Statistics and Application 统计学与应用, 2019, 8(4), 589-597
Published Online August 2019 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2019.84067

Hans汉斯

Study on the Text Categorization of Engineering Geological Investigation

Chaoguo Tang[1], Yongbiao Zhu[2], Bo Xie[3], Yu Wu[3], Jie Meng[3*]

[1]Information Technology Center, China Railway Eryuan Engineering Group Co., Chengdu Sichuan
[2]Geotechnical Engineering Design and Research Institute of Geological Prospecting, China Railway Eryuan Engineering Group Ltd., Chengdu Sichuan
[3]School of Mathematics and Statistics, Yunnan University, Kunming Yunnan
Email: tangcg@ey.crec.cn, *691669246@qq.com

Received: Jul. 16th, 2019; accepted: Jul. 28th, 2019; published: Aug. 5th, 2019

Abstract

With the development of information technology, electronic text information is increasing. Automatic text categorization is a key technology that can facilitate users to obtain the required information accurately in the mass text information resources. It enjoys a wide application in various fields. From the perspective of improving the classification accuracy, this paper used the "Teleological Survey Specification for Railway Engineering Geology" (TB1002-2007) as the training standard. Firstly, based on the word segmentation principle of natural language processing (NLP), the text document is segmented by computer and human. Then the feature reduction technique is applied to the word segmentation results of text documents. The words with high word frequency are selected as the final geological survey corpus. The corpus contained geological terminology. Finally, machine-learning methods are used to automatically classify the text after word segmentation. After comparing the classification results of various classification algorithms, this paper finds that selecting K-nearest neighbor classifier is more ideal than the others due to unevenness of geological exploration data classification.

Keywords

Automatic Text Categorization, Natural Language Processing, Feature Dimension Reduction

工程地质勘察文本的分类研究

唐朝国[1]，朱泳标[2]，解　波[3]，吴　宇[3]，孟　捷[3*]

[1]中铁二院信息技术中心，四川 成都
[2]中铁二院地勘岩土工程设计研究院，四川 成都
[3]云南大学数学与统计学院，云南 昆明
Email: tangcg@ey.crec.cn, *691669246@qq.com

*通讯作者。

文章引用: 唐朝国, 朱泳标, 解波, 吴宇, 孟捷. 工程地质勘察文本的分类研究[J]. 统计学与应用, 2019, 8(4): 589-597.
DOI: 10.12677/sa.2019.84067

收稿日期：2019年7月16日；录用日期：2019年7月28日；发布日期：2019年8月5日

摘　要

伴随着信息技术的不断发展，电子文本信息日益增多，文本自动分类作为处理海量文本信息，方便用户准确搜索所需信息的关键技术，其应用十分广泛。本文从提高分类准确率的角度出发，以《铁路工程地质勘察规范》(TB1002-2007)的电子文本文档为训练标准，运用自然语言处理的分词原理对文本文档进行计算机与人工结合分词，然后针对文本文档的分词结果进行特征降维技术处理，对词条计算词频后，根据词频大小筛选出词频较高的词语作为最终的地质勘察语料库，该语料库包含了地质专业相关术语。最后利用机器学习对分词后的文本文档进行自动分类，在对比多种分类算法的分类结果后，本文发现针对地质勘探数据类别不平衡性，选择*K*近邻分类器对文本文档分类的效果较为理想。

关键词

文本自动分类，自然语言处理，特征降维

1. 引言

随着电子文本信息日益增多，快速、准确地从海量的文本中找到所需信息，并有效地组织和管理这些文本信息成为信息技术面临的一大挑战。文本自动分类作为处理海量文本信息的关键技术，可以在很大程度上解决信息复杂混乱的问题。把相关信息资源归类，这样既方便用户准确地搜索到所需的信息，又能实现数据的有效管理。文本自动分类作为信息过滤、信息检索、信息抽取等的技术基础在各个领域都有着广泛的应用，因此对文本自动分类的研究十分重要。

