# 短文本分析综述

## 摘要

短文本能应用于多个领域。本文分别介绍了短文本的概念以及特点，介绍了文本分类过程。短文本分析的重点是文本分类。本文介绍了一些传统的以及部分最新的短文本分析方法。最后本文以一些案例介绍了短文本分析的应用。

本文的小组分工沿照课堂演讲分工。

刘向昆：引言、短文本概念及特点、短文本分类模型、分类过程

古仁华：文本分类算法、短文本研究现状

刘思杰：短文本分析应用、聚类算法、短文本主题建模方法

## 引言

随着近几年移动互联网的快速发展，互联网中即时消息、微博、短信等短文本信息 的快速增长，人们迫切需要对这些信息进行分类处理，以便准确、快速地定位自己所需的信息，并且从中挖掘出更多有用的信息。目前已有的针对长文本的分类技术取得了丰硕的成果，但是由于短文本自身的特点（如稀疏性、实时性、不规范性等），传统的基于长文本的分类方法对于短文本并不能取得令人满意的效果。

各种短文本素材中，Twitter，微博等社交媒体以其巨大的用户量，内容实时性强，与社会热点内容紧密相关等原因成为短文本分析领域最受欢迎的素材。其中微博是近年兴起的一种信息交流媒体，相比于传 统社会媒体，其发展态势相当之强劲，已逐渐成为人们 生活中不可缺少的一部分。微博2018年第二季度财报发布的数据显示，微博月活跃用户数4.31亿，年度净增用户约7000万，同比增长19%，季度净增用户2000万，环比增长5%，其中，93%为移动端用户；日活跃用户数1.90亿，年度净增用户约3100万，同比增长15%，季度净增用户约600万，环比增长3%，DAU占比MAU 44.1%；微博移动端MAU已持续2年高于总体MAU增速，18Q2移动端MAU为4亿，同比增长21%，用户稳定转向移动社交，本季度移动端用户占比仍维持在93%。可见，微博平台上集聚着非常庞大的用户群体和海量的文本信息资源。这些文本信息资源与其他媒体的商品描述、论坛评论、博客留言和网络即时消息等类似，表现为片段性的描述说明、观点评论或情感抒发，是一种典型的短文本，具有很短的文字内容。

近几年在各种国际国内重要刊物和会议上出现了越来越多的关于微博短文本的研究成果。目前，关于微博短文本的研究主要从两个方面展开：一是研究微博短文本的预处理技术，即研究原始数据的压缩、筛选、表示和选择等一系列过程的处理方法，为后续微博短文本学习及应用研究提供良好的数据环境；二是通过分析微博数据的特点构建分类、聚类或其他学习模型，以充分挖掘出短文本内容之间的内在联系，从而帮助用户发现隐藏在微博数据中的潜在规律或解决生活中的现实问题，例如热点事件发现、意见领袖识别、网络内容监测、消极网络舆情检测等。

本文通过对近几年出现在国内外重要期刊和会议上的短文本研究成果进行连续跟踪、学习和梳理，深入分析和归纳短文本预处理和分类、聚类和情感分析等算法及应用的研究现状，并对短文本研究目前存在的问题进行总结，进一步探讨短文本的研究前景。

## 短文本概念及分类过程

2.1 短文本概念

短文本顾名思义即较短的文本，一般认为200字以内的都可以算做短文本。微博和twitter之前都将其字数限制在140字以内，目前两者都放宽了字数限制，但是还是对字数有一定字数限制。

常见的短文本包括微博推特等社交媒体内容，网络引擎搜索结果，手机短信内容，QQ和微信等即时消息，评论内容，文档摘要，电子邮件主题等。



2.2 短文本的特点

稀疏性：每条短文本形式信息的长度都比较短，都在 200 字以内，因此所包含的有效信息也就非常少，造成样本的特征非常稀疏，并且特征集的维数非常高，很难从中抽取到准确而关键的样本特征用于分类学习。

实时性：在互联网上出现的短文本形式的信息，大部分都是实时更新的，刷新速度非常快，聊天信息、微博信息、评论信息等，并且文本数量非常庞大。

海量性：短文本大量存在于人们的生活中，由于短文本的及时更新和快速传播，使互联网中积累了海量的短文本数据，这要求对于短文本的处理计算必须具有很高的速度。

不规范性：短文本形式的信息用语不规范，其文法通常是非正式的，语言偏向口语化和生活化，还常带有缩写、拼写错误、不规范用语、噪音，而且包含流行词汇较多，造成了噪声特征非常多，如“鸭”代表“呀”，“94”代表“就是”，“88”代表“再见”，“童鞋”代表“同学”，而且更新很快，如流行词“舔狗”、“真香”、“同九何秀”、“老铁” 等等，有的还会包含各种各样表情，增加了用户对信息理解和事件发现 的困难程度[1]。

2.3 短文本分析的难点

短文被分析时因为短文本本身所具有的稀疏性，实时性，海量性，不规范性等特点会面临以下难点：

短文本特征词少，用传统的基于词条的向量空间模型表示，会造成向量空间的稀疏。另外，词频、词共现频率等信息不能得到充分利用，会丢失掉了词语间潜在的语义关联关系。

短文本的不规范性，使文本中出现不规则特征词和分词词典无法识别的未登录词，导致传统的文本预处理和文本表示方法不够准确。

短文本数据的规模巨大，在分类算法的选择上往往更倾向于非惰性的学习方法，避免造成过高的时间复杂度。

因此，短文本分类一般在预处理、文本表示、分类器的构建等环节中进行优化和改进，以提高分类效果和精度。

2.4 研究现状

迄今为止，文本自动分类技术经过20多年的发展，已能较好的解决了现实生活中的部分问题，但在短文本分类领域，国内外只进行了少量的工作，且其效果并不理想。国外对短文本研究开始相对较早，主要集中在概念相似度计算方面，有代表性的是Mehran Sahami等人提出的使用基于web语义核函数的方法和D.Metaler等人提出的基于相似性度量的方法。国内对于短文本的研究起步较晚，目前主要集中在重庆邮电大学、中国科学院等机构，重点在特征处理环节和分类算法上。常用的短文分类算法基本可分为两类：一类是基于某种规则改进分类过 程；另一类是基于外部语义信息扩充短文本的信息量，从而提高分类效果。

