

# 概率潜在语义分析的 KNN 文本分类算法

戚后林, 顾磊

(南京邮电大学 计算机学院, 江苏 南京 210003)

**摘要:**传统的 KNN 文本算法在计算文本之间的相似度时,只是做简单的概念匹配,没有考虑到训练集与测试集文本中词项携带的语义信息,因此在利用 KNN 分类器进行文本分类过程中有可能导致语义丢失,分类结果不准确。针对这种情况,提出了一种基于概率潜在主题模型的 KNN 文本分类算法。该算法预先使用概率主题模型对训练集文本进行文本-主题、主题-词项建模,将文本携带的语义信息映射到主题上的低维空间,把文本相似度用文本-主题、主题-词项的概率分布表示,对低维文本的语义信息利用 KNN 算法进行文本分类。实验结果表明,在训练较大的训练数据集和待分类数据集上,所提算法能够利用 KNN 分类器进行文本的语义分类,且能提高 KNN 分类的准确率和召回率以及  $F_1$  值。

**关键词:**文本分类;KNN 算法;文本表示模型;语义分类;概率潜在主题模型

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2017)07-0057-05

doi:10.3969/j.issn.1673-629X.2017.07.013

## KNN Text Classification Algorithm with Probabilistic Latent Semantic Analysis

QI Hou-lin, GU Lei

(College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

**Abstract:** Traditional KNN Text Classification (TC) algorithm just implements a simple concept matching during calculation of the similarity between texts without taking the semantic information of the text in training and test set into account. Thus it is possible to lose semantic meaning in the process of text classification with KNN classifier as well as inaccurate categorization results. Against this problem, a KNN text classification algorithm based on probabilistic latent topic model has been proposed, which establishes probabilistic topic models of text-theme, theme-lexical item for training set texts beforehand to map the semantic information to low dimensional space of theme and dictates text similarity with probability distributions of text-theme and theme-lexical. The semantic information of low dimensional text can be classified with the proposed KNN algorithm. The experimental results show that in training of large training dataset and unclassified dataset, the proposed algorithm can conduct semantic classification of text with KNN classifier and enhance the accuracy and recall rate as well as  $F_1$  measure in KNN classification.

**Key words:** text classification; KNN; text presentation model; semantic classification; probability latent semantic analysis

## 1 概述

随着互联网的发展,数据呈现爆炸式增长,而作为信息的重要载体,文本已经成为互联网数据的重要组成部分,社区博客,微博短文,门户网站的新闻,无时无刻不在产生着大量文本,如何分类这些文本已经成为数据挖掘的一项重要内容。常用的文本分类算法有贝叶斯分类(Bayes)、K-最近邻算法(K-Nearest Neighbor, KNN)、支持向量机(SVM)以及人工神经网络(Artificial Neural Network)。其中, KNN 因其实现简单、分类性能较高在文本分类领域得到了广泛的研究

与应用。但是由于 KNN 算法的计算量较大,算法在计算文本相似度时不能有效处理一词多义或者同义词等问题,导致其准确率和召回率有时偏低。因此,对于 KNN 算法性能不足的优化成为研究的热点问题。

杜尔斌<sup>[1]</sup>等提出了一种改进的 KNN 文本分类算法。该算法采取统计量与信息增益这两种监督权重分配方法,弥补了原始 KNN 算法中对词项的非监督权重分配不足的缺点,提高了 KNN 分类的准确率,但需要在分类之余对数据集做额外的处理。骆凡等<sup>[2]</sup>提出了一种基于 Apriori 的 KNN 文本分类算法。该算法首

收稿日期:2016-07-18

修回日期:2016-10-26

网络出版时间:2017-06-05

基金项目:国家自然科学基金资助项目(61302157)