国外在文本分类方面的研究起步较早，实际应用成果较多，从 1995 年开始进入快速增长期，应用机器学习做文本分类的算法逐渐增多。国内对文本自动分类的研究起步较晚，从 2000 年才开始进入快速增长期，并且由于中英文之间的差距，国外一系列文本自动分类成果在国内实用性较弱，只能参考其研究思想。因此国内更注重对中文文本自动分类系统的研究。郑州大学张金瑞[1]提出了基于 LDA 的弱监督文本分类算法 VB-LDA (Latent Dirichlet Allocation with Vector and Bigram)并将该算法应用到文本分类中，首先获取主题的高频词和类别的代表词，然后利用词向量化工具将它们都转化成相应的词向量，最后用距离度量来计算出每篇文档中概率最大的主题所对应的类别。广东工业大学黄瑜青[2]提出将支持向量机与文本自动分类器相结合，以解决文本自动分类中维数庞大、线性不可分和分类性能不高的问题。石佳[3]等人提出一种基于 N 元语法的汉语自动分词系统，将分词与标注结合起来，用词性标注来参与评价分词结果。中科院陈建英[4]提出一种改进的面向地名知识库的双向最大匹配算法并设计了一个面向中文地名的知识库。通过对大量中国地名信息的分析和研究，并参考中国的行政地域特点，采用地名词分级思想，将所有的地名词进行层级划分，并结合目前互联网的词库资源和主流的数据存储技术，得出了一个全新的地名知识库。北京交通大学邬启为[5]选用了基于密度方法的聚类算法 OPTICS (Ordering Points To Identify the Clustering Structure)对网页文本聚类，该方法比起其它聚类方法，可以发现不同形状的文本簇，

并且还能过滤离群点。刘海峰[6]等人提出了一种聚类模式下的 KNN，使用一种改进的聚类方法对文本特征集进行初步筛选，随后使用一种基于类别的改进 KNN 分类器进行分类，减少了噪声样本对测试样本类别判定的干扰。以上这些研究大部分是分类算法改进和在计算机系统等方面做研究，而未曾在铁路工程地质勘察数据方面做相关研究。因此本文将从建立地质勘察语料库、对地质勘察文本文档进行分类两方面对地质勘察文本做相关研究。

2. 数据处理

2.1. 数据来源

本文数据主要来源于重庆到怀化增建二线铁路工程地质勘察资料。主要抽取每段里程中说明地质情况的文本文档进行研究分析。其中主要包括工程地质勘察报告、工程地质说明书、地质第四篇、调查表等 21 类文本文档，总的文本数量共有 547 篇文本文档，分为 21 类。为了提高后续分类的准确度，本文部分文件的名称根据文件内容来划分其所属类别，各类地勘文本文档的名称及相对应的数量如下表(表 1)所示：

Table 1. Name and quantity of all geological prospecting documents
表 1. 所有地勘文档名称及数量(单位：篇)

名称	数量	名称	数量
地质说明书	298	地质素材	2
勘察报告	96	断面排版	2
调查表	47	解译说明	2
互提资料单	41	通知单	2
测井报告	15	测试报告	1
地质第四篇	8	地质概况	1
移交登记表	8	情况说明	1
汇总一览表	6	审查表格	1
照片集	6	文字报告	1
区测报告	5	资料说明	1
勘测说明	3	**合计**	**547**

从上表可以看出，总的文本数量有 547 篇，地质说明书 298 篇，工程地质说明书 96 篇，调查表 47 篇，从表中可以看出各类文本文档的数量相差很大。就文本内容而言，短文本和长文本的内容相差较大，为提高文本自动分类准确率，需考虑类别平衡性的划分及文本文档的长度。

2.2. 类别标签的处理

在做分类之前首先要给每个类别系上标签，本文通过文档的文件名截取部分有代表性的字符串作为类标签。但是在截取过程中，由于不同的操作人员撰写报告的方式不同。因此，其报告也是各有特点，命名不规范的各类文本文档的处理方式如下：