基于规则的方法主要是针对短文本特点，在特征提取、文本表示、分类器构 建等多个环节提出创新的方法。J Hynek提出一种基于Apriori的频繁词集分类方法来对数字图书馆中的文档摘要进行分类;Zelikovitz S在短文本分类中使用了潜在语义索引(LSI)，在创建简化向量空间时将训练数据和未标记的测试样本进行组合，使特征空间中包含了对短文本分类有帮助的语义关联;Qiang Pu和Hui He使用基于字的 N-gram 模型抽取了中文短文本中的组块，反映出短文本的语义结构和特征间的依赖关系；王细薇提出了一种基于特征扩展的中文短文本分类方法，利用 FP-Growth 算法挖掘出训练集特征与测试集特征之间的共现关系，用特征共现集来扩展短文本特征，增强短文本特征的表述能力；郭泗辉提出一种基于连接强度扩展特征的贝叶斯网络短文本分类算法，算法在考虑了连接强度的因素后，减少了本来不相关的两个节点被归类为父子关系的错误干扰，使每个节点找到的父节点更加准确，从而使文本的分类准确度得到了提升；高金勇基于迭代的 TFIDF 算法对短文本向量进行了优化；吴薇采用正则表达式作为规则生成工具，对大规模短文本进行过滤；樊兴华提出了基于组合朴素贝叶斯(NB) 和 K-近邻(KNN)分类器的两步中文短文本分类方法;闫瑞提出一种动态组合分类器的方法，每个树节点生成一个分类器，每个类别由多个节点分类器组合成的子树表示，最终得到一个树状组合分类器结构来支持分类。

基于外部语义信息的方法主要是应用通用知识库、领域词典、搜索引擎等补 充短文本中的语义信息。D Song提出一种基于信息流的领域知识库，并在此基础上进行短文本分类；X Phan提出一种半监督学习方法进行短文本分类，通过外部的网络数据源扩展短文本的词条信息，这既解决了词特征的稀疏性问题，也使得训练得到的分类器覆盖的话题范围更加广泛；P Ferragina利用 ODP(Open Directory Project)、WebKB等手工标注知识库计算查询词、网页片段等短文本相似度；M Sahami通过谷歌搜索引擎返回的结果来统计短文本片段相似度，从而丰富文本信息;宁亚辉抽取领域高频词作为特征词，借助知网从语义方面将特征词扩展为概念和义元，通过计算不同概念所包含相同义元的信息量来衡量词的相似度，从而进行分类；王盛利用知网的上下位关系有效补充了短文本语义信息量。

此外，很多学者对不同互联网应用中大量出现且形式多样的短文本数据的分类方法展开了研究。如黄永文将产品评论文本中的产品特征、用户观点作为语 义内容，并将语义内容数量和评论文本长度等加入分类特征进行产品评论的挖掘; 崔争艳结合知网本体库，将关键词映射到语义概念，并用KNN分类算法实现微博分类;王雅蕾基于信息检索思想，提出一种基于类文档排名的分类算法，适用于处理交互问答系统中的大规模问题文本;刘金岭利用对同义关系词汇归并和上下位词汇聚焦以及种子词汇的确定来实现对手机短信文本空间的降维，从而对中文短信文本进行快速的舆情预测。

### 2.5文本分类过程

文本分析过程中一般包括预处理，文本表示和文本分类三个阶段。这三个阶段中又会分别用到一些常用算法，以下介绍一下这三个阶段中经常用到的一些算法。

2.5.1 预处理阶段

文本信息属于非结构化数据，计算机无法直接进行处理，需选取适合的数学模型对文本信息进行描述，将其转化为计算机可直接处理的结构化数据，以便快速对文本进行相关方向的信息挖掘。获取文本信息后，即进入文本预处理阶段。预处理技术主要包含文本过滤、中文分词等相关技术。

2.5.2 特殊符号等干扰信息过滤

采集后的文本信息大部分包含很多特殊符号，如网站中的便签信息、广告、无用数字、标点符号，类似“@#$＆\*()一+，．／<>?”等。很多情况下，这些符号与研究的方向无关，为提高分词效率并避免不相关的干扰，文本信息需要经过过滤才能使用。

2.5.3 分词

中文分词是将由汉字组成的自然语言切分成有意义的词的过程，是搜索引擎中的关键技术之一，也是其他文本处理的基本步骤。在文本分析中，分词精确度会严重影响文本分析的效果。在英文中，有空格将单词与单词隔离开来，而在中文中，词与词之间并没有分隔符，所以中文分词技术较之英文分词技术会有所不同，难度也有所加大。目前，主流的分词技术主要包括以下几种：基于词库(字典)的分词、基于统计的分词、基于理解的分词。在文本分词中，如何识别未登录词与歧义词的分词问题是重点和难点。

(1) 基于词库的文本分词

利用一定的算法将需要分析的语句与词库中的词条相匹配，若能在词库中找到某个字符串，说明成功进行了匹配。根据扫描方向可分为正向和逆向匹配；根据字符串长度分为最大和最小匹配；根据与词性标注结合与否，分为单纯和一体化方法。其中，最常用的集中方法是正向(由左到右)最大匹配法、逆向(由右向左)最大匹配法以及最少切分法。最大匹配法以词库中最长词条长度i为初始字符串大小，取语句前(后)i个字符作为初始字符串进行字典匹配，逐个减少，直至识别匹配成功。但此种方法缺乏对生词的辨识功能，并易产生歧义。李庆虎等[2]整合了“TREE索引树”与“整词二分”方法，对词前两个字构建索引，即双子哈希结构，逐字匹配查询，便于构建与维护索引树；肖红等[3]。提出了一种基于三级索引结构的词库，通过分组和添加三级索引来提高检索速度；段小斌等[4]辅以单字附加库以提高检索速度，同时将词条匹配与词频统计相结合，以增加新词与歧义词的分辨率；吴涛等[5]将改进的双向Markov链统计方法用于更新词库，利用基于博奕树的搜索算法及有穷自动机后串最大匹配算法进行分词，实验证明，该方法可以有效提升分词速度与效率；闻玉彪等[6]对汉字建立索引，建立规范化词库，分词时找到所有以组成汉字为头的词组，再由长到短排序来进行匹配；刘芳芳等[7]提出的动态更新框架能在分词时利用互信息概念动态识别生词，并将符合规则的生词动态更新至词典，适应性良好；何国斌等[8]提出一种分词概率算法，将二分法与哈希法相结合，并引入随机数，基于最大匹配进行分词。

