**分类号**  **学号**

**学校代码10487 密级**

****

**硕士学位论文**

**（学术型□ 专业型☑）**

**面向社会治理的热点话题发现平台**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **学位申请人** | **：** |  | **黄滨** |  |  |
| **学科专业** | **：** |  | **计算机技术** |  |  |
| **指导教师** | **：** |  | **XXX 教授** |  |  |
| **答辩日期** | **：** |  | **XXXX年XX月XX日** |  |  |

**A Dissertation Submitted in Partial Fulfillment of the Requirements**

**for the Master Degree in Engineering/Science （工学/理学硕士）/the Professional Master Degree (专业学位)**

**英文题目，Times New Roman，小二号，  
实词的首字母大写**

**Candidate : xxx（中文习惯，姓在前且姓全部大写）**

**Major : \*\*\***

**Supervisor : Prof. xxx**

**Huazhong University of Science and Technology**

**Wuhan 430074, P. R. China**

**May, 2021**

**独创性声明**

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期： 年 月 日

**学位论文版权使用授权书**

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权华中科技大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保 密□，在 年解密后适用本授权书。

本论文属于

不保密□。

（请在以上方框内打“√”）

学位论文作者签名： 指导教师签名：

日期： 年 月 日 日期： 年 月 日

# 摘 要

一方面，随着web3.0时代的到来，各大新闻门户、微信、微博和知乎等信息发布平台取代了传统媒体。越来越多的人习惯在这些平台发布信息、了解信息。新兴社交媒体成为了政府倾听民生的重要通道，沉淀了大量的有价值的文本信息。这些信息能给政府决策提供参考，有利于提升社会治理水平。

实践发现这些文本信息很大一部分是来自社交媒体的短文本。这些短文本往往指代不明、缺失语义而且数据量庞大。而如何从海量的短文本中实时提取有价值的话题一直是话题发现任务的难点。

另一方面，传统的社会治理方式无法把握新时代社会发展的要求，也无法满足不同社会主体的多样化的需求。大数据时代要求社会治理必须能及时地参考多元信息。而传统的数据挖掘平台难以实时处理海量的文本数据，而且各部门之间数据平台独立且松散，无法有效共享信息。

因此为了解决以上问题，本文研究提出了一种增强语义的文本向量化算法和基于多层级文本聚类的热点话题发现框架，实现了从海量短文本中实时发现并追踪热点话题。最后，应用上述研究成果，设计并实现了基于社会治理的大数据舆情预警平台。本文主要研究结果：

1. 针对特征稀疏的短文本，提出了增强语义的文本向量化算法。一方面，构建社会治理类知识图谱，然后将知识图谱嵌入，给词向量引入外部的人类知识。另一方面，引入一种注意力机制，将句子的向量表示集中在关键词（实体词，共线词、主题词）之上，同时结合共现词矩阵，尽可能消除关键词的歧义。
2. 为了实时发现热点话题，研究实现了基于多层级聚类的热点话题发现框架。基于波利亚坛子概率模型识别新话题，同时通过分类器指定已有话题，实现在线文本聚类。同时引入话题衰减模型和事件更替机制，可以有效地追踪新旧事件的变化。
3. 设计并实现了基于社会治理的大数据舆情预警平台。Scrapy 爬虫获取新闻门户网站的报道。通过动态文本聚类算法，实时发现热点话题。对舆情数据进行精准的情感分析，实现舆情预警分级推送。

**关键词：** 热点话题发现；文本向量化；文本聚类；自然语言处理；语义增强

# Abstract

英文摘要字体为Times New Roman，小四，1.5倍行距。

英文摘要和关键词应与中文相对应。英语摘要用词应准确，使用本学科通用的词汇；摘要中主语（作用）常常省略，因而一般使用被动语态；应使用正确的时态，并要注意主、谓语的一致，必要的冠词不能省略。

**Key words:** Keyword1, Key word2, Keyword3

# 目 录

[摘 要 I](#_Toc57978729)

[Abstract III](#_Toc57978730)

1 绪论（黑体，居中，三号）

[1.1 研究背景与意义（黑体，四号） (1](#_Toc57978733))

[1.2 XXX国内外研究现状（请拟定具体的题目） (2](#_Toc57978734))

[1.3 存在的问题 (2](#_Toc57978735))

[1.4 本文主要内容 (2](#_Toc57978736))

2 系统与控制理论类\*

[2.1 引言（引言标题可选） (4](#_Toc57978738))

[2.2 预备知识（可选，标题可自选） (4](#_Toc57978739))

[2.3 问题的描述（请拟定具体的题目） (5](#_Toc57978740))

[2.4 控制器设计与闭环系统分析（请根据所设计的控制器特点自行拟定具体的题目） (5](#_Toc57978741))

[2.5 数值仿真（请拟定具体的题目） (6](#_Toc57978742))

[2.6 本章小结 (6](#_Toc57978743))

3 理论/算法类研究类论文

[3.1 引言（引言标题可选） (7](#_Toc57978745))

[3.2 \*\*理论/算法 (7](#_Toc57978746))

[3.3 \*\*仿真或算法实现 (7](#_Toc57978747))

[3.4 理论/算法准确性的评估 (7](#_Toc57978748))

[3.5 分析与讨论 (7](#_Toc57978749))

[3.6 本章小结 (7](#_Toc57978750))

4 学位论文写作细则

[4.1 关于图 (8](#_Toc57978752))

[4.2 关于表格 (9](#_Toc57978753))

[4.3 名词、术语 (10](#_Toc57978754))

[4.4 符号、单位的使用 (11](#_Toc57978755))

[4.5 数字的使用 (11](#_Toc57978756))

[4.6 其它应该注意的问题 (12](#_Toc57978757))

[4.7 本章小结 (12](#_Toc57978758))

[5 总结与展望](#_Toc57978759)

[5.1 本文主要内容及结论 (13](#_Toc57978760))

[5.2 本文的主要创新点 (13](#_Toc57978761))

[5.3 展望 (13](#_Toc57978762))

[致 谢 (14](#_Toc57978763))

[参考文献 (15](#_Toc57978764))

