

# 农业信息分类中 $K$ -means 与 SVM 的混合算法研究

赵新苗,冯向萍,赵 涛

(新疆农业大学 计算机与信息工程学院,新疆 乌鲁木齐 830052)

**摘 要:**随着新疆农业信息技术的不断发展和农村互联网的广泛普及,互联网中海量的农业相关知识和信息虽然为工作人员带来了便利,但是与此时也给信息检索增加了难度。在对具有新疆特色的农作物网页信息分类研究的基础上,提出并实现了  $K$ -means 与 SVM 相结合的分类方法,以帮助农业相关工作人员获得更准确有效的信息。该分类方法采用  $K$ -means 对训练样本进行聚类以减少边缘训练样本,并应用 SVM 对删减后的训练集进行训练。为减少训练集边缘样本、节省训练时间,还提出了两种基于中心向量的边缘样本删减方法,分别仅保留中心向量方法和保留中心向量临近样本。实验验证结果表明,所提出的方法均能够同时有效地减少训练样本和训练时间。

**关键词:**农业信息;分类;聚类;边缘样本删减

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2017)06-0178-05

doi:10.3969/j.issn.1673-629X.2017.06.037

## Investigation on $K$ -means and SVM Mixed Algorithm for Agriculture Information Classification

ZHAO Xin-miao, FENG Xiang-ping, ZHAO Tao

(College of Computer and Information Engineering, Xinjiang Agricultural University, Urumqi 830052, China)

**Abstract:** With the continuous development of Xinjiang agricultural information technology and the widespread popularity of rural Internet, the amount of relevant knowledge and information in Internet has been bringing lots of conveniences for people and difficulty for effective information retrieval at the same time. Based on the requirement analysis of Xinjiang Rural Information Acquisition System and aiming at categorization of the web pages which are about characteristic crops in Xinjiang to help display more accurate and effective agricultural information and reduce the number of training sets and save training time, a method combined with SVM and  $K$ -means has been proposed. Its main process contains clustering the training sets with  $K$ -means to delete edge samples and training the SVM on the new deleted training sets. Two methods of deleting edge samples and retaining neighbors of the centers have also been proposed. Experimental results show that these methods can decrease training samples and training time.

**Key words:** agricultural information; classification; clustering; edge samples reduction

## 0 引言

中共中央国务院在《关于积极发展现代农业扎实推进社会主义新农村建设的若干意见》中明确提出:“推动农业信息数据收集整理规范化、标准化”<sup>[1]</sup>。根据“十二五”规划中关于农业方面提出的《全国农业和农村经济发展第十二个五年规划》和《农业科技发展规划(2006-2020年)》<sup>[2]</sup>可以看出,在新时代到来之际,农业要走信息化、科技化的道路<sup>[3]</sup>。

目前有很多学者对农业搜索引擎进行研究。例如,周鹏等就目前搜索引擎在专业特色领域中应用度

低的问题,使用开源的搜索引擎架构 Nutch 搭建了农业信息相关的搜索引擎<sup>[4]</sup>。熊金辉等根据目前农业信息化发展的现状,指出了建立搜索引擎是必要的<sup>[5]</sup>。王晓琴等则针对传统搜索引擎专业性差和查准率低等问题,实现了基于 Nutch 的农业垂直搜索引擎<sup>[6]</sup>。

搜索引擎抓取回来的数据很庞大且杂乱无章,因此对搜索引擎抓取的数据进行有效的管理和分类势在必行。关于文本分类的文献较多,例如, Apte 用决策树技术来获取分类器<sup>[7]</sup>; Yang 等提出了一种邻近算法进行分类<sup>[8]</sup>; Lewis 等采用了一个线性分类器<sup>[9]</sup>; Cohen

收稿日期:2016-04-13

修回日期:2016-07-28

网络出版时间:2017-04-28

基金项目:新疆维吾尔自治区科技计划项目(2015X0108-1)

作者简介:赵新苗(1990-),女,硕士研究生,研究方向为数据库技术;冯向萍,副教授,通讯作者,研究方向为数据库技术及应用。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20170428.1702.010.html>

等设计了一种建立在权值更新基础上的休眠专家算法<sup>[10]</sup>。

新疆农作物在种类、种植方法、农业政策和农业科技上与其他地区都存在差异,但是却没有提供新疆农作物相关的独特的搜索引擎,所以建立新疆特有的农作物信息检索平台显得尤为重要。为此,根据《新疆农村信息采集系统》功能需求,主要通过分析机器学习分类算法,设计农业信息分类模型,以达到对搜索引擎采集的数据进行正文内容抽取和分类的目的。