(2) 基于统计的文本分词

基于统计的文本分词也称为无字典分词。此方法统计相邻字在语料库中共现的频度并计算互现信息，当互现信息大于规定的某个阈值时，即认定为构成词。即认定相邻字共现的几率越大，越有可能构成词。但此种方法易将“我的”、“之一”等字组误分为词，所以需要经常结合词库方法进行分词。费洪晓等[9]介绍了互信息、N元统计模型、t测试原理等3种基于概率的词频统计基本原理，并利用自行设计的中文分词系统进行对比；翟凤文等[10]将字典与统计方法相结合，先利用正反向最大匹配法切分，再基于统计与规则相结合的原理处理单字词，能够消解歧义字段，解决高频未登录词的问题，并提出消除组合型歧义字段是尚待解决的问题；朱小娟等[11]在进行普通词频统计初步分词后，提出了一种利用SVM对歧义字段进一步分割的方法，其中利用互信息表示歧义字段，极大地提高了切割效率；田思虎等[12]首先用隐马尔科夫信号源表示信号源(即二元统计模型)进行粗分，计算出此模型有向边的权值，然后根据词长修定权值，最后运用最短路径法求出分词结果。实验结果证明，该方法可以防止过度拆分。

(3) 基于理解的文本分词

该分词方法利用计算机模拟人对句子的理解过程来识别词，在分词的同时能够通过分析句法和语义来消除歧义，语义研究是实现理论与实践突破的关键。此方法的研究涉及到人工智能与专家系统等相关知识，效果好，但实现复杂，且需借助大量笼统、复杂的汉语知识和信息进行学习训练，机器也很难直接读取。此方法发展尚未成熟，处于试验阶段。其目前包含的人工智能方法主要有专家系统、神经网络和生成-测试法。王彩荣等[13]对自动分词专家系统的框架进行了设计，设计的核心问题是如何组织知识库并设计推理机制。其中，需按照推理机制实现分词，并利用产生式规则表示的启发性知识库切分歧义

字段；人脑理解文字语言实际上是个并行与联想的过程，神经网络具有自组织与自学习性，徐秉铮等[14]据此利用神经网络模拟人脑，重点解决了歧义字段问题；林亚平等[15]将BP神经网络、概率网(PNN)等运用到自动分词系统中，并通过改变神经网络层次、隐含层、训练次数和样本数量大小，查看对分词效果的影响；

2.5.4 去停用词

停用词并没有实际含义，但有可能大量出现在中文文本中，如“我”、“这儿”、“是”等。现已建立了多种中文停用词词库(表)，如“百度停用词表”、“哈工大停用词词库”、“四川大学机器学习智能实验室停用词库”等。为了提高文本数据的处理效率，可以根据所选的停用词库(表)，在分词后适当过滤掉停用词。

### 2.6 文本表示模型

当用户以一定的方式表示出其需求之后，检索系统应根据用户的需求，在表示信息的数据（特别是非结构化的文本数据）中进行检索，获取与用户需求相关的结果集并按一定次序输出，对这个过程建模就产生了各种不同的信息检索模型。

一般地，一个信息检索模型要确定文档的表示方式、用户查询的表示方式及用户查询与文档间相关度计算的方法包括文档表示、匹配函数和输出结果等。其中，文档表示反映的是文档在系统中的存储形式，查询表示反映了用户想要表达的信息需求，匹配函数用于把经过处理的文档表示和查询表示进行匹配计算并得到结果集，结果输出则是将检索结果按照其和用户需求的相关性排序输出。信息检索模型是将文档、查询及他们之间的关系进行建模的框架，一般可由三元组F（D，Q，R（qi，dj））来进行表示，其中D是文档逻辑视图，Q是用户信息需求的逻辑视图，R（qi，dj）是一个与查询qi∈Q和文档dj ∈D有关的函数，以便决定结果集的输出顺序。

在文本的特征表示中，最常用的有模型向量空间表示模型VSM(Vector Space Mode1)，布尔检索模型，语言模型和概率检索模型。布尔检索模型特点是简单实用，但是检索结果不便按照相关性排序，空间向量模型是把每一个标引词看作一个向量，可对每一个标引词定义权重，再通过某种计算（比如计算内积）得到向量间的相似度，然后可以根据相似度进行排名。

2.6.1 传统布尔检索与扩展布尔检索模型

布尔模型是一种简单但实用的检索模型，它建立在布尔逻辑和集合论基础之上。在传统的布尔检索模型中，文档由一组标引词进行描述，标引词权重只有0和1两种取值，分别表示文档中包含和不包含该标引词，而用户查询是有一组标引词构成的布尔表达式，有布尔运算符来链接运算分量，由此注册表达用户需求和提问的提问逻辑式。检索则根据倒排检索机制找出那些文档是与查询准确匹配的。布尔检索模型以其结构简单、容易实现、快速等特点广泛地应用在检索系统中使用，大多数的检索系统都支持布尔逻辑运算。

布尔检索模型简单易用，但是其缺点是逻辑表达式较死板，检索结果不便按照相关性排序输出，为此，研究人员提出了多种改进方案。有研究人员将检索出的文档集合和用户的检索需求进行模糊逻辑比较，并按照相关性给出检索结果。有人提出一种扩展布尔检索模型，他在保持传统布尔检索结果的同时，吸引了模糊检索和向量检索的长处，不仅允许对文档向量和查询向量中的标引项甲醛，还允许对布尔表达式的连接符加权，并通过一个特殊的参数p来控制布尔操作负的严格程度，是对布尔检索和矢量处理的一种折中。

尽管布尔模型存在不足之处，但这不妨碍其在商业上的成功应用，这是因为布尔检索模型一般尽管新检索项在文档中出现与否，因而通过适当的索引其检索效率一般较高。

2.6.2 向量空间表示模型(VSM)

20世纪70年代提出的VSM模型是一个非常经典的模型。在此模型中，每个维度代表一个词(设共有i维)，向量的边长代表词的频度，每一篇文章则是一个i维矢量空间。计算两个矢量空间之间的相似度，即可得到文本相似度。但该模型未考虑到多义词与同义词问题。相似性度量函数常用的有余弦公式、TF—IDF(词频一倒文本频度)等。

向量空间的优点之一是可以通过对查询向量中的词条富裕权值，因而便于通过相似度对检索结果进行排序。但是该模型认为和查询最相关的文档是在用词规律上和查询类似的文档，这种强加的“序化”不尽合理。此外。文档之间的相似度用向量内积或余弦相似度来表示，而这种方法的理论前提是词条键应该是相互独立的，显然这种假设忽略了此条件固有的关联性及他们之间可能存在的语义联系，因此研究人员提出了很多改进措施。LSI（潜在语义索引）以奇异值分解方法来降低向量模型的维度，已被用于多种实际应用系统中，其理论假设是文档中的单词用法包含了潜在的结构，可以通过统计学方法来估计这种结构，并且词条、文档和用户查询都可以用这种潜在的结构进行描述。