[附录1 攻读硕士学位期间取得的研究成果 (17](#_Toc57978765))

[附录2 其它附录 (18](#_Toc57978766))

# 绪论

## 研究背景与意义

一方面，随着web3.0时代的到来，新闻门户、头条新闻和微博等信息发布平台取代了传统媒体。越来越多的人习惯在这些平台发布信息、了解信息。这些新媒体成为了政府倾听民生的重要通道，沉淀了大量有价值的文本信息。这些信息能给政府决策提供参考，有利于政府提升社会治理水平。

另一方面，传统的社会治理方式无法把握新时代社会发展的要求，也无法满足不同社会主体的多样化的需求。大数据时代要求社会治理必须能及时地参考多元信息。而传统的数据挖掘平台难以实时处理海量的文本数据，而且各政府部门之间数据平台独立且松散，无法有效共享信息。因此，构建高效的大数据管理平台，自动发现和跟踪互联网中各热点事件、按统一格式分级存储海量数据，对于社会治理具有重要的意义。

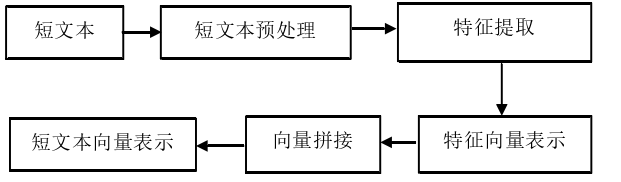
事件发现与追踪（Topic Detection and Tracking，简称 TDT）是自然语言处理技术（ Natural Language Processing, 简称NLP）的经典课题。研究人员和政府机构可以利用该技术识别并追踪新闻报道中的热点事件。

之前的话题发现算法都多认为话题是一个词语，所以通常以词为最小粒度进行词向量化，通过统计或者机器学习的方法选取出最新的热点词语。但是仅用单个词难以准确概括热点话题，而且容易产生歧义。通过引入语义增强的文本向量化算法，可以提高话题发现算法的准确度和可解释性，有利于相关部门和研究者了解舆情态势。

因此，面对社会治理对于社交媒体和新闻报道的热点话题发现与追踪的强烈需求，以及现有的以词为粒度的文本聚类算法存在上述问题。本文提出了一种正文抽取算法，从互联网中提取新闻报道等长文本和社交媒体等短文本信息。聚焦语义增强的文本向量化算法和多层在线文本聚类算法进行研究，并将成果运用于实际系统中，取得了较好的实践效果。

## 国内外研究综述

### 文本向量化



传统的文本向量化的过程， 【对比3.1 改进后的文本向量化流程】

文本向量化之前一般需要经过文本分词、取停用词等预处理的过程。经典的文本分词，包括基于词典匹配[1]、基于字标注法分词[2]等。字标注法通过上下文学习一个词出现的概率模型来实现分词。另一个重要的过程就是文本向量化。文本向量化（又称“词向量模型”、“向量空间模型”）即将文本表示成计算机可识别的实数向量，根据粒度大小不同可将文本特征表示分为字、词、句子或篇章几个层次。文本向量化的方法主要分为离散表示和分布式表示。

离散模型包括词集模型和词袋模型。词袋模型简单会导致文本向量的维度膨胀到词典的大小。（词典中有数十万个单词标记，但每个文档中只有数百个单词。）这严重加大了计算的开销，导致维度灾难。为了改进，Ramos[3]提出TF-IDF算法，采用基于降维的词典模型、提取关键的特征，将高维度的稀疏文本向量降低维度后进行处理。在利用TF-IDF进行特征提取时，若词α在某篇文档中出现频率较高且在其他文档中出现频率较低时，则认为α可以代表该文档的特征，具有较好的分类能力，那么α作为特征被提取出来。此外常用的降维方法还有奇异值分解（SVD）[4]，非负矩阵分解（NMF）[5],主成分分析（PCA）[6]和随机近邻嵌入（SNE）[7]等。

但是词袋模型失去了单词的顺序，也忽略了单词的语义。例如，“系统”和“算法”可能很遥远，但是“系统”和“存储”也一样遥远。引入近义词库做语义替换可以一定程度上缓解这个问题。但是无法解决词袋模型忽略单词顺序的弊端。

还有一些学者采用基于深度学习的词嵌入表示的向量化方法，例如word2vec模型。word2vec模型是一个双层的神经网络，利用语义信息用来构建词的向量。word2vec初始化时输入的是词袋模型one-hot编码，通过猜测相邻位置的输入词来加强文本向量之间在上下文中的联系。2013年Mikolov[8]提出了word2vec中两个重要的模型：CBOW模型和Skip-gram模型。CBOW 模型根据上下文预测当前词的概率。相比之下，Skip-gram 模型根据当前词预测周围词的概率。Word2vec 利用这两种模型将单词表达为词向量，可用于对词向量进行聚类、分类和计算相似度。一般通过word2vec获得词向量之后，对所有的词向量进行加权平均可以得到一个文本的向量表示。但是这样丢失了较多的语义信息。2014年Quocle和Tomas[9]又在word2vec方法的基础上，又给出了doc2vec的训练方法。本质上和word2vec一致，都是训练一个神经网络，用一个密集的向量来表示每个文档，该向量被训练来预测文档中的单词。