## 1 相关算法

### 1.1 支持向量机

支持向量机(Support Vector Machine, SVM)方法的提出在很多领域应用广泛。它是一种基于统计学习理论的新型学习方法。最大特点就是根据 Vapnik 结构使风险达到了最小,提高了学习机的泛化能力<sup>[11]</sup>。通过以上几个特点可以看出, SVM 优于其他一些算法,通过大量的研究表明支持向量机分类器是最优越的分类器之一<sup>[12-13]</sup>。但是对于大规模的网页分类数据而言,训练样本庞大, SVM 需要训练的时间很长。

(1) 支持向量机分类的过程。

SVM 通过非线性变换,将输入量映射到一个高维空间  $H$  上,在  $H$  中构造最优的分类超平面,进而得到良好的泛化能力<sup>[14]</sup>。其详细的算法步骤如下:

(1) 求解约束条件下的二次最优问题,得到使分类间距  $\text{Margin} = 2/\|w\|$  最大的  $w = \sum_{i=1}^n a_i^* y_i d_i$  和分类阈值  $b$ 。

其中,  $d_i$  是由  $a_i$  不为 0 而确定的支持向量;  $b = 0.5(wd_1 + wd_2)$  是分类阈值,  $d_1$  和  $d_2$  分别是两个类中任意一个  $a_i > 0$  所对应的样本向量。

(2) 将待分类向量  $d$  和支持向量  $d_i$  用核函数  $K(d, d_i)$  映射到线性空间。常用的核函数有:

① 多项式核:  $[(d, d_i) + 1]^q, q \in \text{自然数}$ 。

② 径向基核(RBF):  $\exp\left\{-\frac{\|d - d_i\|}{\sigma^2}\right\}$ 。

③ 两层神经网络核:  $S(\alpha(d, d_i) + t)$ 。其中,  $S$  是 sigmoid 函数,  $\alpha$  和  $t$  是常数。

(3) 用核函数  $K(d, d_i) = \Phi(d)(d_i)$  来代替点积  $(d, d_i)$ , 然后用  $a_i^* y_i$  加权, 从而构成  $\sum a_i^* y_i K(d, d_i)$ 。

(4) 判断  $f(d) = \text{sgn}(\sum_{i=1}^n a_i^* y_i K(d, d_i) + b)$  是否大于等于+1 或小于等于-1, 决定文档  $d$  属于哪一类。其中,  $\text{sgn}$  为函数,  $a_i$  为各个样本所对应的拉格朗日乘子。

万方数据

### 1.2 $K$ -means 聚类算法

$K$ -means 算法是聚类分析中一种经典的基于中心向量的聚类方法。该算法以其原理简单、收敛速度快以及适应性强而得到广泛应用。

$K$ -means 算法核心思想是将  $n$  个数据对象划分成  $k$  个聚类, 生成的每个聚类满足: 同一聚类中的对象相似度较高, 而不同聚类中的对象相似度较小, 即类内紧凑, 类间独立。

## 2 $K$ -means 和 SVM 结合分类算法

SVM 的时间复杂度线性时为  $O(nd)$ , 非线性时为  $O(nd^2)$ , 其中  $d$  为训练样本数,  $d$  为特征维度。所以随着样本数量和特征数量的增加, SVM 的时间复杂度和空间复杂度也会增加。

$K$ -means 算法经常以局部最优结束, 适合处理大数据集, 特别是当数据呈现球形分布时效果较好, 但由于数据的分布往往是分散和不规则的, 因此该方法聚类速度快, 但准确率低。

根据  $K$ -means 算法聚类速度快和 SVM 算法准确率高的特点, 将两者结合起来, 既能提高训练速度, 又能保证分类准确率。针对这一问题, 提出了一种  $K$ -means 和 SVM 相结合的分类模型。

### 2.1 $K$ -means 与 SVM 结合的分类模型

具体流程如下: 首先使用  $K$ -means 对训练样本进行聚类, 然后删减掉边缘训练样本, 使用这种方法对数据集进行删减, 以达到准确率基本不变并且提高分类模型速度的效果; 然后使用 SVM 对删减过的训练样本进行训练, 并对测试样本进行分类测试。