1) 地质勘察报告

在地质勘探文本文档中，地质勘察报告一类的数量相对较多，本文把文件名中含有“地质勘察报告”、

“工程地质说明书”、“地质说明书”、“地质第四篇”、“第四篇(送审稿)”、“第四篇地质”、“第四篇(正式稿)”字符串的文本文档归属于地质勘察报告，该类别文件中对文件的命名相对规范，需要处理的文本文档较少。

2) 调查表

调查表一类文本文档的命名最复杂，本文把文件名中含有以下字符串的文本文档归为调查表一类，分别是：“调查表”、“人行天桥”、“公路公交”、“调查表(2)”、“公路立交(2)”、“人行天桥(2)”、“调查表(3)”共七类不同的文件命名方式。

3) 综合测井报告

该类把文本文档名称中含“综合测井报告”字符串的文档归为综合测井报告一类。

4) 电子文件移交登记表

本文把文本文档名称中含“电子文件移交登记表”、“电子文件移交登记表(2)”、“电子文件移交登记表(3)”、“电子文件移交登记表(4)”四类字符串的归为电子文件移交登记表一类。

5) 地质素材

地质素材一类主要包含三类字符串的文本文档，分别是：“地质素材”、“地质素材(已处审)”、“地质素材(正式)审后”。

6) 其他

该类主要是文本文档数量最少、或者内容最少的文档，包含：“断面排版”、“地质概况”、“情况说明”、“审查表格”、“文字报告”、“资料说明”共六类文本文档。其中“断面排版”文本文档记录里程数，内容里面的文字甚少；“情况说明”和“资料说明”仅记录一句话，上述六类文本文档被归为其他类别。

剩余九类文本文档的命名较为规范，尚未做任何处理，截取部分字符串作为类别标签即可。

3. 建立地勘语料库

3.1. 文本处理的理论基础

3.1.1. 文本预处理

中文文本预处理过程主要是对文本文档进行分词，分词过程中所采用的技术主要是自然语言处理。自然语言处理相关技术有分词、句法分析、语义分析、篇章分析等[7]。在自然语言处理过程中,因为词是最小可以独立活动的有意义的语言成分[8]，因此在进行上述分析之前，最重要的是进行分词。目前中文分词算法主要分为三种：基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法[9] (主要运用 N 元文法模型、互信息和隐马尔科夫模型等统计模型)。下面主要叙述隐马尔科夫模型的分词原理。

马尔科夫模型是隐马尔科夫模型的基本条件，马尔科夫模型是用来描述随机过程随着时间的变化而随机变化的一类随机过程[10]。假定某一序列有 N 个可列举的状态 $S=\{s_1,s_2,\cdots,s_N\}$，则随着时间的改变，该序列将会从一种状态转为另外一种状态。若随机变量序列为 $Q=\{q_1,q_2,\cdots,q_T\}$，该随机变量序列的取值为状态集 S 中的某个状态。若在时间 t 时其状态标记为 q_t，若要对该序列作描述，则需要给出当前时刻的状态，该时刻前各个时刻所对应的状态之间的关系，随机变量序列在 t 时刻处于 s_j 状态的概率取决于时刻 $1,2,\cdots,t-1$ 的状态，其概率用公式表示为：

$$P\left(q_t=s_j \mid q_{t-1}=s_i,q_{t-2}=s_K,\cdots\right)$$

假设在某一特定条件下，观测变量的取值在 t 时刻的状态仅与在 $t-1$ 时刻的状态相关，与 $t-2$ 时刻的状态无关，这就是所谓的马尔科夫链(Markov chain)，用公式表示如下：

$$P\left(q_t = s_j \mid q_{t-1} = s_i, q_{t-2} = s_k, \cdots\right) = P\left(q_t = s_j \mid q_{t-1} = s_i\right)$$