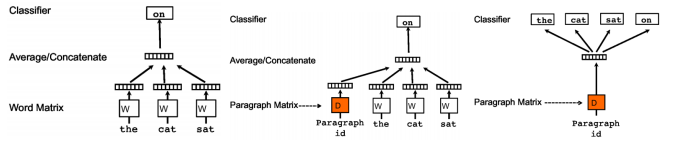


图 1.1 Word2vec和Doc2vec原理

### 知识图谱嵌入算法

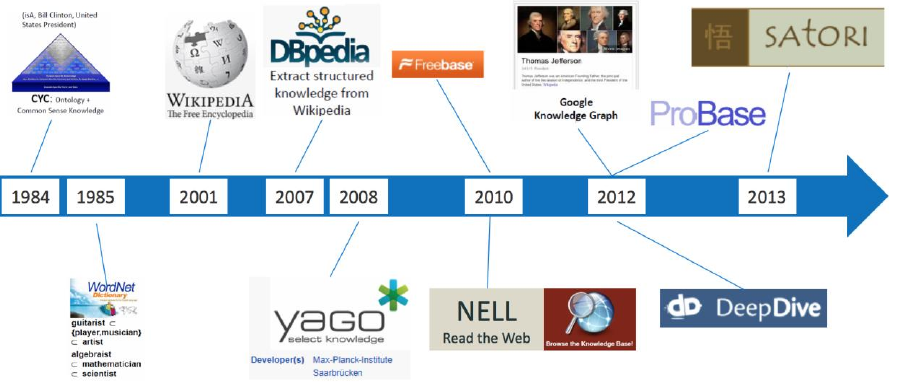
2012年Google[10]率先提出了知识图谱Knowledge Graph（KG）的概念。知识图谱使用（实体-关系-实体）三元组的概念描述物理世界的概念，形成了结构化的网状结构的知识库。知识图谱的实体抽取技术和知识表示都是这一领域的主要研究方向。知识图谱以知识驱动算法，被广泛应用于智能搜索[11,12,13]、智能问答[14,15,16]、个性化推荐[17,18]等领域。从 1984 年开始出现了类似知识库的概念。2012 年 5 月份，Google 花重金收购 Metaweb 公司，并向外界正式发布其知识图谱(knowledge graph)。自此，知识图谱正式走入公众视野。从最初的靠人力搭建的小规模的词典 WordNet[19]，到目前大规模的 YAGO[20]、 Freebase[21]和ProBase[22]等等。知识库的种类也是越来越多，例如，树形结构的 WordNet、有向图结构的 FreeBase 以及二部关系图的 ProBase。2019年Github开源了目前最大的中文知识图谱知识库OwnThink。

图 1.2知识图谱库发展

### 热点事件发现算法

获得向量化的文本之后就是选用主题模型（TopicModel）抽取其主题。有很多学者使用监督方法来处理特定领域的TDT任务。Sakaki[32]等人提出来二分类支持向量机 (SVM)从推特数据集实时检测地震和预测地震中心。陈燕、Hadi[33]也使用使用二分类支持向量机 (SVM)发现新事件。除了经典的SVM，神经网络也在这一领域被广泛运用。Nguyen[34]等人引入递归神经网络RNN来检测事件模型。Wang[35]等人提出了一种基于LSTM（递归神经网络RNN的一个变种）的事件发现和总结的联合模型。Maksim、Hady[36]等提出了引入了半监督的CompareLDA算法用来鉴别指定主题下的文档。目的是研究CompareLDA是否可以从同一语料库中获得不同的主题。这种方法确实提高了聚类的效果，但有主题之间必须是对立的，而且同样需要预先标记样本训练模型。但是监督方法需要给指定特征、标记样本。而为此需要给定一个稳定的静态环境，事件是给定的而且不能变动。这在处理大规模数据集会很遇到很多困难。

还有一些研究者专注于研究无监督的事件发现方法，因为社交媒体数据不能包含所有预先定义的事件，无监督算法也不需要事先指定标签。其中聚类算法应用最为广泛。常用的聚类算法有基于相似度的聚类，例如single-pass、层次聚类（HCA）、DBSCAN和K-Means等。还有基于概率模型的聚类算法例如潜在的狄利克雷分布（LDA）[37]模型和狄利克雷多项混合（DMM）模型。Hansi[38]结合了基于预测的词嵌入和层次聚类（HCA）的特征，在社交媒体中进行事件发现。邱兴发、邹桥沙[39]使用BERT对语料集进行预训练，获得训练过的编码器，将文档和事件映射到同样的低维语义空间，然后计算他们之间的相似度。最后提出了能动态更替关键词的single-pass算法在公共的Twitter数据集上取得优越的聚类效果。

有一些学者使用混合方法结合了监督学习和无监督学习的优点进行事件发现。Teitler [40]训练一个分类器来区分与事件相关的推文和与事件无关的推文。它可以帮助减少为聚类提供的数据量，还使聚类算法能够并行计算，从而提高系统的效率。而Becker等人[41]首先进行聚类，然后使用SVM对聚类结果中是否包含关于相关信息进行分类。

## 本文主要研究内容和方向

综合当前国内外研究现状，总结现有研究中存在的问题，本文的研究和工作内容如下：

（1）针对单个词语容易产生歧义的问题，提出了语义增强的文本向量化算法。由于以word2vec为代表的空间向量模型以词为最小粒度，将词表示为一个低维向量。但是它把有多个含义的单词能映射为一个向量，对短文本聚类会造成较大的误差。而且大部分的句子的语义实际上只集中在一两个词上，其他词起到的作用很小。

所有，该算法引入一种注意力机制，将句子的向量表示集中在关键词（实体词，共线词、主题词）之上，同时通过词共现矩阵尽可能消除关键词的歧义。

（2）针对快速更新的短文本信息，本文研究提出了多层在线文本聚类框架MOTC（Multilayer online text clustering）。该模型以先分类后聚类的方式，利用分类器指定话题，实现在线聚类。基于概率模型的进行新话题识别，在每个文档都有一定的概率会单独成为一个集群，聚类时不需要手动定义已有集群和新文档之间的差异阈值。对已有的话题再次进行聚类，发现其中的子话题和新话题，并设计热度衰减的话题模型形成树状模型和事件更替机制。

## 本文组织结构

本文共分为6章内容，主要内容如下：

第一章，绪论：简要介绍论文的研究背景和意义、国内外研究现状，给出全文的主要研究内容和贡献点，最后给出了全文的组织结构。

第二章，相关背景知识。首先介绍了文本向量化算法的相关研究，包括word2Vec和Doc2Vec等，以及知识图谱嵌入算法及其原理。然后介绍了常见的基于文本聚类的热点话题发现算法。

第三章，语义增强的文本向量化算法。本章提出了一种带有注意力机制的文本向量化算法。通过多种方式对已有的词向量进行加权，并调整权重达到消除语义歧义的效果。结合知识图谱嵌入，给向量化的文本引入外部知识。并通过实验验证这种文本向量化算法对比word2Vec对于下游文本聚类或者文本分类效果的优化。