### 2.2 $K$ -means 方法删减边缘样本

#### 2.2.1 删减边缘样本的原理

对于训练集中的某个样本  $d$  来说, 要么是第  $i$  类文本, 要么是第  $i$  类和其他类交叉区域的文本, 交叉区域的文本大多是多类别属性的文本, 还有不属于任何类的文本, 这些边缘化的样本在分类过程中不仅增加了训练过程的时间开销和计算开销, 还会影响分类结果。因此对训练文本进行删减, 删除一些边缘样本以达到准确率基本不变并减少训练时间和计算量的目的。

根据上述思想对所采用的  $K$ -means 方法删减边缘样本, 提出了两种基于中心向量的解决方法, 即仅保留中心向量或者保留中心向量邻近的文本作为训练样本, 并对其进行实验验证。

#### 2.2.2 保留中心向量

该方法将训练集向量表示后用  $K$ -means 方法对每类的训练集聚类, 聚类簇数为  $n \leq k \leq m$ , 即聚类簇数大于 1 并且小于该类样本数。因为每个簇的中心向

量都是具有代表性的有用数据,因此取每个簇的中心向量作为新的训练集,以达到减少训练样本数的目的。

假设训练集为  $S$ , 类别数为  $n$ , 类别为  $c = (c_1, c_2, \dots, c_n)$ , 每类训练集数为  $m$ ,  $\bar{x}_i$  代表类  $c_i$  的中心向量。

(1) 读入训练集。假设每类的聚类簇数为  $k$ , 随机取出  $k$  个文本向量作为初始向量, 其中  $1 \leq k \leq m$ 。

(2) 将新到样本归纳到距离最近的类中。

(3) 重新计算簇的中心向量  $\bar{x}_i$ :

$$\bar{x}_i = \frac{1}{|c_i|} \sum_{x_j \in c_i} x_j \quad (1)$$

(4) 重复上述步骤, 直到收敛, 取每类中每个簇的中心向量作为 SVM 的新训练集  $S_1$  进行训练。

(5) 对生成的分类器进行分类测试, 如图 1 所示。

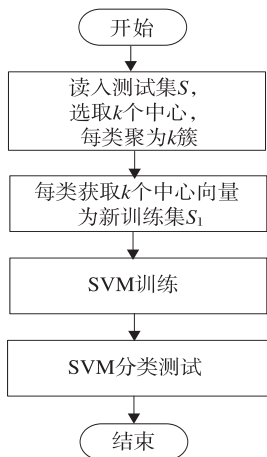


图 1 保留中心向量的混合模型流程图

### 2.2.3 保留中心向量临近文本

将训练集向量表示后, 使用  $K$ -means 方法将每类文本聚为  $k$  簇, 其中  $n \leq k \leq m$ , 即聚类簇数大于 1 并且小于该类样本个数。假设训练集为  $S$ , 类别数为  $d$ , 类别为  $c = (c_1, c_2, \dots, c_n)$ , 每类训练集数为  $m$ ,  $\bar{x}_i$  代表类  $c_i$  的中心向量。 $d_j$  为第  $j$  簇中某个样本到中心向量的距离, 取所有样本到簇中心的距离的平均值为半径  $d_r$ , 最后每一簇的训练集就是分布在以  $\bar{x}_i$  为中心、 $d_r$  为半径的圆内。每一类的训练集即为每簇样本个数之和。

(1) 读入训练集。假设每类的聚类簇数为  $k$ , 随机取出  $k$  个文本向量作为初始向量, 其中  $1 \leq k \leq m$ 。

(2) 将新到文本归纳到距离近的类中。

(3) 重新计算簇的中心向量  $\bar{x}_i$  (同式(1))。

(4) 重复上述步骤, 直到收敛, 取出每簇中以  $\bar{x}_i$  为中心、 $d_r$  为半径的样本为该类的样本, 汇总每一类的样本作为新的测试集  $S_1$ , 并使用 SVM 进行训练;

