一种基于情感词典和朴素贝叶斯的中文文本情感分类方法 *

杨 鼎[1,2], 阳爱民[1,3]

(1. 湖南工业大学 计算机与通信学院, 湖南 株洲 412008; 2. 湖南省教育考试院 信息处, 长沙 410001; 3. 广东外语外贸大学 信息科学与技术学院, 广州 510006)

摘 要: 基于朴素贝叶斯理论提出了一种新的中文文本情感分类方法。这种方法利用情感词典对文本进行处理和表示,基于朴素贝叶斯理论构建文本情感分类器,并以互联网上宾馆中文评论作为分类研究的对象。实验表明,使用提出的方法构成的分类器具有分类速度快、分类准确度高、鲁棒性强等特点,并且适合于大量中文文本情感分类应用系统。

关键词: 文本情感分类; 朴素贝叶斯; 情感词典

中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2010)10-3737-03

doi:10.3969/j.issn.1001-3695.2010.10.035

Classification approach of Chinese texts sentiment based on semantic lexicon and naive Bayesian

YANG Ding[1,2], YANG Ai-min[1,3]

(1. *Institute of Computer & Communication, Hunan University of Technology, Zhuzhou Hunan* 412008, *China*; 2. *Dept. of Information, Hunan Provincial Education Examination Board, Changsha* 410001, *China*; 3. *School of Informatics, Guangdong University of Foreign Studies, Guangzhou* 510006, *China*)

Abstract: This paper provided a new classification approach of Chinese texts based on naive Bayesian. The approach reached its goal by applying semantic lexicon on text processing and expressing, constructing sentiment classifier based on naive Bayesian and experimental data obtained from hotel's Chinese reviews through Internet service. Backed with the experimental data, this approach demonstrates its efficiency, accuracy and robustness, which makes it applicable as well in sentiment classification for plenty of Chinese texts.

Key words: text sentiment classification; naive Bayesian; semantic lexicon

人们对事物的情感都是有两面性的,如正面与反面、褒义与贬义等。一般认为,文本的情感倾向分类是一个两分类问题,就是把文本分成正面和负面两类。中文文本情感分类研究是国内近年来的研究热门,大多数研究是利用机器学习方法构建分类器,采用不同的特征权值公式、特征选择方法对采集来的语料进行分类实验,并比较不同方法的优缺点和分类性能。文献[1]比较了不同的特征选择和使用多种机器学习构建分类器;文献[2]利用多种特征选择方法和权重计算方法、五种停用词表以及用 SVM 构建分类器对汽车语料的文本情感类别进行了研究。

互联网作为一个巨大的语料库,是中文情感分类应用的实验场,但面对海量的中文文本信息,如何能快速有效地进行情感分类,挖掘互联网信息,增强用户体验等,值得深入研究。本文介绍了使用情感词典作为分类特征,利用朴素贝叶斯方法构建分类器,对大量中文宾馆评论进行情感分类研究。实验结果表明,由这种方法构建的分类器具有分类速度快、分类准确度高、鲁棒性强等特点,并且可以快速地进行大量的中文文本情感分类。本文提出的中文情感分类器的构建过程主要包括中文文本处理及表示、特征选择、分类器训练和分类器评测等。

1 中文情感文本处理及表示

情感文本处理和表示是分类器构建的一个重要过程,包括对文本进行的中文分词、特征选择、特征权值计算及文本向量表示等工作。无论是训练语料还是测试语料,都需要先进行文本处理才能输入分类器进行分类。

本文提出使用情感词典对文本进行表示,这个过程可以在中文分词阶段就能完成,不需要单独的特征选择步骤,处理流程如图 1 所示。

1.1 中文分词

中文不像英文那样每个词汇之间有空格分开,需先进行分词才能进一步处理。本文采用了最大匹配算法对中文文本进行分词,该方法属于基于字符串匹配的分词方法,需要分词词典支持。分词词典本文采用了国家语言文字工作委员会发布

收稿日期: 2010-03-11; **修回日期**: 2010-04-21 **基金项目**: 湖南省教育厅科学研究资助项目(07B014);广东省自然科学基金资助项目(9151805707000010);广州市社科规划项目(08Y59)

作者简介:杨鼎(1982-),男,河南禹州人,硕士,主要研究方向为文本情感分类、数据挖掘(dean@hut.edu.cn);阳爱民(1970-),男,湖南永州人,教授,博士,主要研究方向为模式分类、智能计算.