### 2.7短文本主题建模方法

主题提取模型通常包含多个流程，比如文本预处理、文本向量化、主题挖掘和主题表示过程。每个流程中都有多种处理方法，不同的组合方法将会产生不同的建模结果。

本文主要介绍LDA(Latent Dirichlet Allocation), NMF(Non-Negative Matrix Factorization)

#### 2.7.1 LDA

LDA 是最常用的主题提取模型之一，因为该模型能够处理多种类型的文本数据，而且模拟的拟合效果非常易于解释。

直观上来看，LDA 根据不同文档中词语的共现频率来提取文本中潜在的主题信息。另一方面，具有相同主题结构的文本之间往往非常相似，因此我们可以根据潜在的主题空间来推断词语之间的相似性和文档之间的相似性。

LDA 算法中主要有两类参数：

* 每个主题中各个关键词的分布参数
* 每篇文档中各个主题的分布参数

与 SVD 模型不同的是，LDA 模型所提取的主题非常好解释。以 clearcut-topics 数据集为例，LDA 模型中每个主题都有明确的关键词，它和SVD主要有以下两个区别：

* LDA 模型中可能存在重复的主题
* 不同的主题可以共享相同的关键词，比如单词 “we” 在所有的主题中都出现了。

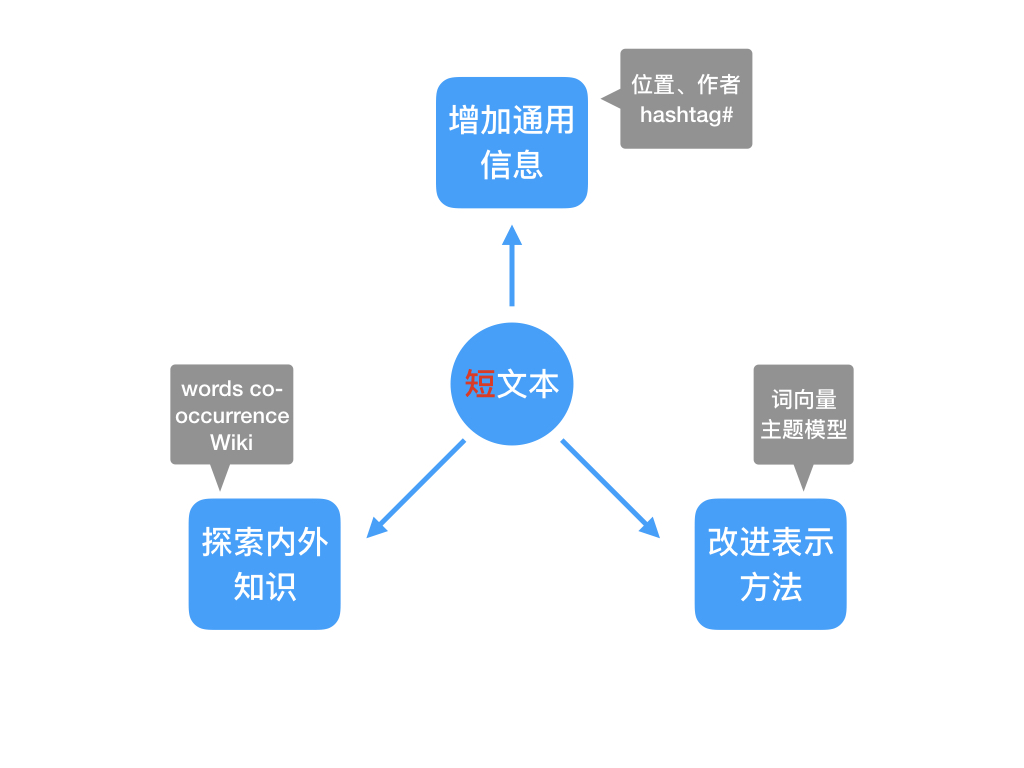
此外，对 LDA 模型来说，采用不同的文本向量化方法也会得到不同的结果。

#### 2.7.2 NMF

NMF 可以视为 LDA模型的特例，从理论上来说，这两个模型之间的联系非常复杂。但是在实际应用中，NMF 经常被视为参数固定且可以获得稀疏解的 LDA 模型。虽然 NMF 模型的灵活性不如 LDA 模型，但是该模型可以很好地处理短文本数据集。

另一方面，NMF 最大的缺点是拟合结果的不一致——当我们设置过大的主题个数时，NMF 拟合的结果非常糟糕。相比之下，LDA模型的拟合结果更为稳健。

## 文本分类算法

短文本分类算法相对于传统文本算法主要从三个方面做改进。

文本分类算法主要包括两种:一种是有指导的训练，用带有类标识的样本进行训练，也称为有监督分类;另一种是无指导的训练，训练文本无类别标识，也称无监督分类。本文主要研究有指导的文本自动分类算法。分类算法就是通过构造某种分类模型(也称为分类器)，并以此来判断待分类文本所属的类别。常见的文本自动分类算法包括:类中心分类法(K-means)、K-近邻算法(KNN)、贝叶 斯算法(Naive Bayes)、决策树(Decision Tree)、神经网络(NNS)和支持向量机算法(SVM)等。Yiming Yang 对几种常用的文本分类算法进行了分析和比较，结果表明，朴素贝叶斯、K-近邻、支持向量机是三种较好的文本分类算法[[1]](#endnote-1)。

### 3.1 KNN算法

KNN （K-Nearest Neighbor） 代表 k 个最近邻分类法通过 K 个最与之相近的历史记录的组合来辨别新的记录KNN 是一个众所周知的统计方法 在过去的 40 年里在模式识别中集中地被研究 KNN 在早期的研究策略中已被应用于文本分类 是基准 Reuters 主体的高操作性的方法之一 其它方法 如 LLSF 决策树和神经网络等。

该算法的基本思路是 在给定新文本后 考虑在训练文本集中与该新文本距离最近 最相似 的 K 篇文本 根据这 K 篇文本所属的类别判定新文本所属的类别 具体的算法步骤如下：

(1)将测试文本集合表示为特征向量空间，将待分类文档表示为文本的特征向 量。

(2)计算待分类文本与训练文本集中每篇文本的相似度，相似性计算公式可以采用余弦夹角、Dice 系数、Jaccard 系数、欧式(Euclid)距离、明 氏(Minkovski)距离、曼氏(Manhattan)距离等度量方法。然后从训练文档集中选出与待分类文档最相似的 k 个。 k 值是一个经验值，一般先确定一个初始值， 然后根据实验测试的结果进行调整。

(3)在待分类文本的 k 个近邻中，依次按下式计算对类的归属值 p :