作者简介:戚后林(1990-),男,硕士研究生,研究方向为中文文本分类;顾磊,副教授,硕士生导师,研究方向为中文信息处理、机器学习。

网络出版地址: <http://jns.cnki.net/kcms/detail/61.1450.TP.20170605.1507.040.html>

先利用 Apriori 算法计算训练集样本的频繁项集,找出各频繁项集对于某个文章分类的置信度作为文本中词项的权重,因此在进行文本相似度计算时,待分类样本包含了词项之间的语义关联信息。樊存佳等<sup>[3]</sup>为了提高 KNN 算法的分类速度,提出了一种基于  $K$ -Medoids 的文本分类算法。该算法利用  $K$ -Medoids 聚类剔除对于 KNN 分类贡献较小的样本,从而减少了 KNN 算法在执行分类时的计算量,提高了算法运行速度。刘海峰等<sup>[4]</sup>利用  $K$ -means 聚类算法对初始训练集样本进行聚类降维,以此达到减少噪声样本对于测试样本类别判定干扰的目的;但该算法需对数据集进行预处理,在数据量较大时,算法的性能有所下降。同样为了优化 KNN 算法的执行速度,文献[5]提出了一种基于密度 KNN 分类的算法。该算法利用分治法将训练集中的每一类分成更小的类簇,然后估算这些更细小类簇对于整个类的影响权重,裁剪掉影响较小的噪音数据,然后再把剩下的类簇合并。由于剔除了噪音数据,因此数据量较处理之前变小,加快了文本分类的速度。Lin Yung-Shen<sup>[6]</sup>考虑了文本特征在整个训练集中出现、在单个文献中出现,以及在文本中未出现的三种不同情况,提出了一种新的文本相似度计算方法。Jing H 等<sup>[7]</sup>提出了一种语义朴素贝叶斯算法对文本进行分类,利用双线性文本模型把文本词项转换成矩阵来提取文本的语义信息,利用主成分分析法抽取每个文本的语义特征作为文本集合的语义空间。

相对于数字而言,文本是一种特殊的数据,已经发展了近千年,从口语到纸上记录的字符,再到计算机中存储的数据,文本中词项所带有的含义越来越复杂。例如,对于同样出现词项“苹果”的文本来说,一篇文本中指的是水果-“苹果”,而在另一篇中可能指的是计算机公司-“苹果”。同理,如果两篇文本中分别出现了“真相”,“实情”词项,在利用传统的分类算法计算这两个文本的相似度时,很有可能把这两个意思相近的词汇以完全不同的含义来处理。出现以上种种问题的原因在于:算法对于文本中的词汇只做了简单的概念匹配,丢失了语义信息,而这种一词多义或者同义词的现象在互联网产生的文本中很常见。语义信息是在文本产生时被作者赋予的,因此对文本生成过程进行建模对于语义信息的提取至关重要。近年来,由于机器学习等技术的发展,结合传统的文本分类算法与机器学习中的文本分类算法取得了很好的成效。文献[8]总结了机器学习在文本分类中的进展以及各种思想方法的局限性和优势。文献[9]提出了一种新的文本生成和表示模型,概率潜在主题模型(Probability Latent Semantic Analysis, PLSA),将文本中的高维词项映射到低维数据主题当中。该模型加入了文本的主题