假设我们只考虑独立于 t 时刻的随机过程。除了上述结构信息之外，若要确定一个隐马尔科夫模型，需考虑下列三组参数：

1) 状态转移概率：模型在各个状态之间相互转换的概率，通常记为 A_{ij}，其中 $A_{ij} = P\left(q_t = s_j \mid q_{t-1} = s_i\right), 1 \le i, j \le N$

上述式子表示任意 $t-1$ 时刻，假设状态为 s_i，求下一时刻状态为 s_j 的概率，同时 A_{ij} 需满足下列几个条件：① $A_{ij} \ge 1$；② $\sum_{i=1}^{N} A_{ij} = 1$

2) 输出观测概率：模型从状态 s_j 观察到符号 V_k 的概率分布矩阵，记为：$B = \left\{b_j(k)\right\}$，其中 $b_j(k) = P\left(O_t = V_k \mid q_t = s_j\right), 1 \le j \le N, 1 \le k \le M$

上式表示在任意时刻 t 的状态为 s_j，观测值 V_k 被获取额概率。

$b_j(k)$ 满足两个条件：① $b_j(k) \ge 0$；② $\sum_{k=1}^{M} b_j(k) = 1$

3) 初始状态概率，模型在最开始时刻各个状态的概率。记为：$\pi = \left\{\pi_1, \pi_2, \cdots, \pi_N\right\}$，其中 $\pi_i = P\left(q_1 = s_i\right), 1 \le i \le N$

上述式子表示模型初始状态为 s_i 的概率。

π_i 需满足两个条件：① $\pi_i \ge 0$；② $\sum_{i=1}^{N} \pi_i = 1$

一般情况下，我们把一个隐马尔科夫模型记为 $u\left(S, K, A, B, \pi\right)$，其中 S 为某时刻处于某种状态的集合，K 是输出符号的集合；A 为状态转移概率矩阵，决定状态序列，B 为输出观测概率矩阵，决定观测序列，π 为初始状态概率向量，决定状态序列。

3.1.2. 特征降维技术

在文本自动分类中，一般一个词条就被看作一个维度，对于篇幅相对较长的文章来说，文本文档分词后可能出现上万个词条，把每个词条都看作一个维度，最终将会形成一个高维度的特征空间，在分类过程存在矩阵的稀疏性问题。在原始的长文本文档中，中文分词后的词语个数有上万个，表示成特征空间后其维数相对较高，实现分类算法时其难度较大。由于所有的文本文档都要计算其特征空间的值的大小，因此，特征降维技术可以大大提高分类算法的效率，还可以降低噪声，提高分类效果的准确率。下面主要讲述 TF-IDF 法：

在文本文档中，一个词条的词频(Term Frequency, TF)是指文本文档分词之后，该词条出现的频数。为防止其偏向内容较长的文档，一般词频通常需要归一化，在归一化过程中，分子一般小于分母区，同一个词语在长文本文档中出现的频数可能会比在短文本文档中出现的频数大，而不管该词条的重要性有多大。

对于给定文本文档中的词条 t_i 来说，该词条的重要程度可用如下式子表示：

$$\mathrm{TF}_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

其中 $n_{i,j}$ 是词条 t_i 在文本文档 d_j 中的出现频数，而分母是所有文本文档中所有词条频数的总和。

逆向文档频率(Inverse Document Frequency, IDF)是指词条在少数文本中出现的频数比在多数文本中出现的频数大[5]。逆向文件频率对词条在文本文档中普遍重要程度的度量。某个词条的 IDF 计算公式如下：

$$\mathrm{IDF}_{i,j}=\log\frac{|D|}{\left|\left\{j:t_i\in d_j\right\}\right|}$$

其中$|D|$是文档集合中文本文档的总数，$\left|\left\{j:t_i\in d_j\right\}\right|$是词条$t_i$出现过的文本文档的数目，在实际计算过程中，$n_{i,j}\neq 0$。由于考虑到词语不在文档集合中，导致分子为零的情况。因此考虑分母+1，即分母写为：$1+\left|\left\{d\in D:t\in d\right\}\right|$