其中，d为待分类文本的特征向量，为d与训练文档的相似度，是类作用函数，当属于类时，该函数值为1，否则为0。

(4) 比较各类的 p 值，将待分类文档分到 p 值最大的那个类别中。

KNN算法是一种简单有效的分类算法，它重新训练的代价较低，计算时间和空间线性于训练集的规模（在一些场合不算太大），由于KNN方法主要靠周围有限的邻近的样本，而不是靠判别类域的方法来确定所属类别的，因此对于类域的交叉或重叠较多的待分样本集来说，KNN方法较其他方法更为适合。所以该算法比较适用于样本容量比较大的类域的自动分类，而那些样本容量较小的类域采用这种算法比较容易产生误分。

但是它也有一些明显的缺点，KNN算法是懒散学习方法，计算量较大。目前常用的解决方法是事先对已知样本点进行剪辑，事先去除对分类作用不大的样本。当样本不平衡时，如一个类的样本容量很大，而其他类样本容量很小时，有可能导致当输入一个新样本时，该样本的K个邻居中大容量类的样本占多数。该算法只计算“最近的”邻居样本，某一类的样本数量很大，那么或者这类样本并不接近目标样本，或者这类样本很靠近目标样本。无论怎样，数量并不能影响运行结果。可以采用权值的方法（和该样本距离小的邻居权值大）来改进。

### 3.2 朴素贝叶斯算法

朴素贝叶斯(Naive Bayes，NB)分类器是一种基于贝叶斯分析的分类器。该算法是基于概率统计的算法，它以词在文本中出现的比率作为它属于某个类别的概率，并综合考虑诸特征词属于各个类别的概率来计算文本属于各类别的概率，最后将文本归入所属概率最大的类别。具体计算步骤如下:

(1)计算每个特征词属于各个类别的概率。在训练阶段，对于每个类别 j 就形成了向量，其中表示第i个特征词属于第j类的概率，计算方法为:

其中|V|表示类 中不同的特征词的个数，|D|表示类的训练文本的个数， 表示第i个特征词。

(2)在分类阶段，根据新文本中每个特征词属于各类别的概率综合计算属于类的概率。计算公式如下:

其中，表示类出现的概率。是类包含文本的概率。|C|为类的总数，n为特征词总数，为特征词在

(3)比较新文本属于各个类的概率，将文本分到所属概率最大的那个类别中。

### 3.3支持向量机

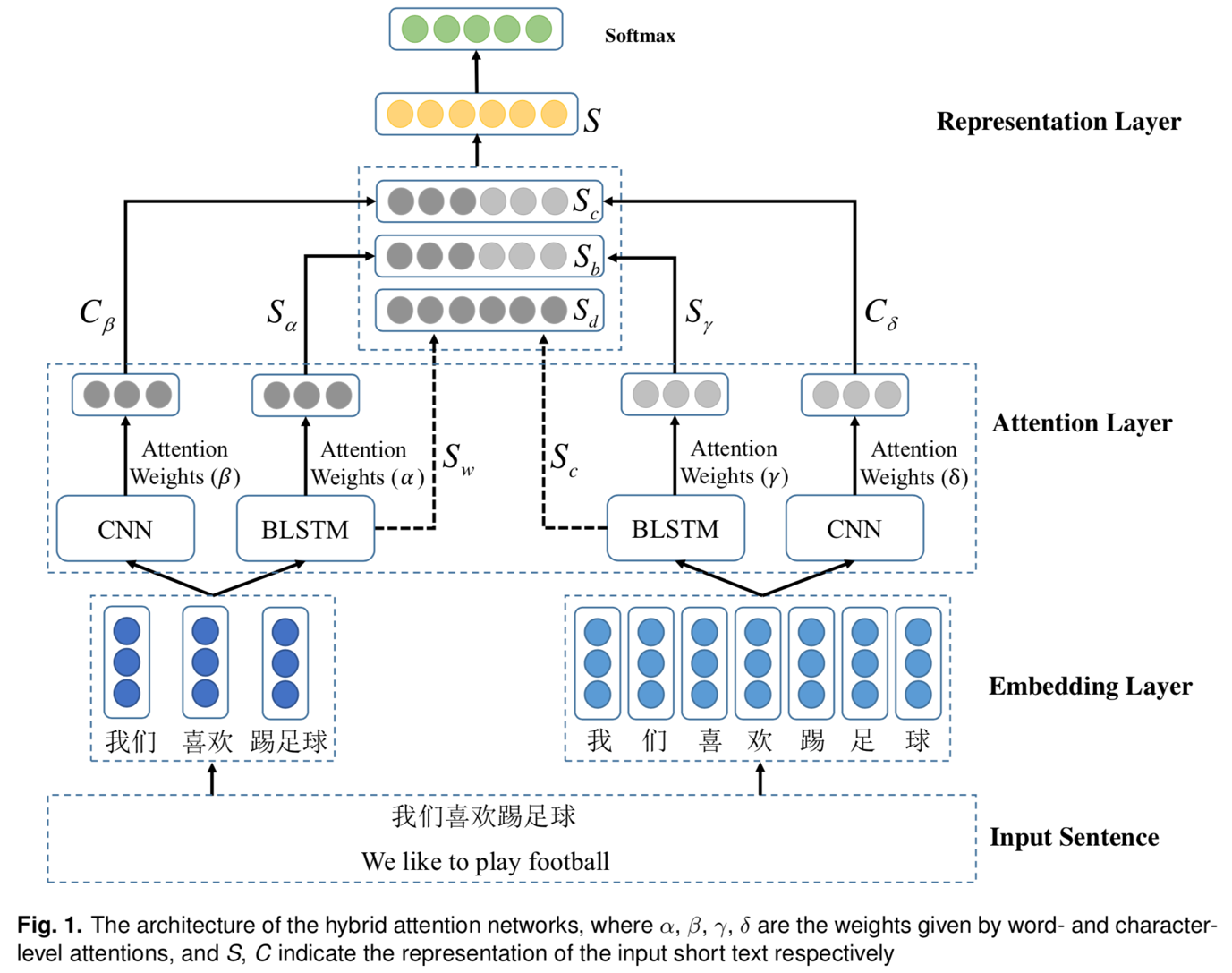
支持向量机(Support Vector Machine, SVM)是建立在统计学习理论基础上发展 而来的一种机器学习方法[[2]](#endnote-2)，由 Vapnic在1995 年提出，它基于结构风险最小化 原理，将原始数据集合压缩到支持向量集合，学习得到分类决策函数。给定一组训练实例，每个训练实例被标记为属于两个类别中的一个或另一个，SVM训练算法建立一个将新的实例分配给两个类别之一的模型，使其成为非概率二元线性分类器。SVM模型是将实例表示为空间中的点，这样映射就使得单独类别的实例被尽可能宽的明显的间隔分开。然后，将新的实例映射到同一空间，并基于它们落在间隔的哪一侧来预测所属类别。