信息,利用概率分布精确描述了文本-主题,主题-词项三层结构之间的关系,而且由于传统的空间向量模型(Vector Space Model, VSM)不能消除在文本分类过程中由于一词多义和多词一义带来的偏差,在实现降维的同时也实现了语义匹配。文献[10]提出了一种 LF-SVM(Latent Factor SVM)文本分类算法。该算法隐含的主题利用 PLSA 建模提取,然后利用提取出的隐藏变量作为训练集样本的分类代表,利用 SVM 进行文本分类。Chen Yewang 等<sup>[11]</sup>为了克服中文短文本中无自然分隔符的问题,提出了一种显式语义的文本分类算法。潜在的语义利用互联网上的 BaiduBaik 工具提取,然后利用 PLSA 提出的文档模型对提取出的语义进行主题建模,再利用 SVM 等算法对文本进行分类,提高了算法的执行速度与分类的准确率。Lu Youwei 等<sup>[12]</sup>提出了一种多标签的半监督文本分类算法(semi-supervised Latent Dirichlet allocation, ssLDA)。该算法在不需要指定训练集样本标签的情况下,把训练集样本和测试集样本训练出一个模型,然后进行分类。吕超镇等<sup>[13]</sup>提出了一种基于 LDA 特征扩展的文本分类算法。该算法将提取到的主题中的词扩展到原文本中,以此来丰富文本的语义信息,然后利用 SVM 进行文本分类。史庆伟等<sup>[14]</sup>克服了 LDA 在计算文本-主题、主题-词项概率分布过程中由于利用词袋模型所带来非作用词对结果的影响,利用 mRMR 算法预先过滤掉特征空间的非作用词,从而更精确地提取到文本的主题标签。

## 2 基于 VSM-KNN 的语义文本分类

### 2.1 传统的文本表示模型-空间向量模型

文本的 VSM 模型,由 Salton 等提出,并在文献检索系统成功应用。它可以把文本表示成向量,然后把对文本的处理转换成向量运算,在此基础上进行文本相似度计算直观简便,目前已经成为常用的文本表示模型。VSM 模型描述如下:

(1) 词典  $W = \{w_1, w_2, \dots, w_N\}$ ,  $N$  为词项的个数。

(2) 文本集合  $D = \{d_1, d_2, \dots, d_M\}$ ,  $M$  为文本的个数。

(3)  $d_i = \{w_{i1}, w_{i2}, \dots, w_{iN}\}$ ,  $w_{in}$  表示词项  $w_{in} \in W$  在文本  $d_i$  中的权重,常用 TF-IDF 值表示为:

$$w_{in} = \text{tf}_n \times \ln\left(\frac{M}{\text{df}_n}\right)$$

其中,  $\text{tf}_n$  表示词  $w_{in}$  在文本  $d_i$  中出现的频率;  $\text{df}_n$  表示整个文本集合中包含此词项的文本数目。

### 2.2 基于空间向量模型的 KNN 算法

KNN 算法的核心思想是找出与测试样本最相近的  $K$  个训练集中的样本,然后判断这  $K$  个样本中的大

多数属于哪一类,所谓的物以类聚,就是把此测试样本划分到此类中。基于空间向量模型 KNN 文本分类算法的步骤如下:

(1) 把文本  $d_i$  表示成以词项 TF-IDF 值的空间向量模型  $d_i = \{w_{i1}, w_{i2}, \dots, w_{iN}\}$ 。

(2) 待分类样本  $d_i = \{w_{i1}, w_{i2}, \dots, w_{iN}\}$  与训练集中任意样本  $t_j = \{w_{j1}, w_{j2}, \dots, w_{jN}\}$  的欧氏距离为:

$$\text{sim}(d_i, t_j) = \sqrt{\sum_{n=1}^N (w_{in} - w_{jn})^2} \quad (1)$$

除了欧氏距离外,另一种常用的方法是余弦距离:

$$\text{sim}(d_i, t_j) = \frac{w_{i1}w_{j1} + w_{i2}w_{j2} + \dots + w_{iN}w_{jN}}{\sqrt{\sum_{n=1}^N (w_{in})^2} \times \sqrt{\sum_{n=1}^N (w_{jn})^2}} \quad (2)$$

其中,  $N$  为词典中词项的个数。

(3) 待分类样本  $d_i$  与任意类别  $c_k$  的权重计算为:

$$p(d_i, c_k) = \sum_j^K \text{sim}(d_i, t_j) \tau(t_j, c_k) \quad (3)$$

其中,  $K$  为此类别包含的样本数目;  $\tau(t_j, c_k)$  是判断文本归属类别的函数,当  $t_j$  属于  $c_k$  时,  $\tau(t_j, c_k) = 1$ , 否则  $\tau(t_j, c_k) = 0$ 。