然后

$$\text{TF-IDF}_{i,j}=\mathrm{TF}_{i,j}\times\mathrm{IDF}_{i,j}$$

在一个文本文档中，高频率的词语及该词语在文本文档中的低频率文本文档，将会产生权重较高的TF-IDF。TF-IDF 是一直常见的权重计算方法，它旨在过滤常见且重要性不大的词语，保留相对重要的词语。

3.2. 构建语料库的基本步骤

1) 格式化文件。由于 Python 软件的格式化要求，首先把.doc 文本文档转换为.txt 文本文档。

2) 提取文本文档的文字部分。对转换后的.txt 文本文档，通过正则表达式去除各类标点符号、数字及英文字母等特殊符号，提取最终的文字部分。

3) 分词。利用 Python 软件中的 jieba 分词包，根据隐马尔科夫模型的分词原理对文本文档进行分词。

4) 去除停止词。去除相对不重要的语气词、助词等。

5) 连接词。根据分词的结果自定义步长(本文定义的不长为 1)，把相邻两个位置的词条拼接在一起组成一个词语。

6) 计算词频。计算拼接后的词语的词频，并按照词频降序排列。

7) 确定语料库。根据输出词频的大小，提取词频大于阈值 5 的词条作为最终语料库。

语料库中按照词频的大小降序排列后，其位置词频图(图 1)如下：

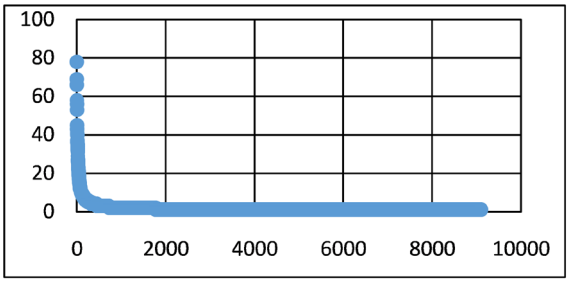

Figure 1. Location of corpus entry—word frequency graph
图 1. 语料库词条的位置——词频图

从上图可以看出，经过上述算法计算后，地质勘察规范经过一系列处理后，最终剩余九千余词条作为文本向量，把每个词条看作一个维度，这样就形成一个高维的向量空间，为降低噪声对文本文档分类准确率的影响，需要去除词频较低的词，选择词频大于 5 的词条作为最终的语料库，该语料库中涵盖 250 个词条。

4. *K* 近邻模型分类

在文本自动分类过程中，分类器的选择是影响分类准确率的一大关键因素，现在流行的文本自动分

类算法有很多，涵盖：朴素贝叶斯分类算法、支持向量机算法、K 近邻分类算法、神经网络分类算法、决策树分类算法等，这些分类算法也各有优缺点[11]。本文尝试了贝叶斯分类器、神经网络分类器、K 近邻分类器三大分类器的分类效果，从分类正确率来看，K 近邻分类器的分类正确率最高，故主要叙述 K 近邻分类的分类原理及分类结果展示。

4.1. 分类原理

K 近邻分类模型是一种最简单且最基本的分类方法，其工作原理是：给定一个测试数据集，首先定义一种距离的度量方式，在训练数据集中找出距离最近的 K 个训练样本，在这 K 个训练样本中，某个类中的训练样本最多，根据投票法则把输入训练样本归为这个类标签。其具体算法如下：

输入：训练数据集 $T=\{(x_1,c_1),(x_2,c_2),\cdots,(x_N,c_N)\}$

其中 $x_i\in X$ 为输入训练样本的特征向量，$c_i\in C=\{c_1,c_2,\cdots,c_K\}$ 为其所属的类别，$i=1,2,\cdots,N$，训练样本的特征向量为 x。