除了进行线性分类之外，SVM还可以使用所谓的核技巧有效地进行非线性分类，将其输入隐式映射到高维特征空间中。

### 3.4 Hybrid Attention Networks[[3]](#endnote-3)

本文的主要贡献是一种新颖的神经结构，即混合注意网络（HAN），旨在提供两种见解，分析哪些字符或单词有助于中文短文的分类决策。 首先，由于中文短文由单词或字符组成，我们分别用单词和字符嵌入为每个句子建立基于RNN和CNN的表示，然后将它们连接成句子表示。 其次，可以观察到通常存在一组显着的字符和单词来表示每个句子类。 此外，相同的单词或字符在不同的句子中可能是不同的重要。

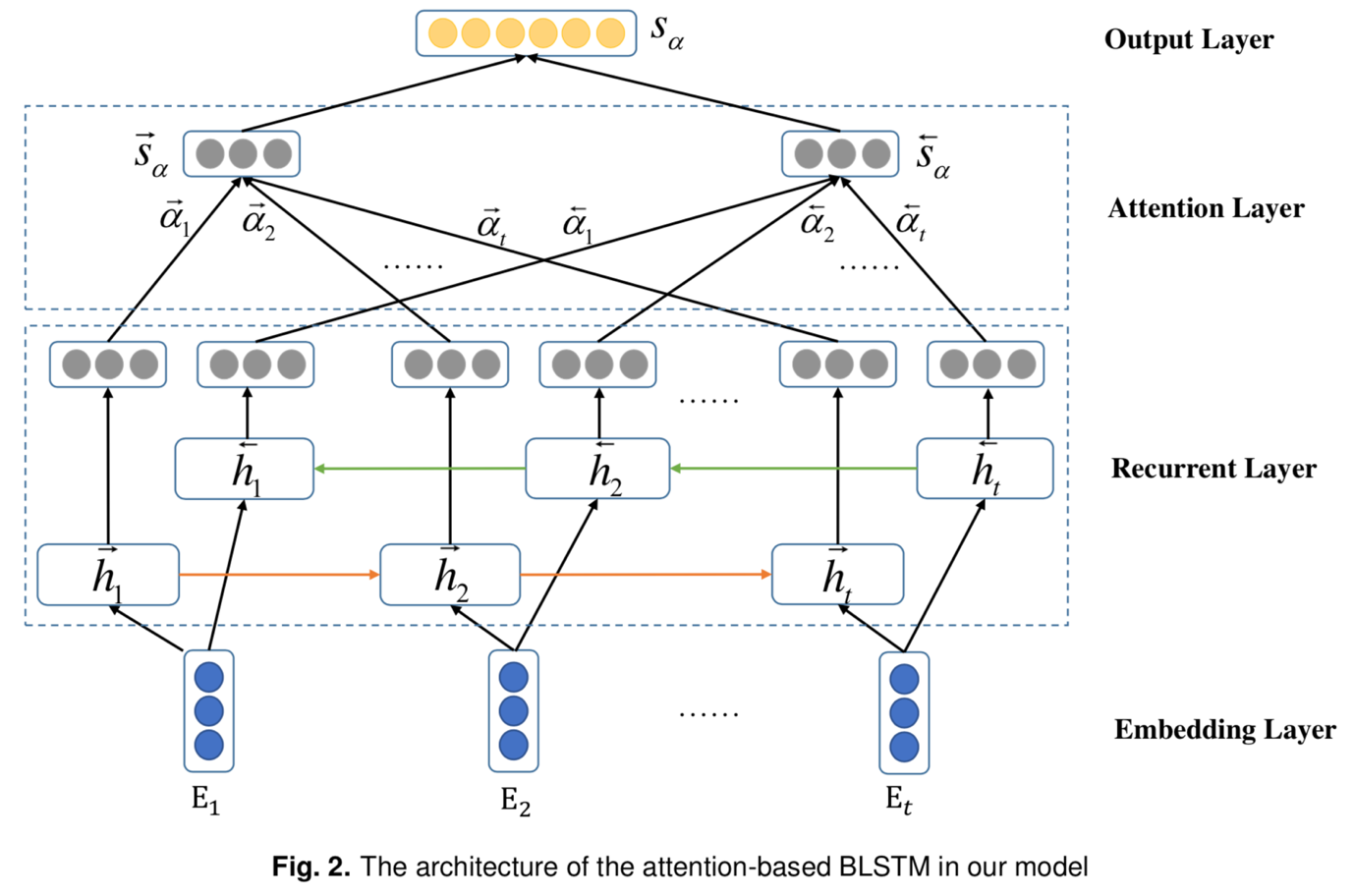
考虑到这种敏感性，我们的模型引入了两种机制，即上下文，即字符和单词级别的注意力，旨在在构造字符和单词级别表示时或多或少地注意特殊字符或单词，然后将这两个级别的注意力模型连接成句子的细致表示。

模型包括四个部分：嵌入层，关注层，表示层和分类层。如下图所示。

(1)在模型中，我们应用RNN和CNN分别用单词和字符的序列来进行中文短文本表示。RNN背后的关键思想是利用顺序信息。它可以通过递归地输出依赖于先前计算的隐藏状态向量将任意长度的输入文本映射到固定大小向量。 然而传统的RNN存在梯度爆炸或消失的问题，其中梯度向量随着它们传播到较早的时间步长而呈指数增长或衰减。 因此难以训练RNN以捕获序列中的长距离依赖性。 为了解决这个问题，Hochreiter和Schmidhuber提出了一种特殊类型的RNN长短期记忆（LSTM），它引入了一个存储单元和三个门（即输入门，忘记门和输出门）控制信息在网络中的传递。

然而，标准LSTM网络按时间顺序处理序列，当预测输入文本的开头或中间的语义含义时，其不能捕获来自未来上下文的语义依赖性。双向LSTM提供了一种有效的方法，它可以通过涉及两个单独的隐藏层来访问前一个和后一个上下文，即一个是前向LSTM而另一个是后向LSTM。因此，该模型能够捕获过去和未来的上下文信息。在该模型中，输入短文本序列被馈送到前向LSTM层，并且输入序列的反向被馈送到后向LSTM层。

(2)注意力网络

我们注意到并非所有的单词或字符都对句子意义的表示有同等作用。 当读一个句子时，人们通常可以粗略注意到有些字或者词更重要。 我们在模型中使用注意力机制从词语和字符两个层面实现这个想法。注意力网络如下图所示。