第四章，多层在线文本聚类框架MOTC（Multilayer online text clustering）。首先介绍这种框架的设计如何实现在线聚类和话题热度衰减。然后通过实验进行验证对比传统文本聚类算法的效果和性能。

第五章，基于正文抽取算法和在线热点话题发现算法的文本大数据分析系统设计。从实践中社会治理智能化的需求出发，设计并实现了基于分布式爬虫和正文抽取算法的文本大数据收集系统，并应用本文提出的文本向量化和多层聚类算法设计了热点事件发现系统。首先介绍了项目的背景和系统的整体架构，以及核心功能的设计。最后对功能进行测试和展示。

第六章，总结与展望。总结了全文的主要内容，提出了对未来工作的设想。如何减少模型的复杂度，以及如何提高训练效率。

# 相关背景知识

由于本章介绍了自然语言处理中的基本理论知识，通过描述经典的文本向量化和知识图谱表示模型的发展过程以及优缺点，具体说明了进行词嵌入的一般思路，为话题发现算法打下基础。随后引入常见的基于文本聚类的话题发现算法。

## 预备知识

### 文本向量化

文本向量化（又称“词向量模型”、“向量空间模型”）即将文本表示成计算机可识别的实数向量，根据粒度大小不同可将文本特征表示分为字、词、句子或篇章几个层次。文本向量化的方法主要分为离散表示和分布式表示。

离散模型包括词集模型和词袋模型。这两种方法都是基于词之间保持独立性、没有关联为前提，将所有文本中单词按字典排序形成一个单词集合。词集模型的单词向量化采用简单的One-Hot编码，标识当前文本是否包含字典中的词语。词集模式中的文本向量就是文本中所有的单词向量的叠加，最后得到的向量每个分量代表该位置对应单词在文本中的出现次数。词袋模型简单会导致文本向量的维度膨胀到词典的大小。（词典中有数十万个单词标记，但每个文档中只有数百个单词。）这严重加大了计算的开销，导致维度灾难。

TF-IDF算法，采用基于降维的词典模型、提取关键的特征，将高维度的稀疏文本向量降低维度后进行处理。在利用TF-IDF进行特征提取时，若词A在某篇文档中出现频率较高且在其他文档中出现频率较低时，则认为A可以代表该文档的特征，具有较好的分类能力，那么α作为特征被提取出来。此外常用的降维方法还有奇异值分解（SVD），非负矩阵分解（NMF）,主成分分析（PCA）和随机近邻嵌入（SNE）等。

但是词袋模型失去了单词的顺序，也忽略了单词的语义。例如，“系统”和“算法”可能很遥远，但是“系统”和“存储”也一样遥远。引入近义词库做语义替换可以一定程度上缓解这个问题。但是无法解决词袋模型忽略单词顺序的弊端。

还有采用基于深度学习的词嵌入表示的向量化方法，例如word2vec模型。word2vec模型是一个双层的神经网络，利用语义信息用来构建词的向量。word2vec初始化时输入的是词袋模型one-hot编码，通过猜测相邻位置的输入词来加强文本向量之间在上下文中的联系。word2vec中有两个重要的模型：CBOW模型和Skip-gram模型。CBOW 模型根据上下文预测当前词的概率。相比之下，Skip-gram 模型根据当前词预测周围词的概率。

Word2vec 利用这两种模型将单词表达为词向量，可用于对词向量进行聚类、分类和计算相似度。一般通过word2vec获得词向量之后，对所有的词向量进行加权平均可以得到一个文本的向量表示。但是这样丢失了较多的语义信息。

### 知识图谱表示模型

知识图谱表示模型，其中最广泛使用的是平移距离模型。它采用基于距离的评分函数，典型代表有TRANSE及其拓展模型GTRANSE模型，高斯嵌入模型KG2E和TransG等。其中TRANSE模型能把知识图谱中的实体和关系嵌入到低维的文本空间中变成低维向量，同时保留原有的知识信息。该模型认为每个关系都是主体向量到客体向量的一个平移映射。要求知识三元组（h，r，t）在低维空间的向量表示满足h + r ≈ t。通过机器学习的方法学习对象的语义信息，将语义信息转化为稠密的实值向量，同时保留语义的完整信息。 通过删除正例三元组(h, r, t)中的h(或t)，然后从实体集中随机选择一个实体对删除h(或t)不完整的三元组进行填充来生成负例三元组。构建浅层的神经网络，将正采样和负采样结合作为损失函数，最后通过训练的神经网络模型可以把实体和关系映射成为低维向量。

### 聚类算法

K-Means 的数学原理为给定样本集%FontSize=12
%TeXFontSize=12
\documentclass{article}
\pagestyle{empty}
\begin{document}
\[
 D = \{X_1,X_2,\cdots,X_P\}
\]
\end{document}。K-means 对聚类所得簇的划分为%FontSize=12
%TeXFontSize=12
\documentclass{article}
\pagestyle{empty}
\begin{document}
\[
 \{C_1,C_2,\cdots,C_K\}
\]
\end{document}，每个 Ci 是观测的指标的集合，称之为类。将聚类要求用数学语言描述出来，不难发现这些类应该满足下列两个性质 ：即每个观测必然属于这K 个类中其中一个类，且任意两个类与类之间没有重复的样本数据。用数学语言描述如下

(1) %FontSize=12
%TeXFontSize=12
\documentclass{article}
\pagestyle{empty}
\begin{document}
\[
C_1\cup C_2\cup\cdots\cup C_k=\{1,2,\cdots,n\}
\]
\end{document};

(2)%FontSize=12
%TeXFontSize=12
\documentclass{article}
\usepackage{amssymb, amsthm, amsmath}
\pagestyle{empty}
\begin{document}
\[
 C_i\cap C_j = \emptyset
\]
\end{document} 对每个 (0 < i < j < K +1) 都成立。

设第 *k* 类的类内差异是对第 *k* 类中观测互不相同程度的度量%FontSize=12
%TeXFontSize=12
\documentclass{article}
\pagestyle{empty}
\begin{document}
\[
W(C_k)
\]
\end{document}。因此简要解决如下最小化问题:

%FontSize=12
%TeXFontSize=12
\documentclass{article}
\pagestyle{empty}
\begin{document}
\[
\min_{c_1,\cdots,c_k}\left\lbrace\sum\limits_{t=1}^K W(C_t)\right\rbrace
\]
\end{document}

为了解决这个 NP 难问题，K-Means 算法采用了贪心策略：确保每一步都比上一次“更好”的 K-Means 算法，通过迭代的算法来逼近真实解。K-Means 采用的这样的启发式方式[30]就很简单，大大减少求解这个问题的难度。

K-Means 算法描述

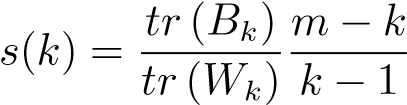
1. 将每一个样本随机分配到1到k的某个类中。

2. 重复下列操作，直到类中心不再更新或者运行轮数达到阈值为止

(a) 分别计算每一个类的类中心，一共要计算k次，中心的定义一般默认为该类中全部向量的均值；

(b)计算每个样本到这k个类中心的距离。选取距离它最近的那个类，并将该样本归于该类 (可用欧式距离定义 “最近”)。

用于评价 K-means 算法的常见指标有轮廓系数 Silhouette Coefficient [31]，Calinski-Harabasz Index[32]以及Davies-Bouldin。其中，Calinski-Harabasz Index 计算简单直接，得到的指标数值越大通常代表着聚类效果越好。这一指标数学计算公式是：



其中 m 为样本数，k 为类别数。Bk 为类别之间的协方差矩阵，Wk 为类别内部数据的协方差矩阵。tr 为矩阵的迹。

层次聚类 (hierarchical clustering)，也叫系统聚类。本文使用的是最广泛的一种实现方法：“自底向上”的“凝聚法”。“凝聚法”的根本目的是自下而上慢慢归类，画出一个系谱图。首先，要把每一个样本作为初始聚类，然后任意选取两个样本，挑选其中最接近的进行合并操作。如此循环，直到结束。算法思想与哈夫曼树算法有异曲同工之妙。

描述两类集合的距离的方法共五种, 比较常用的是前四种：

Ward距离法：计算A 类和 B 类合并后的类的方差作为距离，这是Scikit-learn（0.23.1）层次聚类的默认方法；

最短 (single) 距离法：计算两个集合中所有观测的距离，记录其最小值；

最长 (complete) 距离法：计算两个集合中所有观测的距离，记录其最大值；

平均 (average) 距离法：计算两个集合中所有观测的距离，记录其平均值；

重心 (centroid) 距离法，计算 A 类重心 (平均向量) 和 B 类重心的距离，较少使用。

除此以外，还可以选择多种“距离”，比如欧氏距离，曼哈顿距离等。

系统聚类算法描述

1. 先把每个观测看成一个类，然后任意两类为一对，共%FontSize=12
%TeXFontSize=12
\documentclass{article}
\pagestyle{empty}
\begin{document}
\[
n\choose 2
\]
\end{document} 对，计算其距离。

2. %FontSize=12
%TeXFontSize=12
\documentclass{article}
\pagestyle{empty}
\begin{document}
\[
i=n,n-1,\cdots,2
\]
\end{document}

(a) 在这 i 类中，比较任意两对的距离，找到最小的那一对，将他们结合起来。用这两个类的距离作为这两类在谱系图中交汇的高度。

(b) 计算剩余所以类中任选两类的距离。

## 问题描述

## 本章小结

Todo。

# 语义增强的文本向量化算法

传统的文本向量化算法，例如Doc2Vec在处理新闻报道类长文本时往往有比较好的效果。因为长文本的词语量较短文本更充足，加权后的文本往往包含了更多的信息。但是短文本来自于社交媒体的博客、新闻的评论区等，通常篇幅更短，指代不明。如果直接采用传统向量化方法势必会减少文本对应的向量的语义信息，因此本章提出了基于语义增强的文本向量化算法，解决短文本特征稀疏和一词多义的问题。为了验证算法的效果，在两个经典的中英文本数据集中应用了该词嵌入算法。最后，对比该算法和传统的算法Doc2Vec对下游的文本分类、文本聚类任务的影响，进一步验证了该算法的有效性。

## 问题的描述和定义

文本向量化是指将一条长（短）文本信息映射为一个定长的数值向量。在处理NLP任务时，不管是早期的机器学习方法，还是现在流行的深度神经网络，首先都需要从文本中提取语义、语法和上下文特征。理论上来讲，文本特征的提取效果往往决定了处理文本的各类算法的上限。如果文本特征提取时语义缺失，那么必然会将误差传递到以文本为单位的下游NLP任务中去。因此，为了更好地改善文本向量化的效果，基于传统的文本向量化方法，本文提出了引入外部知识图谱知识和自适应关键词权重的方法。

为了提升效果，文本向量化算法往往会通过引入外部知识来拓展文本的语义，尤其是信息较少的短文本。引入外部知识的方法大致有三种：通过百科网站的文本增加训练集；利用知识图谱的知识来丰富文本的语义；还有就是微调已经训练好的深度预训练网络例如Bert，复用其参数。前两种从数据集的角度对文本数据进行了增强，有着更广泛的适用性。

另一方面，本章为了解决较短文本的特征稀疏和一词多义的问题，提出了基于语义增强的文本向量化算法。分别从引入外部知识、利用上下文共现词的角度增强词向量的语义。最后，通过自适应调整关键词的权重的方式，对文本中的关键词进行加权平均获取到最后的文本向量，使得文本能包含更多关键信息。

## 自适应关键词权重的文本向量化

### 关键词选取算法

【为什么要选取关键词】长文本中包含极多的词语和语义，换而言之，包含了极大的噪声。对于长文本中的全量词语，不管是拼接还是加权平均其词向量，都会导致维度灾难或者语义缺失。在深度学习领域广泛使用的注意力机制已经证明通过调整特征的权重可以大幅提升表示学习的效果。本文认为如果能选取出关键词，然后由关键词拼接而成的向量会更具有代表性，也可以在保持维度较少的情况下，尽可能多地获得语义信息。传统的关键词选取的流程如下图所示：