输出：训练样本 x 所属的类别 c。

1) 依据给定计算距离的方式，在训练数据集 T 中找出与 x 距离最近的 K 个点，涵盖这 K 个点的区间称为 x 的领域，记为：$N_K(x)$

2) 在 x 的领域 $N_K(x)$ 中根据投票规则(少数服从多数)判断 x 所属的类别 c：

$$c=\arg\max_{C_j}\sum_{x_i\in N_K(X)}I(c_i=C_j),i=1,2,\cdots,N;j=1,2,\cdots,K$$

其中 $I=\begin{cases}1, & c_i=C_j\\ 0, & c_i\neq C_j\end{cases}$。

在 K 近邻模型中，决定该模型的三大基本元素主要是：K 值的选择、距离的度量、分类决策规则。

在进行 K 近邻分类过程中，其特征空间属于较高维度的实数向量空间，计算两向量空间的相似度一般采用欧氏距离的计算法则，也可以采用其他计算距离法则来计算两向量空间的相似程度，常用的计算距离的方式有：欧氏距离、曼哈顿距离、Minkowski 距离等多种算法，本文主要利用欧氏距离的度量方式来计算向量空间相似度，其计算公式为：

$$d(x_i,x_j)=\left(\sum_{l=1}^{n}\left|x_i^{(l)}-x_j^{(l)}\right|\right)^{\frac{1}{2}}$$

其中 x_i,x_j 是 n 维实数向量空间 $x_i=\left(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(n)}\right)^{\mathrm{T}}$，$x_j=\left(x_j^{(1)},x_j^{(2)},\cdots,x_j^{(n)}\right)^{\mathrm{T}}$。

对于 K 值的选择会极大程度影响 K 近邻的分类效果。若 K 值选择过大，意味着用较大范围内的训练样本作预测，在学习过程可以减小误差，此时只有输入距离训练样本较远的特征向量才会对预测结果起作用，这样预测往往会发生错误。若 K 值选择过小，即用较小范围内的训练样本进行预测，这样得出的误差很小，这样只有输入与训练样本较为相似的特征向量才会对预测结果起作用。若减小 K 值，则模型从整体上变得更加复杂，容易出现过拟合现象。在 K 近邻分类工作中，往往采用多数表决的方式来决定最终的决策规则，也就是由输入训练样本的 K 个近邻训练样本中的大多数决定其所属的类标签。

4.2. 结果分析

基于以上 K 近邻算法及文本文档中自然语言处理的相关描述，对渝怀二线的 547 篇文本文档进行 K 近邻分类。由于 K 近邻分类算法属于有监督的学习，因此先对文本文档设计类标签，总的分 11 个类标签，每类的类别名称分别是：地质勘察报告、互提资料单、工程地质说明书、调查表、综合测井技术报告、

区测报告、电子文件移交登记表、地下水侵蚀汇总一览表、地质第四篇、地质素材、其他共 11 类。通过训练后，仅有三篇文本文档被判错，以下是判错文档名称及相应的预测类别表(表 2)：

Table 2. Incorrect text name and corresponding prediction category table
表 2. 判错文本名称及对应预测类别表

文档名称	实际类别	预测类别
新黄家湾隧道区测报告	区测报告	工程地质说明书
C3K49 新黄家湾隧道工程地质勘察报告	地质勘察报告	工程地质说明书
新界牌坡隧道工程地质勘察报告(设施)	其他	工程地质说明书

在 K 值取 11 的情况下，随机从原始数据中选择 80%的样本的作为训练集，20%的样本的作为测试集，训练 2000 次后，测试集的正确率达到 97.27%，在大量的实验过程中，只有当 K 值取 11 时，即把文本文档分为 11 类时，测试集的正确率才达到最高，具体名称及相应的数值如下表(表 3)所示：