$$d_r = \frac{1}{|c_i|} \sum d(\bar{x}_i, d_j) \quad (2)$$

$$d(\bar{x}_i, d_j) = \sqrt{\sum_{d_j \in c_i} (\bar{x}_i - d_j)^2} \quad (3)$$

(5) 对生成的分类器进行分类测试, 如图 2 所示。

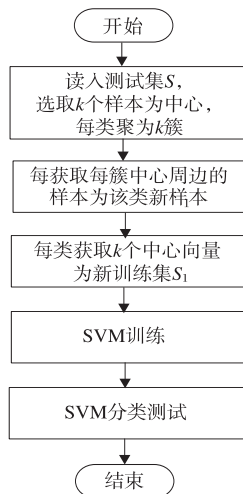


图 2 保留中心向量周边样本流程图

### 2.3 分类结果评价标准

国际上通用的评价指标有: 查准率 (Precision)、查全率 (Recall) 和  $F_1$  测度<sup>[15]</sup>。

$$\text{Precision} = \frac{A}{A + D} \quad (4)$$

$$\text{Recall} = \frac{A}{A + B} \quad (5)$$

其中,  $A$  表示样本集中原本是正例, 被模型判断为正例的样本数;  $B$  表示样本集中原本是正例, 却被模型判断为反例的样本数;  $D$  表示样本集中原本是反例, 却被模型判断为正例的样本数。

$F_1$  测度是对查准率和查全率两个指标进行加权和平均后形成的一个综合指标。

$$F_1 = \frac{\text{Precision} \times \text{Recall} \times 2}{\text{Precision} + \text{Recall}} \quad (6)$$

## 3 实验结果与分析

通过八爪鱼采集器从中国农业网 (www. agronet. com. cn)、中国兴农网 (www. xn121. com)、中国玉米网 (www. chnym. com)、中国棉花网 (www. cncotton. com)、葡萄网 (http://grape. forestry. gov. cn)、中国小麦网 (www. xiaomai. cn)、红枣网 (www. zao. com. cn) 和核桃网 (www. cnhetao. com) 上抓取了 21 450 条相关网页数据。

此处将对保留中心向量和保留中心向量周围样本的边缘样本删减方法进行对比实验, 若样本太多训练时间会太长。为了便于实验, 将从每类样本中抽取 500 篇作为训练集, 即总共有 4 000 篇文章作为实验样本。

### 3.1 保留中心向量实验

实验中对每类训练样本数为 500 的训练集进行了聚类 (即共 4 000 个训练样本), 具体如下:

- (1) 每类分别聚为 100 簇,并获得每类 100 个中心向量作为该类新的测试集,共 800 个新测试样本。
- (2) 每类分别聚为 200 簇,并获得每类 200 个中心向量作为该类新的测试集,共 1 600 个新测试样本。
- (3) 每类分别聚为 300 簇,并获得每类 300 个中心向量作为该类新的测试集,共 2 400 个新测试样本。
- (4) 每类分别聚为 400 簇,并获得每类 400 个中心向量作为该类新的测试集,共 3 200 个新测试样本。
- (5) 训练集分别保存在 train(原训练集,仅使用 SVM),train\_100,train\_200,train\_300,train\_400 中,然后对这几个训练集进行训练和分类测试。

表 1 为仅使用 SVM 和改进后的 K-means&SVM 混合模型在训练文本个数、 $F_1$  测度以及训练时间上的对比。

表 1 改进前和改进后实验结果对比

样本聚类	样本数	$F_1$ 测度	$t$ /ms
原样本仅使用 SVM	4 000	84.110 1	86 049
每类聚类成 100 个中心	800	59.294 1	16 230
每类聚类成 200 个中心	1 600	66.384	26 948
每类聚类成 300 个中心	2 400	76.933 8	35 585
每类聚类成 400 个中心	3 200	80.649 1	46 110

3.2 保留中心向量结果分析

如图 3 所示,改进前每类训练集样本个数为 500 时,仅使用 SVM 分类, $F_1$  测度大概为 84% 左右,训练时间为 86 049 ms;当把每类训练集聚类成 100 簇时, $F_1$  测度大约为 60% 左右,训练时间下降为 16 230 ms, $F_1$  下降明显。原因是因为训练集样本删减个数太多导致  $F_1$  测度直线下降,并且时间的下降速度快于  $F_1$  测度的下降速度。