### 多维度的词向量融合

在进行文本分类、文本相似度检测等以文本为粒度的NLP任务时，传统的语言模型通过机器学习或者深度学习方法获得词的向量表示之后往往通过加权平均或者拼接的方式获得文本的向量表示。文本中全部词的向量拼接对于长文本来说会产生“维度爆炸”和维度不均衡的问题。尽管可以通过PCA、TSNE等降维算法调整维度，但是仍然不可避免地带来精度的损失。实践过程中，往往会采用加权平均的方式。加权平均包括有直接相加、相加取平均值、与TF-IDF结合将词频作为权重进行加权平均等方式。

从实践上看，TF-IDF结合Word2Vec的文本向量化算法已经能取得较好的效果，但是仅仅通过词频来调整权重是不完善。特别是当关键性人名容易被代词所替代之后，尽管名词的词频有所下降但实际上的重要性依旧很高。直观来看，一段文字中的主要语义往往集中在了人名、地名和动词上，助词和副词往往分担了较少语义，显然不同词性的词不应该有相同语义。为了解决这一问题，本文计划结合词频、词性因素调整权重达到最佳的文本向量。

### 基于知识图谱嵌入的知识拓展

## 实验设计及结果分析

## 本章小结

本章简要给出理论/算法类研究论文的基本框架。在每章的最后，都需要对该章的内容进行小结，不宜太长，建议1/2-2/3页版面。主要小结一下本章用什么理论或方法、做了什么事、得到的重要结果或结论。

# 基于多层聚类的话题识别模型

学位论文很多的错误源于凌乱的格式。

为规范学位论文写作，本章结合工科学位论文的特点，参照学术出版规范，就图、表制作，名词、术语、单位、符号、数量等的使用规范化进行了说明。

## 关于图

参照CY/T 170-2019《学术出版规范 插图》标准执行。

论文中的图一般居中放置，大小要合适，宽度不得超过文字边缘，图中文字一般不大于五号；图像分辨率应尽量不低于300 dpi；若为自画图，尽量采用中文标识，图中文字经缩放后，字号不得大于五号；图注也只需加注中文，宋体，五号。若引用文献中的图，应在图注最后标注引用，图注后无需加标点符号。文中所有的图都应予编号，图序号按“章”进行编号，如图4-1所示：

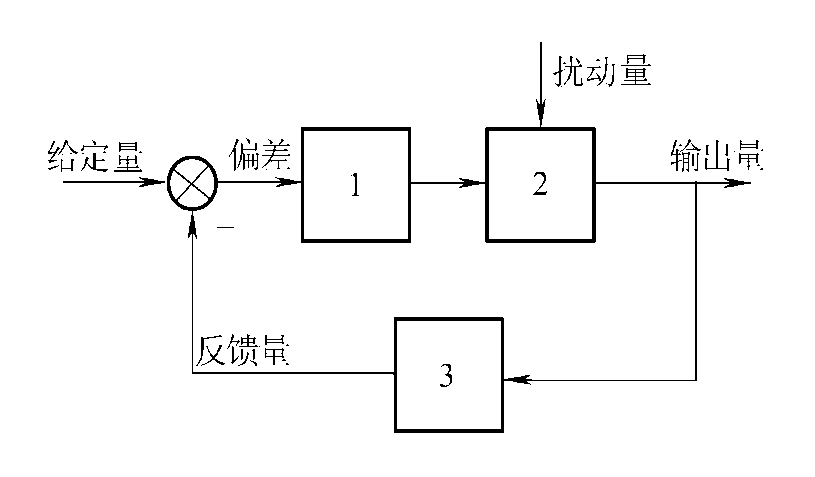


图4-1 闭环系统示意图[1]

制作图时请注意:

（1）内容与形式应力求统一。风格、体例应一致。线条图应清晰，线型选用、线条粗细应规范，色调准确，图形布局合理、大小适当。连续色调图应清晰，层次感和饱和度适当。

（2）每幅都得有图题。图题应准确、简明地阐释该图内容。

（3）图应与正文内容相关。应选择能有效传达关键信息的图，应具有自明性、简明性、科学性和艺术性。

（4）结构示意图、原理示意图和流程图的设计制作应符合现行的国家标准或行业标准。

（5）坐标曲线图的坐标轴、标值线的画法应规范，标目、标值、坐标原点应标注完整、规范、统一。

（6）图中涉及标志用图形符号、设备用图形符号和技术文件用图形符号应符合现行的国家标准。图中的术语、数值、符号等应与正文其它图中的表述一致。

（7）图应尽量与相应的文字靠近，根据排版可放在文字的下方或上方。各章节不能因为图的原因出现大段留白的现象。

（8）图中应有对于需特别说明问题部分的标识说明，如图像中需给出比例尺，在彩图中给出Color bar等。说明字体与比例尺的字体应至少比正文小一号。

## 关于表格

学位论文中的表格要使用三线格，居中放置。表格题目应居中放置于表格顶端。仅需提供中文表头，表头及表中文字应小于正文字号，一般为宋体，五号；若其中含英文或数字时，字体为Times New Roman，五号。表格题目与表格尽量不要分页。实在需要分页时，请在下一页续表。若该表完全源于文献，应在表格题目最后标注引用。标头最后不加标点。如表4-1所示：

表4-1 表格示例[2]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Aaa\* | Bbb | Ccc | Ddd | Eee[3] |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