(4) 分类决策函数为:

$$f = \text{argmax}_k(p(d_i, c_k)) \quad (4)$$

即找到最大的  $p(d_i, c_k)$ , 把此文本划分到此类当中。

在用 KNN 算法进行文本分类时,对于文本相似度的计算一般采用余弦距离或欧氏距离,但是在空间向量模型的基础上对待测试文本计算训练集的相似度时,就像上文提到的那样,基于该模型的 KNN 算法只能匹配基本的深层次共同出现的概念,由于文本中词的多义性与同义性,算法对于深层次的语义信息匹配无能为力,而在互联网产生文本中,一词多义和同义词现象非常普遍。同时,算法在计算样本与训练集间的相似度时,需要同训练集中的每个文本进行计算,当训练集样本维度较高时,算法的性能会下降。

### 3 基于 PLSA-KNN 的语义文本分类

#### 3.1 PLSA 文本生成和表示模型

概率潜在语义分析模型假设文本集合中的每个文本是由多个满足一致概率分布的主题混合而成,而每个主题又是由多个词汇的概率分布混合组成。由于可以观察到文本集合,所以根据观察到的文本集合中的文本和每个文本中的词项来推导该文本的主题分布以及每个主题相应的词项分布。因为 PLSA 文本表示模型是把文本用文本-主题分布以及主题-词项分布表示,区别于传统的利用 TF-IDF 值把文本表示成空间向量模型,利用概率潜在语义分析模型进行文本分类,

不仅可以对文本进行主题降维,还可以提取到文本的语义信息。基本定义如下:

定义 1:  $p(t_i)$  表示文本集合中某篇文本被选中的概率。

定义 2:  $p(w_j | t_i)$  表示词  $w_j$  在指定文本  $t_i$  中出现的概率。

定义 3:  $p(z_k | t_i)$  表示某个主题  $z_k$  在文本  $t_i$  下出现的概率。

定义 4:  $p(w_j | z_k)$  表示某个词项  $w_j$  在主题  $z_k$  下出现的概率。

定义 5:  $p(t_i, w_j)$  表示词项  $w_j$  在文本  $t_i$  中的词频。

PLSA 模型假设文本集合中每个文本的词项生成过程如下:

(1) 按照概率分布  $p(t_i)$  选择一篇文本;

(2) 在选定文本  $t_i$  后,按照文本-主题分布  $p(z_k | t_i)$  选择一个主题;

(3) 在选定主题  $z_k$  后,根据主题-词项分布  $p(w_j | z_k)$  选择一个词;

(4) 重复以上几步,直到生成文本集合。

如图 1 所示,  $M$  表示文本个数,  $N$  表示文本中词的个数。由于每篇文本  $t$  和文本中的任意词项  $w$  都是可以观察到的,对于任意一篇文本  $p(w_j | t_i)$  是已知的。因此,  $p(w_j | t_i)$  可以利用期望最大化算法(Expectation Maximization, EM)<sup>[22]</sup>训练出主题  $z_k$ 。

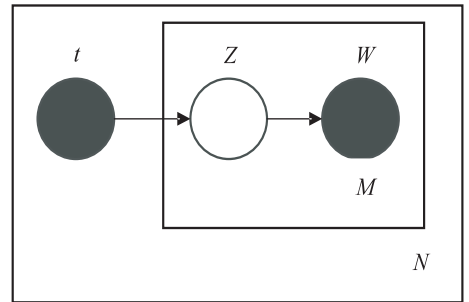


图 1 概率潜在主题模型

文本中每个词的生成概率为:

$$p(t_i, w_j) = p(t_i)p(w_j | t_i) = p(t_i) \sum_{k=1}^K p(w_j | z_k) p(z_k | t_i) \quad (5)$$

