1:5,])
}
latest_new_tf_idf<-sqldf("select keywords,sum(tf_idf) as tf_idf from latest_5_keywords group by keywords")
latest_top10<-latest_new_tf_idf[order(latest_new_tf_idf$tf_idf,decreasing = T),][1:10,]
barplot(latest_top10$tf_idf,names.arg = latest_top10$keywords,ylab = "TF-IDF",main = "2020物联网热点领域",las=2)
```

最新热点分析

截至目前，2020年共有相关文献17篇，分析各领域关键词的重要程度如下：



- 就文献方面而言，目前研究热度比较高的两个热点问题为工业智慧水务问题以及泛在电力物联网问题。
- 由于2020年的数据并不完整，且受疫情影响目前文献发布数量较少，该数据仅能代表已经发布新文献的各领域。
- 结合2019年数据来看，2020年仍有很大可能研究热度较高的问题有：物联网中的算法设计、大数据、区块链、信息安全、泛在电力物联网等。

6

结语

结语

对文献的摘要进行分词操作，并将分词结果与作者关键词整理为整洁文本格式，利用R计算词项的词频、逆文档频率等参数，对文献进行层次聚类分析，深入挖掘该领域文献的组成结构。以TF-IDF数值为依据提取重要程度较高关键词并人工标注关键词所属领域，计算各领域每年文献数量的变化情况作为各领域研究热点与趋势变化的分析依据。

我国物联网发展至今，技术应用已经较为普遍且成熟。就上述研究热点的构成来看，我国物联网行业仍在朝着高精尖的核心技术不断突破。

基于IT行业的区块链、大数据等领域正飞速发展，物联网与这些领域的联系也越发紧密，这些IT技术在物联网中的应用也是当前以及未来的热点研究方向。工业制造是我国经济发展中的重要方面，利用物联网技术可以改善当前的工业生产模式，通过智能技术的引入，还能提高生产效率，增加企业的经济效益，这也将是物联网领域研究的热点问题。除此之外，但凡涉及到网络的领域也必然会重视信息安全，信息安全与隐私保护是网络技术中的重要内容。为了应对技术的革新与攻击手段的发展，在物联网领域的信息安全技术也必然是未来的热点研究问题。

Thank you!

GitHub源码链接：
https://github.com/zxynju2017/IOT_Analyse