\*注：存在注释时，请采用宋体、5号

一般情况下，表应放在文字的下面，有时依据排版的需要，也可以放在文字上方。但无论如何，图应尽量与文字靠近。

（1）表格序号按“章”进行编号，如第3章第1个表为“表3-1”，以此类推。

（2）表格内容与正文配合应相得益彰，内容适合用表表达。

（3）表格应具有自明性和简明性，栏目设置应科学。

（4）表格中的数据应具有完整性和准确性。

（5）表格中连续数的分组应科学，不得重叠和遗漏。

（6）表格中的数值修约和极原数值的书写应符合GB/T 8170的规定。

（7）表格中的量和单位的名称、符号及书写应符合国家标准GB3100～3102—93量和单位的系列标准和相关行业标准。

（8）表格中数字形式的使用应符合GB/T 15835规定。

（9）表格中的科学技术名词应符合CVT 119的规定。

（10）表格中的术语、数值、符号等应与正文以及同一文本中其他表格中的表述一致。

（11）全书的表格的表号、表题，表头、表身、表注的格式应统一。

## 名词、术语

学位论文存在大量的专业名词与专业术语，针对同一内容的名字、术语有多种表达方式时，原则上以关系最为密切的学科为准，并尽量符合最新的国家或行业规程、规范。若属全国科学技术名词审定委员会最新公布的各类学科的“名词”，则须严格执行，且在全文中统一。

外国专有名称在释文中首次出现时，应附原文和简称。例如：“美国垦务局（United States Bureau of Reclamati，USBR）”、“美国大坝委员会（United States Committee On Large Dams，USCOLD）”。在同一条目中再次出现时可采用中文或英文缩写。

## 符号、单位的使用

标点符号的使用，应符合国家标准GB／T 15834—2011《标点符号用法》。

（1）论文中的计量单位应采用法定计量单位（简称法定单位），应符合国家标准GB 3100～3102—93量和单位的系列标准和相关行业标准。

（2）表格及插图中，使用单位符号，不使用单位名称和单位中文符号。叙述性文字中，优先使用单位符号。必要时，可使用单位名称，但不可使用单位中文符号。如：流量为11400 m3/s，不可写作流量11400米３／秒。

（3）两个物理量（量值加单位）在表示范围时，两个量值用波浪线“～”连接后使用一个计量单位，如：应写作800～1500 m3/s，而不写作800m3/s～1500 m3/s。

## 数字的使用

数字的使用应符合国家标准GB/T15838—2011《出版物上数字用法》，同时应符合相关行业标准。

以下情况应使用阿拉伯数字：

（1）统计表中的数值，如：正负整数、小数、百分数、分数、比例。

（2）物理量量值，如：150 m3/s，200 kg，注意数量与单位之间应有空格。

（3）非物理量量值。如：21.35元，480人。

（4）当阿拉伯数字与汉字数字混用时，要顾及上下行文的协调一致。

（5）两个百分数表示范围时，要使用两个百分号，如15%～20%，不可写成15～20%。

（6）专业性科技出版物上的多位数字，应从小数点算起，每三位留空半个数码位置，不采用传统的以千位撇“，”分节的办法，如3800000应写成3 800 000，或写成380万，而不要写成3，800，000。

以下情况应使用汉字数字：

（1）定型的词、词组、成语、惯用语或具有修辞色彩的词语中作为词素的数字。如：星期六、四氧化三铁、五省一市、“八五”计划、第三季度等。

（2）相邻两个数字并列连用表示概数，连用的两个数字间不得用顿号“、”隔开。如：二三米、十三四吨、一千七八百元。

（3）不是出现在具有统计意义的一组数字中的整数一至十。如：一个人、三本书、五个百分点等。

（4）带有“几”的数字，表示概数。如：十几天、几千年等。

## 其它应该注意的问题

### 关于文献引用

在学位论文撰写中，凡是文字、图、表来自于参考文献的，必须要加注文献信息。国际惯例：与他人文章有20个字连续雷同的就是抄袭。抄袭是必须严禁的行为，否则害人害己。建议大家最好用自己的理解来撰写相关内容，特别是第一章绪论中更应注意。

在引用参考文献时，不要将文献标识直接写在某一小节的标题上，这将被视为整节内容都是引用；如有某几句话完全来自文献的，则在这部分内容的最后一句的结束加文献标注。

在课题组内部，经常出现多位同学做相近的课题，或课题的关联性较大。每个学生介绍研究背景或实验室材料与方法时，请尽可能用自己理解后的语句去撰写，以避免雷同，千万不要图省事将前一届学生的论文大段地照搬过来。有些图若是能自己画的，最好自己重新画一遍，以示区分。

### 公式

公式中主要字母的字号与正文一样（即小四号），尽量避免图片转贴过来的公式。公式编号根据章节按顺序进行编排，如第2章第3个公式，标注为2-3，将公式编号以右对齐方式排列，但注意公式是以居中的格式排列的（类似教材中的通用格式）。

## 本章小结

本章主要介绍学位论文写作的规范化要求，包括图、表制作及其与正文的对应关系；专业名词、术语、单位、符号及数字的使用等。

# 总结与展望

## 本文主要内容及结论

对全文进行全面地总结，并根据各章节归纳出若干有机联系的论点。按正文的内容分段描述，包括本研究“做了什么（提出\*\*新理论/算法、设计或研发\*\*工艺/仪器）、获取什么结果、得出什么结论”。

请特别注意，全文总结与摘要及各章的小节要有所区分，不能简单的拷贝。这里的重点是结论，结论应该准确、完整、明确、精练。

## 本文的主要创新点

通常情况下，学位论文的创新点应放在最后一章。

创新点要凝炼，表述要清晰明了，如提出了什么创新的思路，主要特点是什么，相比现有理论或技术的提高是什么、或者有什么新的发现，是否具有重要的科学意义或应用前景。既不能过于简单，也不要太细。

硕士学位论文创新点不宜太多，一般为2个左右即可，要注意归纳创新点，千万不要以为越多越好。论文的创新不以创新点的多少来评定的，而以其创新性的价值来评定。几章的工作合在一起凝炼成一个创新点也不是不可以的。

## 展望

对本研究成果的意义、推广应用的现实性或可能性加以论述。同时，描述本文研究中尚存在的不足，或因时间尚未完成但又必须继续的工作，对进一步的工作进行展望。

# 致 谢

对在课题研究及论文写作过程中给予指导和帮助的导师、校内外专家、实验技术人员、同学等表示感谢。

在致谢时建议具体，不同的人如何助力完成你的论文，都需要特别注明。如导师、其他老师或实验技术人员、以及同学对你论文的贡献是不一样的，有指引课题方向、修改论文，也有具体教会实验操作，也有协助你做了哪方向的实验，或者给你精神安慰、陪你度过紧张的研究生生涯。