其中,  $p(w_j | t_i) = \sum_{k=1}^K p(w_j | z_k) p(z_k | t_i)$ 。

由于含有潜在的未知主题,故可以根据已观察到的信息,利用 EM 算法来最大化文本-主题分布  $p(z_k | t_i)$ 、主题-词项分布  $p(w_j | z_k)$ 。

(1) 文本-主题分布为:

$$p(w_j | z_k) = \frac{\sum_{i=1}^M n(t_i, w_j) p(z_k | t_i, w_j)}{\sum_{j=1}^N \sum_{i=1}^M n(t_i, w_j) p(z_k | t_i, w_j)} \quad (6)$$

其中,  $n(t_i, w_j)$  为文本  $t_i$  中词项  $w_j$  出现的频率。

(2) 主题-词项分布为:

$$p(z_k | t_i) = \frac{\sum_{j=1}^N n(t_i, w_j) p(z_k | t_i, w_j)}{n(t_i)}$$

(7)

其中,  $n(t_i)$  为文本  $t_i$  中词项的总数。

(3) 主题的后验概率为:

$$p(z_k | t_i, w_j) = \frac{p(z_k | t_i) p(w_j | z_k)}{\sum_{k=1}^K p(z_k | t_i) p(w_j | z_k)}$$

(8)

3.2 基于 PLSA 的 KNN 分类算法

传统的 KNN 算法在计算文本相似度时只是简单地匹配词项,未能融合文本之间的语义信息,而利用概率潜在主题模型把训练集样本加入潜在主题之后,可以利用文本-主题、主题-词项分布与测试样本进行语义层面的相似度计算。在找出文本训练集的潜在主题后,利用低维的文本-主题、主题-词项文本表示与测试集样本中的词项进行匹配,具体的算法步骤如下:

(1) 利用期望最大化算法找出训练集中的潜在主题  $z_k$ 。

(2) 把训练集中的文本集合映射成文本-主题  $p(z_k | t_i)$ 、主题-词项  $p(w_j | z_k)$  的概率潜在主题模型。

(3) 统计待分类文本集合中文本  $d_p$  的词项  $w_j$  出现的频率  $\text{tf}_j$ 。

(4) 计算  $d_p$  与测试集任意文本  $t_i$  的相似度(对于任意待分类文本,遍历训练样本,利用主题分布与此项分布以及词频乘积来计算相似度)。

$$\text{sim}(d_p, t_i) = \sum_{j=1}^N \sum_{k=1}^K \text{tf}_j \times p(w_j | z_k) p(z_k | t_i)$$

(9)

(5) 计算带分类样本  $d_p$  与每个类  $c_k$  的权重:

$$p(d_p, c_k) = \sum_i^K \text{sim}(d_p, t_i) \tau(t_i, c_k)$$

(10)

当  $t_i$  属于  $c_k$  时,  $\tau(t_i, c_k) = 1$ , 否则  $\tau(t_i, c_k) = 0$ 。

(6) 找出最大的  $p(d_p, c_k)$ , 将文本  $d_p$  划分到类  $K$  中。

4 实验及结果分析

4.1 性能评估指标

使用以下三个指标来检验算法的有效性:

(1) 查准率:某类中样本被正确归类的准确率。

$$\text{pre} = \frac{A}{B} \times 100\%$$

(11)

其中,  $A$  表示在某类中被正确分为此类的数目;  $B$  表示被错误分到此类的数目。

(2) 召回率。

$$\text{re} = \frac{A}{C} \times 100\%$$

(12)

其中,  $A$  同式(11);  $C$  表示此类中的文本被错误划分到其他类的数目。

(3)  $F_1$  值。

$$F_1 = \frac{2 \times \text{pre} \times \text{re}}{\text{pre} + \text{re}} \times 100\%$$

(13)