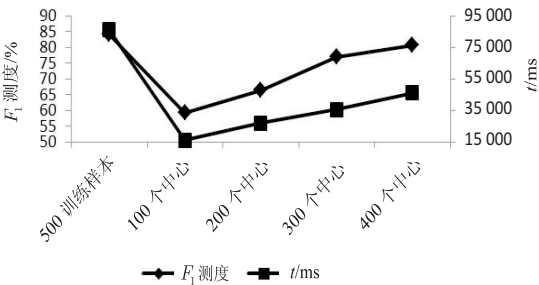


图 3  $F_1$  测度和训练时间随样本数增加的变化图

当每类训练集聚类成 200~400 簇时,随着样本数量的上升, $F_1$  测度有所提升,但是训练时间也会出现翻倍增长的趋势。当训练集聚为 400 簇时, $F_1$  测度达到了 80% 左右,但时间的增长速度明显快过  $F_1$  测度的增长速度。

如图 4 所示,训练时间随着取得中心向量个数变化的时间差值如下所述:从每类 100 变为 200 时,时间差为 10 718 ms,从每类 200 变为 300 时,时间差为

8 637 ms;从每类 300 变为 400 时,时间差为 10 525 ms;从每类 400 变为原 500 个训练样本时,时间差为 39 939 ms。

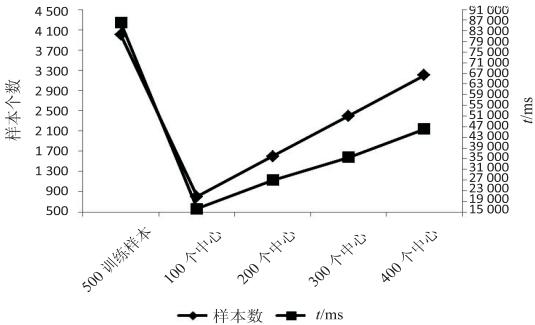


图 4 随着训练样本增加的训练时间变化图

从上述仅使用 SVM 和改进后的时间差可以发现,保留中心向量的边缘样本删减法的确可以减少训练时间,训练样本数和训练时间是成正相关的,而且随着样本数的增加,时间增加的幅度更大。

3.3 保留中心向量临近样本结果

实验中对每类训练样本数为 500 的训练集进行聚类(即共 4 000 个训练样本),具体如下:

每类分别聚为 100 簇,并获得每类以这 100 簇的中心向量为圆心,以所有样本到该簇中心的平均距离为半径的圆,作为该类新的测试集,据统计共 2 652 个新测试样本。

表 2 为仅使用 SVM、每类仅保留 300 个中心向量的 K-means&SVM 混合模型、每类取 100 中心向量临近样本的 K-means&SVM 混合模型在训练文本个数、 $F_1$  测度以及训练时间上的对比。

表 2 实验结果对比

每类样本	样本数	$F_1$ 测度	$t$ /ms
原样本	4 000	84.11	86 049
300 个中心	2 400	76.93	35 585
100 个中心临近样本	2 652	83.11	56 727

3.4 保留中心向量临近样本分析

每类取 100 个簇的中心向量周围的文本为训练集时,训练集为 2 652 个, $F_1$  测度为 83%,训练时间为 56 727 ms。

根据表 2 的实验结果可得:

- (1) 每类取 100 中心向量临近样本的 K-means&SVM 混合模型的训练样本数比仅使用 SVM 的训练样本数少 1 348 个,训练时间比 train 少 29 322 ms,但是  $F_1$  测度仅比 train 少大约 1%。
- (2) 每类取 100 中心向量临近样本的 K-means&SVM 混合模型的训练样本数比每类仅保留 300 个中心向量的 K-means&SVM 混合模型的训练样本数多 252 个,但是训练时间比每类仅保留 300 个中

心向量的  $K - \text{means}$  & SVM 混合模型多 21 142 ms,  $F_1$  测度比 train\_300 高大约 3%。

总之,对训练集中边缘样本进行删减是以牺牲部分文本信息为代价的,过多的文本剪裁虽然可以节约计算开销,但是必然会引起准确率的下降。上述方法的目的是删除与中心向量较远的那些文本,因此在计算时间和准确率之间会有适当的折中取舍,这种根据簇分布删减边缘样本的方法,只要删除数量适当,对分类结果产生的影响较小。

保留中心向量和保留中心向量邻近样本的方法各有优缺点:前者训练时间较快,但是  $F_1$  测度较低;后者  $F_1$  测度较高,但是训练时间相对较慢。但是两种方法均可以起到准确率基本不变、训练时间缩短的效果。