## 聚类算法

目前聚类算法的种类多种多样,按照逻辑划分主要包括基于网格的方法、基于层次的方法、基于划分的方法、基于密度的方法和基于模型的方法

1、基于网格的聚类方法

该类算法的典型特点是将处理对象由原始数据点转化为自行划分的网格单元,因此称为 基于网格的方法。该类算法主要流程有:首先通过将数据空间的每一维平均分割成等长的区间段,再将数据空间划分成不相交的网格单元,由此可见同一单元中的点属于同一类的可能性比较大,所以落入同一网格中的点就被视为一个对象来进行处理,以后所有的聚类操作都是基于网格单元进行的,因为一般来说网格单元的数量要小于原始数据点的数量,从而使用网格聚类,其效率得到了很大的提高,但网格聚类算法也有劣势,该算法更依赖于参数

2、基于层次的聚类方法

该类算法是对给定的数据集进行分层次的分解,直到某种条件满足为止,该条件由数据处理者定义,类似于数据分析中设置的阈值。根据不同的数据,该类方法采用相应不同的相似性计算方法,层次聚类方法可以是基于距离的或基于密度或连通性的,同时该类方法的 些扩展也考虑了子空间聚类,为了减少计算成本,层次聚类方法拥有一个严格的规则,即 旦一个步骤(合并或拆分)完成,它就不能被撤销,这也是该类算法的劣势所在

3、基于划分的聚类方法

划分方法的主要内容包含了选代的思想。该类算法指的是给定一个有N个元组或者纪 录的数据集,分裂法将构造K个分组,每一个分组就代表一个聚类,K<N。而且这K个分 组满足下列条

(1)每一个分组至少包含一个数据纪录

(2)每一个数据纪录属于且仅属于一个分组。

对于给定的K,算法首先给出一个初始的分组方案,以后通过反复选代的方法在保持K 值不变的情况下不断的完善分组情况,使得每一次改进之后的分组方案都较前一次好,而划 分法所谓好的标准就是:同一分组中的数据相似度越高越好,而不同分组中的数据相似度越

皆名的k- means算法属于基于划分的聚类算法。其中,k代表聚类算法中类的个数, means表示该算法是一种均值算法,k- means算法从字面理解即用均值算法把文本聚成k个簇。K- means算法的具体内容在下文再进行详细描述

4、基于密度的聚类算法

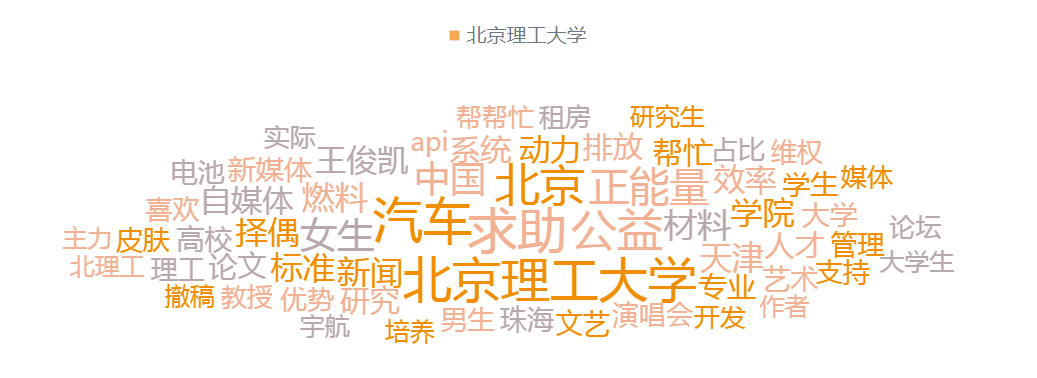
该类算法的典型特点是它不是基于各种各样的距离的,而是基于密度的。使用基于距离的相似度计算方式聚类的一大特点就是数据往往呈现“类圆形”,而基于密度的方法打破了这一局限。该类方法的核心思想是,只要一个区域中的点的密度大过某个阅值,不管是形成什么样的形状,都把它加到与之相近的聚类中去

5、基于模型的聚类算法

该类方法的主要思想是先为每个类簇定下框架或模型,然后去寻找满足该模型的数据集 进行填充。该方法的一个潜在假定就是:待处理数据集是由一系列的概率分布所决定的。这 样一个模型可能是数据点在空间中的密度分布函数或者其他。通常有两种研究方向:基于统计的方案和基于神经网络的方案。

## 短文本分析应用

短文本分析能应用于诸多领域。限制在140字以内的微博是典型的短文本。此外还有电商网站下的商品评论、搜索引擎中输入的搜索关键词、互联网聊天信息、电子邮件主题、商品描述信息、文献摘要等。本文重点以前两者做重点介绍。