4.2 实验环境

实验选取了网易提供的新闻数据集,包含体育、经济、医药、文化、汽车五大类别。其中每个类别取 4 000 个文本,共计 20 000 个文本。每个文本中的词数大约 300 ~ 3 000 不等。从每个类中随机抽取 400 个不重复的文本构成训练集,共计 2 000 个文本。而测试集也是从每个类中抽取 400 个,既不自身重复也不与训练集的同类文本重复,共计 2 000 个。分词算法采用中科院技术研究所研发的汉语词法分词系统 ICTCLAS,实验在原生 Java 环境下进行。

首先利用概率潜在主题模型对训练集样本建模,把训练集表示成文本-主题、主题-词项结构,再对待测试样本集合进行分类,在进行 PLSA 建模时,选取了 5, 10, 15, ..., 50 个主题分别对训练集样本进行了 10 次测试,三个性能评估指标取 10 次测试的平均值。为了验证算法的有效性,与文献[11]中的算法进行对比。

4.3 实验结果

实验结果见表 1。

表 1 实验结果对比

类别	原始 KNN 算法			文献[11]算法			基于概率主题的 KNN 算法		
	准确率	召回率	$F_1$ 值	准确率	召回率	$F_1$ 值	准确率	召回率	$F_1$ 值
体育	0.67	0.712	0.69	0.85	0.822	0.86	0.923	0.942	0.932
经济	0.725	0.743	0.73	0.825	0.817	0.82	0.897	0.919	0.907
医药	0.683	0.697	0.69	0.792	0.776	0.783	0.91	0.893	0.901
文化	0.717	0.752	0.734	0.814	0.803	0.808	0.95	0.9	0.924
汽车	0.774	0.706	0.738	0.823	0.86	0.841	0.861	0.909	0.884

从实验结果可以看出,基于概率主题模型的 KNN 分类算法无论是查准率还是召回率都比另外两种算法要好,由于是每

本作为初始的训练集样本,故在性能评估上较有说服力。体育与医药是比较独特的两个类别,在选取的训练集和待测试样本集中,出现了诸如”球”、”胜”、”



负”、“健康”、“养生”等具有代表本类特性的词汇,在进行相似度计算时,这类词汇在一致的主题-词项分布中,所占的概率相比于其他词汇较大。因此,提出算法可以很好地辨别,准确率和召回率以及  $F_1$  值较高。同理,由于文化类的文本所包含的内容非常宽泛,在其内容中可以找到包含其他类别的词项,因此文化类的分类结果性能较低。

图2和图3分别为召回率-主题数目以及准确率-主题数目的变化趋势

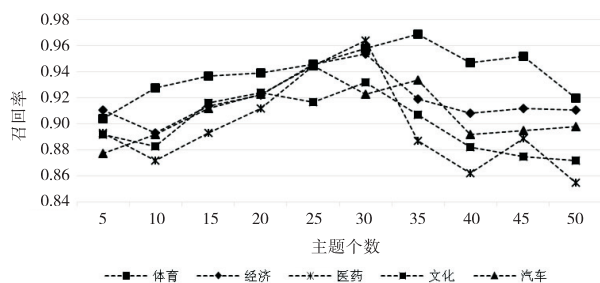


图2 召回率-主题数目变化趋势

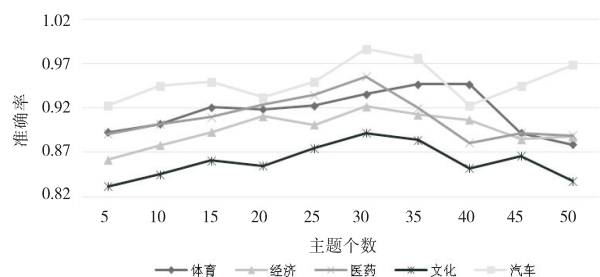


图3 准确率-主题数目变化趋势

从图2和图3中可以看到,算法在选取20~40个主题时,算法的准确率、召回率相比其他主题数目较高。因为实验所选取的数据集是在每个类别中选取400个文本作为训练集样本,每个类别提取的主题在4~8个可以反映出整个类别的主题,故在准确率以及召回率方面,算法对于待测试样本中的每个文本可以较好地分辨,因此准确率和召回率较高。