## 4 结束语

新疆农业信息技术不断加速发展,互联网中海量的农业相关信息虽然为工作人员带来了便利,但与此同时也给信息检索增加了难度。依据《新疆农村信息采集系统》的需求,针对具有新疆特色的农作物网页信息进行分类研究,帮助农业相关工作人员获得更准确有效的信息。为此,提出并实现了  $K - \text{means}$  和 SVM 的农业信息网页分类模型。由实验结果及其分析可得:与仅采用 SVM 相比,保留中心向量和保留中心向量临近文本的方法均可达到准确率基本不变且提高分类的速度的目的;训练集的删减程度需要加以控制,否则会直接影响到分类器的性能;采用的方法满足大数据运行的要求。

## 参考文献:

- [1] 胡金有,张健,游龙勇.我国农业信息网站现状分析[J].农机化研究,2005(6):38-40.
- [2] 黄建全,解翠平,黎凌.新疆农业信息化发展现状与建议[J].新疆农业科技,2013(5):1-4.

(上接第 177 页)

- models for text and citations[C]//Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining. [s. l.]:ACM,2008:542-550.
- [10] 李明,王占宏,鲁明.基于 J2EE 框架的混合模式治安管理信息系统研究与应用[J].计算机工程,2003,29(1):252-252.
- [11] Garrett J J. Ajax: a new approach to web applications[EB/OL]. (2005-02-18)[2011-02-18]. <http://www.adaptive-path.com/publications/essays/archives/000385.php>.
- [12] 郭剑飞.基于 LDA 多模型中文短文本主题分类体系构建与分类[D].哈尔滨:哈尔滨工业大学,2014.
- [13] Thomas H. Probabilistic latent semantic indexing[C]//Pro-

- [3] 董婷婷,方萍.浅议计算机在农业中的应用及前景[J].农业网络信息,2009(6):118-119.
- [4] 周鹏,吴华瑞,赵春江,等.基于 Nutch 农业搜索引擎的研究与设计[J].计算机工程与设计,2009,30(3):610-612.
- [5] 熊金辉,张海雷,余波,等.中文农业信息资源整合平台的设计与实现[J].中国农学通报,2005,21(12):407-410.
- [6] 王晓琴,李书琴,景旭,等.基于 Nutch 的农业垂直搜索引擎研究[J].计算机工程与设计,2014,35(6):2239-2243.
- [7] Velasco E, Thuler L C, Martins C A, et al. Automated learning of decision rules for text categorization[J]. ACM Transactions on Information Systems, 1994, 12(3): 233-251.
- [8] Yang Y. Expert network: effective and efficient learning from human decisions in text categorization and retrieval[C]//International ACM SIGIR conference on research and development in information retrieval. Dublin, Ireland; ACM, 1994: 13-22.
- [9] Lewis D D, Schapire R E, Callan J P, et al. Training algorithms for linear text classifiers[C]//International ACM SIGIR conference on research and development in information retrieval. [s. l.]: ACM, 1999: 298-306.
- [10] Cohen W W, Singer Y. Context-sensitive learning methods for text categorization[J]. ACM Transactions on Information Systems, 2002, 17(2): 307-315.
- [11] Andrew A M. An introduction to support vector machines and other kernel-based learning methods[J]. AI Magazine, 2000, 32(8): 1-28.
- [12] 平源.基于支持向量机的聚类及文本分类研究[D].北京:北京邮电大学,2012.
- [13] 罗瑜.支持向量机在机器学习中的应用研究[D].成都:西南交通大学,2007.
- [14] 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.
- [15] 宋枫溪,高林.文本分类器性能评估指标[J].计算机工程,2004,30(13):107-109.

ceedings of SIGIR. Berkeley, CA, USA: [s. n.], 1999: 50-57.

- [14] Griffiths T, Steyvers M. Probabilistic topic models[M]//Latent semantic analysis: a road to meaning. Hillsdale, NJ: Laurence Erlbaum, 2006.
- [15] Philp R, Eric H. Gibbs sampling for the uninitiated[R]. [s. l.]: [s. n.], 2010.
- [16] 张晨逸,孙建伶,丁轶群.基于 MB-LDA 模型的微博主题挖掘[J].计算机研究与发展,2011,48(10):1795-1802.
- [17] 胡吉明,陈果.基于动态 LDA 主题模型的内容主题挖掘与演化[J].图书情报工作,2014,58(2):138-142.
- [18] Ma D, Rao Lan, Wang Ting. An empirical study of SLDA for information retrieval[J]. Information Retrieval Technology, 2011(1): 84-92.