越具体越能表达你真实的感受，否则就是毫无意义的套话。

# 参考文献

硕士学位申请人的文献阅读量一般不少于40篇，其中外文文献一般不少于1/3；近五年的论文一般不少于1/3；绪论部分应对所读文献加以分析和综合。

参考文献标注法，在正文中引用文献内容处注明参考文献编号。参考文献目录按正文中引用先后顺序排列，重复引用的文献，按第一次出现的顺序编号。

可列于参考文献表的文献类型包括图书、期刊、会议论文集、专利和学位论文等。

当论文中有些术语、公式、背景或数据来源需要解释或说明，以及援引他人的原话、数据等资料而必须指明资料来源时，可用脚注。脚注要按顺序编号。脚注按每一页单独编号。脚注的标识可以用数字 1，2 等，也可以用符号①，②等。脚注的资料来源表示方法同参考文献。

**参考文献格式**

中文书刊：作者按中文写法，姓在前、名在后；英文书刊：作者按英文习惯写法，如在前、姓在后，名用首字母缩写、姓用全称。一般6人以内须列出全部作者，6人以上写6人再加“等”（英文加“et al”））。每个参考文献的最后不加标点符号。

（1）图书：最多列出6个作者，作者与作者之间用逗号分隔. 书名. 版本（第×版）. 译者. 出版地: 出版者, 出版年. 起页-止页（可选）

（2）期刊：最多列出6个作者，作者与作者之间用逗号分隔. 文章名. 期刊名（全称）. 年号, 卷号（期号）: 起页-止页或论文编号

（3）会议论文集：最多列出6个作者，作者之间用逗号分隔. 文章名. 见（英文用“in”）：会议名称（或论文集）. 会议城市, 国家, 会议时间, 出版者, 出版年: 起页-止页

（4）专利：专利申请者. 专利题名. 专利国别, 专利文献种类, 专利号, 出版年

（5）学位论文：作者. 题名：[博士（或硕士）学位论文]. 保存地点: 保存单位（如华中科技大学, 年份）

**参考文献（举例）**

[1] 闫明礼, 张东刚. CFG桩复合地基技术及工程实践（第二版）. 北京: 中国水利水电出版社, 2006

[2] M. Chalfie, S. R. Kain. Green fluorescent protein: properties, applications, and protocols. Hoboken, New Jersey: Wiley-interscience, 1998

[3] 詹向红, 李德新. 中医药防治阿尔茨海默病实验研究述要. 中华中医药学刊, 2004, 22(11): 2094-2096

[4] E. S. Lein, M. J. Hawrylycz, N. Ao, M. Ayres, A. Bensinger, A. Bernard, et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature, 2007, 445(7124): 168-176

[5] M. L. Bouxsein, S. K. Boyd, B. A. Christiansen, R. E. Guldberg, K. J. Jepsen, R. Müller. Guidelines for assessment of bone microstructure in rodents using micro–computed tomography. Journal of Bone and Mineral Research, 2010, 25(7): 1468-1486

[6] Y. Shunsuke, A. Masahide, K. Masayuki, M. Yoshizawa. Performance evaluation of phase-only correlation functions from the viewpoint of correlation Filters, in: 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, HI, USA, 12-15 Nov. 2018, Proceedings of the IEEE, 2019: 1361-1364

[7] T. Yao, J. Wan, P. Huang, X. He, F. Wu, C. Xie. Building efficient key-value stores via a lightweight compaction tree. ACM Transactions on Storage, 2017, 13(4): 1-28

[8] 刘加林. 多功能一次性压舌板: 中国, ZL92214985. 2 [P]. 1993

[9] 李清泉. 基于混合数据结构的三维GIS数据模型与空间分析研究[博士学位论文]. 武汉: 武汉测绘科技大学, 1998

**注释体例的基本内容、结构与位置**

（1）基本内容与结构

“注释体例”含“资料性注释”和“内容性注释”两方面，合一编排。

（2）位置

正文内需注释之处依次排注号，释文于当页下部逐条依次编排。可在正文和页下注之间划一道分隔线，或通过不同的字体将二者区分开来。

（3）排版

【字体】中文：小五，宋体；英文：times new roman 9号字体；

【行距】单倍行距；

【段落】顶格写，无首行缩进，也无左缩进；

【序号】用①这种格式，序号后空一个字符，每页重新编序；

【页码】中文：第х-х页，如 第16-17页。英文：pp.х~х，如pp.5~8, 单页用pх，如p19.

【标点符号】中文使用中文状态下标点符号，英文使用英文状态下标点符号，切忌混用。

# 附录1 攻读硕士学位期间取得的研究成果

**发表与接收论文**

[1] 参照参考文献列出学术论文相关信息（含期刊、会议、或参编书稿），但无论有多少个作者，都必须列出全部作者名；若为英文论文，则名在前、姓在后，姓名均为全称；在本人的名字加粗，以示区别（若为第一作者，则需在最后特别注明署名华中科技大学是否为第一单位）

[2] 若已发表，按参考文献给出页码；若只是online,给出链接；若接受或修改或投稿或拟投，也必须分别注明

[3] 一般情况，一作或重要的论文放在前面

**专 利**

[1] 全部作者的姓名全称，本人的名字加粗. 专利题名. 专利国别，专利文献种类，专利号或申请号

**标 准**

[1] 全部作者的姓名全称，本人的名字加粗. 标准题名. 哪种层次的标准，发表年

**科技奖励**

[1] 全部作者的姓名全称，本人的名字加粗. 题目. 国家级/省部级科技类奖，获奖年

[2] 全部作者的姓名全称，本人的名字加粗. 题目. 国际/国内竞赛类奖，获奖年

# 附录2 其它附录

可包括详细的公式推导、实验数据、计算程序、援引他人的原始资料、数据及其设备条件等。

**参考文献**

[1] 郭越. 融合情感的微博热点话题发现与演化分析\_郭越[D]. 大连理工大学, 2021.