## 5 结束语

针对基于概率主题模型的 KNN 文本分类算法着重解决 KNN 文本分类过程中的语义丢失而导致文本分类性能较差的问题,从空间向量模型出发,对基于此模型的 KNN 文本分类算法的弊端进行分析,引入概率潜在主题模型,把训练集表示成文本-主题、主题-词项结构,从而在计算文本间的相似度时加入文本之间的语义信息。实验结果表明,采用新的概率潜在主题模型的算法对于分类结果的准确率、召回率、 $F_1$  值都有较大改进。同时,该算法也研究了在不同主题数目下基于概率潜在主题模型的文本分类算法性能。由于

万方数据

文本在表示成概率主题模型后,主题对于词项分布非常敏感,如何更精确地分词对于文本分类的结果至关重要;同时,在给定文本测试集个数的情况下,不同的主题个数对于文本分类的结果也有重要影响。因此,以后的工作中将利用更加精确的分本分词算法,优化主题个数,以期达到更好的分类效果。

## 参考文献:

- [1] 杜尔斌,李翔,林祥.改进的 KNN 文本分类算法[J].信息安全与通信保密,2011,9(4):38-39.
- [2] 骆凡,彭艳兵.一种基于 apriori 算法改进的 knn 文本分类方法[J].电子设计工程,2016,24(7):1-3.
- [3] 樊存佳,汪友生,边航.一种改进的 KNN 文本分类算法[J].国外电子测量技术,2015,34(12):39-43.
- [4] 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].微计算机信息,2010,26(1-3):18-20.
- [5] Jing Yongxia, Gou Heping, Zhu Yaling. An improved density-based method for reducing training data in KNN[C]//International conference on computational and information sciences. [s. l.]:[s. n.], 2013:972-975.
- [6] Lin Yung-Shen, Jiang Jung-Yi, Lee Shie-Jue. A similarity measure for text classification and clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(7): 1575-1590.
- [7] Jing H, Tsao Y, Chen Kuan-Yu, et al. Semantic Naïve Bayes classifier for document classification[C]//International joint conference on natural language processing. [s. l.]:[s. n.], 2013:1117-1123.
- [8] 苏金树,张博锋,徐昕.基于机器学习的文本分类技术研究进展[J].软件学报,2006,17(9):1848-1859.
- [9] Hofmann T. Probabilistic latent semantic analysis[C]//Proceedings of the fifteenth conference on uncertainty in artificial intelligence. [s. l.]:[s. n.], 1999:289-296.
- [10] Zhou Xiaofei, Guo Li, Liu Ping, et al. Latent factor SVM for text categorization[C]//IEEE international conference on data mining workshop. [s. l.]:IEEE, 2014:105-110.
- [11] Chen Yewang, Wang Jiongliang, Cai Yiqiao, et al. A method for Chinese text classification based on apparent semantics and latent aspects[J]. Journal of Ambient Intelligence and Humanized Computing, 2015, 6(4):473-480.
- [12] Lu Youwei, Okada S, Nitta K. Semi-supervised latent Dirichlet allocation for multi-label text classification[C]//26th international conference on industrial, engineering and other applications of applied intelligent systems intelligence. [s. l.]:[s. n.], 2013:351-360.
- [13] 吕超镇,姬东鸿,吴飞飞.基于 LDA 特征扩展的短文本分类[J].计算机工程与应用,2015,51(4):123-127.
- [14] 史庆伟,从世源.基于 mRMR 和 LDA 主题模型的文本分类研究[J].计算机工程与应用,2016,52(5):127-133.