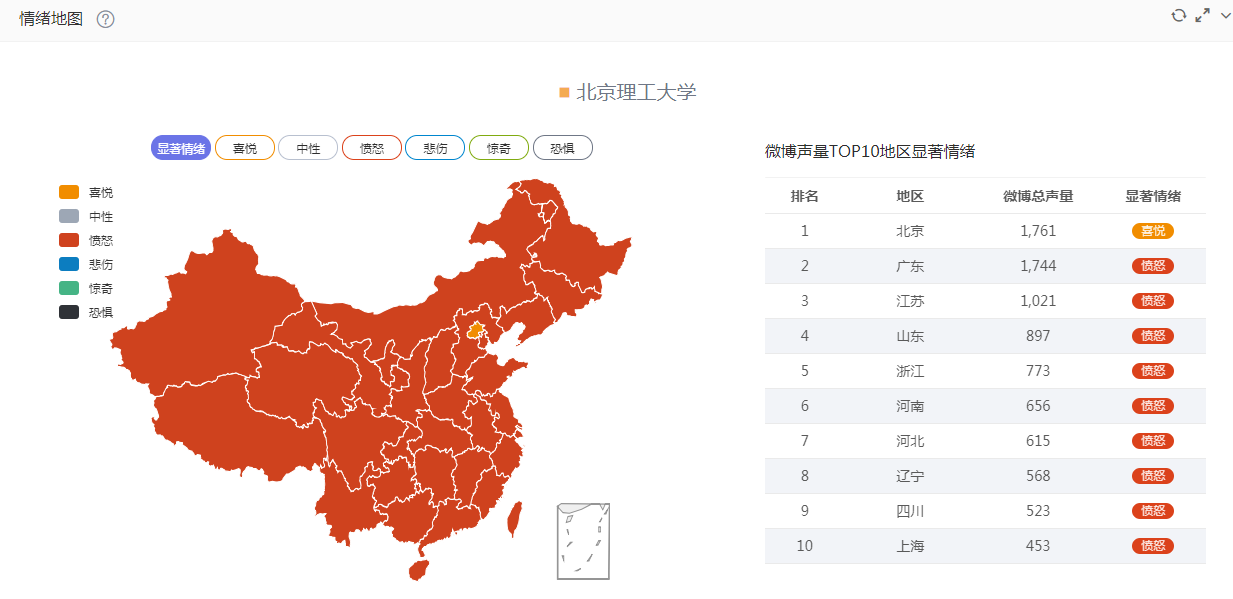


以“北京理工大学”为关键词建立的词云

词云能直观反映出对象实体的一些关键特性。



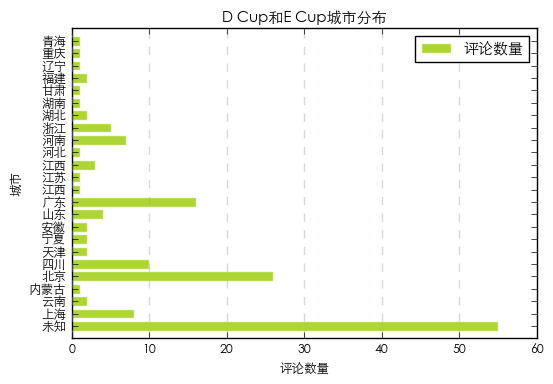
教育话题下的热度排名



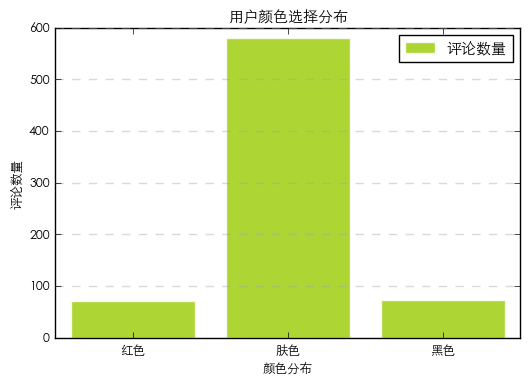
对北京理工大学建立情绪地图

对全国各地与北京理工大学有关的微博做情感分析以后就可以以此建立一个情绪地图。情绪地图能够直观反映出用户对某一事件或者个体的情感态度。

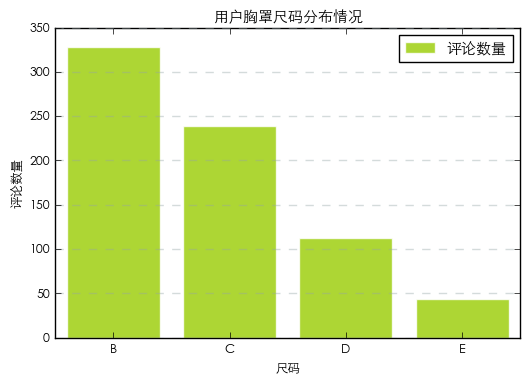
此外，短文本分析也应用于对电商网站的商品评论分析上。以下案例为使用爬虫工具爬取京东网站上的评论后，分类别画出的图示。



分省区的商品购买分布



商品颜色分布



尺码分布

对这些商品评论分类别做出分析后就可以做出相应的市场调整。

参考文献

[1] 王连喜.微博短文本预处理及学习研究综述[J].图书情报工作,2013,57(11):125-131.

[2] 李庆虎，陈玉健，孙家广．一种中文分词词典新机制——双字哈希机制[J]．中文信息学报，2002，17(4)：1318．

[3] 肖红，许少华，李欣．具有三级索引词库结构的中文分词方法研究[J]．计算机应用研究，2006(8)：49—51．

[4] 段小斌，林雯，阮百尧，等．一种基于三级索引词库结构的中文分词方法研究[J]．计算机与数字工程，2007，35(7)：47—49．

[5] 吴涛，张毛迪，陈传波．一种改进的统计与后串最大匹配的中文分词算法研究[J]．计算机工程与科学，2008，30(8)：79—82．

[6] 闻玉彪，贾时银，邓世昆，等．一种改进的最大匹配中文分词算法[J]．计算机技术与发展，2011，21(10)：92—94．

[7] 刘芳芳。王晶，沈奇威．一种引入动态词库更新的中文分词架构

[J]．计算机系统应用，2013，22(3)：100～103.

[8] 何国斌，赵晶璐．基于最大匹配的中文分词概率算法研究[J]．计算机工程，201o，36(5)：173—175.

[9] 费洪晓，康松林，朱小娟，等．基于词频统计的中文分词的研究[J]．计算机工程与应用，2015(7)：67—68．

[10] 翟凤文，赫枫龄，左万利．字典与统计相结合的中文分词方法[J]．

小型微型计算机系统，2006，27(9)：1766—1771．

[11] 朱小娟，陈特放．基于SVM的词频统计中文分词研究[J]．微计算机信息，2007(30)；205—207．

[12] 田思虎，李德华，潘莹．一种改进的基于二元统计HMM分词算法[J]．计算机与数字工程，2011，39(1)：14—16．

[13] 王彩荣．汉语自动分词专家系统的设计与实现[J]．微处理机，2004(3)：56-57．

[14] 徐秉铮，詹剑，贺前华．基于神经网络的分词方法口]．中文信息学报，1992，7(2)：3644．

[15] 林亚平，李彦，童调生，等．汉语自动分词中的神经网络技术研究[J1．湖南大学学报，1997，24(6)：95—101.

[16] 张宁, 贾自艳, 史忠植. 使用 KNN 算法的文本分类[J]. 计算机工程, 2005, 31(8).

[17] Yang Y M，Liu X. A re-examination of text categorization methods[C]. In:

Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, BerkeIey, USA. August, 1999.

[18]Kononenko I, Simec E. Induction of Decision Trees Using RELIEFF[J]. Mathematical and statistical methods in artificial intelligence. Springer Verlag, 1995:1-19.

[19]Vapnik V. The Nature of Statistical Learning Theory[M]. Springer-Verlag, New York: 2000.

[20]ZHOU Y, XU J, CAO J. Hybrid Attention Networks for Chinese Short Text Classiﬁcation[J]. Computación y Sistemas, 2018, 21(4).

1. Yang Y M，Liu X. A re-examination of text categorization methods[C]. In:

   Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, BerkeIey, USA. August, 1999. [↑](#endnote-ref-1)
2. Vapnik V. The Nature of Statistical Learning Theory[M]. Springer-Verlag, New York: 2000. [↑](#endnote-ref-2)
3. ZHOU Y, XU J, CAO J. Hybrid Attention Networks for Chinese Short Text Classiﬁcation[J]. Computación y Sistemas, 2018, 21(4). [↑](#endnote-ref-3)