Table 3. Accuracy table of test set divided into 10 classes
表 3. 分 10 类的测试集正确率表

名称	数值
测试集样本量	110
判错样本量	3
判错率	2.73%
测试集正确率	97.27%

基于数据本身考虑，由于各类文本文档内容相差较大，现考虑控制文本长度本身，即降低词条空间维度，再选择分类器训练。在最终的稀疏矩阵中涵盖 9477 个词条，属于较高纬度空间，现每篇文章选择 8000 个词语即，并计算词条的 TF-IDF，利用每个文档的 TF-IDF 为特征向量，同样从 547 篇文本文档中随机选择 80%的样本作训练集，选择 20%的样本的作测试集，训练 2000 次，其测试集的判错率如下表(表 4)所示：

Table 4. Accuracy table of test set after dimension reduction
表 4. 降维后测试集的正确率表

名称	数值
测试集数量	110
判错数量	0
判错率	0%
测试集正确率	100%

从上表可以看出：测试集的判错文本数量为 0，判错率为 0，测试的正确率为 100%，因此，在文本自动分类过程中，若文本文档类别均衡，文档内容的长短相差甚少，测试集的预测效果将会达到最理想状态。

综上所述，在文本自动分类过程，分类正确率的高低将取决于多重因素，包括分词的准确性，分类

器的选择，文本类别的平衡程度等。针对地质勘探数据类别不平衡性，选择 K 近邻分类器对文本文档分类的效果较为理想，而 K 近邻分类受多重因素的影响，当 K 取 11 时，分类效果最好。若降低文本向量空间的维度，即控制词条为 8000 甚至比此更小，分类的过程中预测正确率达到 100%，属于最理想状态。

5. 结论

本文从地质勘察的文本数据出发，深入研究了自然语言处理的分词过程，利用 Python 软件中的 jieba 分词包，根据准确率较高的隐马尔科夫模型的分词原理对文本文档进行分词。在降维技术过程中，选择词频大于阈值 5 的词条作为最终于语料库。在文本自动分类过程中，本文采用 K 近邻分类器对文档进行自动分类，分类效果较为理想。

参考文献

[1] 张金瑞. 基于 LDA 的文本自动分类研究及其应用[D]: [硕士学位论文]. 郑州: 郑州大学, 2016.

[2] 黄瑜青. 基于支持向量机的文本自动分类器的研究与应用[D]: [硕士学位论文]. 广州: 广东工业大学, 2012.

[3] 石佳, 蔡皖东. 基于 N 元语法的汉语自动分词系统研究[J]. 微电子学与计算机, 2009, 26(7): 98-101.

[4] 陈建英. 面向中文地址的分词引擎设计及实现[D]: [硕士学位论文]. 北京: 中国科学院大学(工程管理与信息技术学院), 2015.

[5] 邬启为. 基于向量空间的文本聚类方法与实现[D]: [硕士学位论文]. 北京: 北京交通大学, 2014.

[6] 刘海峰, 姚泽清, 刘守生, 等. 基于聚类降维的改进 KNN 文本分类[J]. 计算机科学, 2009, 36(11): 18-20.

[7] 化柏林. 基于 NLP 的知识抽取系统架构研究[J]. 现代图书情报技术, 2007, 2(10): 38-41.

[8] 刘建培. Chinese Split Word Design Based on Delphi 基于 Delphi 的中文分词设计[J]. 计算机系统应用, 2009, 18(3): 156-160.

[9] 吴巧玲. 中文分词算法在自然语言处理技术中的研究及应用[J]. 信息与电脑: 理论版, 2011(12): 39-40.

[10] 郭武, 朱明明, 杨红兵. 基于隐马尔科夫模型的 RCS 识别方法研究[J]. 现代雷达, 2013, 35(3): 37-40.

[11] 杨丽华, 戴齐, 郭艳军. KNN 文本分类算法研究[J]. 微计算机信息, 2006, 22(21): 269-270.