的《现代汉语常用词表(草案)》(LCWCC)[3],该词典搜集了现在日常生活中使用频率较高的 56 008 个词汇,基本能够满足分词的需要。

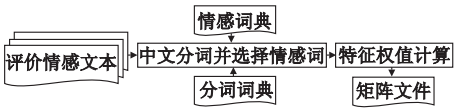

图1　文本处理的流程

在特征选择方法上,本文采用了情感词典作为特征选择的依据,所以在分词时,实际是使用了 LCWCC 和情感词典的并集作为了分词词典。其中最大匹配的步长设置为 4 个汉字,只对中文内容进行分词处理。

1.2　情感词典

在人工对文本的情感倾向进行标注时,文本中往往少数带有情感倾向的词汇起决定性的作用,如褒义词和贬义词。中文情感词汇非常复杂,情感词汇的词性很多,主要为形容词、名词、副词等,但仅仅考虑词性选择情感词并不科学,如“垃圾”“棒槌”都带有负面情感,但大量的名词并不带情感色彩,如果选用势必会降低分类的性能。

情感词是最好的表示文本情感的特征。文献[4]基于多种词典资源构建情感词表,使用加权线性组合方法对句子情感进行分类,实验表明情感词表有助于提升情感分类的效果。本文是在朴素贝叶斯分类器中应用情感词典进行情感特征选择,这种方法构建的分类器可以快速稳定地工作,并且分类结果也较传统的特征选择方法好。

本文使用了文献[5]中的基础情感词(BSL)作为本文选用的情感词典,该词典拥有 5 281 个情感词汇。其中褒义词2 807 个,贬义词 2 474 个。这个词典是在 HowNet 情感词语集的基础上,利用其提供的义原计算两个词的相似度,根据词与正向和负向种子词的平均相似度的差,来判定词的情感倾向得到的一个情感词典。

1.3　特征选择

特征选择是选择文本中能显著表示文本类别的词汇,去掉无用的词汇。因为分词后得到的是一个稀疏矩阵,也是为了便于计算,一般都要降低矩阵的维度。如果使用情感词典作为特征选择的依据,而情感词典词汇数量不是很多,就不需要进一步降维,所以也可以视做使用情感词典是一种降维的方法。在本文中使用了 CHI 统计方法选择特征和使用情感词典作为特征进行了实验比较,使用情感词典比使用 CHI 统计方法作为特征选择进行情感分类效果有所提升。

CHI 统计方法是一种应用比较广泛而且性能不错的特征选择方法,使用 chi-max 作为特征选择在文本分类方面表现比较好[6]。采用 CHI 统计作为特征选择方法,每个特征词 w 对于类别 c 的 x^2 值计算方法如式(1)。

$$x^2(w,c)=\frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{1}$$

其中:A 是特征词 w 和类别 c 共同出现的次数,B 是 w 出现但 c 不出现的次数,C 是 c 出现但 w 不出现的次数,D 是 w 和 c 都不出现的次数,N 是文本总数。采用式(2)作为特征选择的依据,即选用 chi-max 值最大的若干个词作为特征词。

$$\text{chi-max}: x^2_{\max}(w)=\max_{c_i\in C}\{x^2(w,c_i)\} \tag{2}$$

1.4　特征权值计算

特征权值是指特征词在文本中的权重,也可称为词的向量,是分类器分类的重要依据。本文使用词频、布尔型(Bool)两种权值进行情感分类对比。一般在分词处理完成后就可以计算特征权值,然后特征选择后输入分类器。然而使用情感词典作为特征选择,因为分词中就可以完成特征选择,所以特征权值计算在特征选择之后进行。

词频和 Bool 型权值是比较简单的特征权重表示方法,适合用于朴素贝叶斯分类器,Bool 型权值计算方法如式(3)。

$$\text{Bool}:\begin{cases}1 & \text{freq}(w_i,d_j)>0\\ 0 & \text{freq}(w_i,d_j)=0\end{cases} \tag{3}$$

其中:$\text{freq}(w_i,d_j)$ 是词 w_i 出现在文本 d_j 中的频次,即为词频。

1.5　情感文本集向量矩阵

文本 d 可以表示成若干词的集合 $d=\{w_1,w_2,\cdots,w_n\}$,词的特征权值为向量,则文本集 D 可视做文本为行、词汇为列的矩阵。因为每篇文本词汇量不多,是一个稀疏的矩阵,为了节省存储空间,在实际存储时采用 word-index:weight 的格式存储。其中:word-index 为词的索引值,测试语料矩阵中的词索引必须与训练语料中的相互对应;weight 为词在文本中的权值,每两个不同词的向量之间用空格隔开。一个文本一行,就形成了一个矩阵文本。

2　文本情感分类器构建

情感文本集经过处理后得到的向量矩阵,就可以作为情感分类器训练和评测的数据,本文采用了朴素贝叶斯方法来构建情感分类器[7]。朴素贝叶斯是一种基于概率的学习算法,它基于假设的先验概率,给定假设下观察不同特征的概率。对于文本 $d=\{w_1,w_2,\cdots,w_n\}$ 的情感倾向属于 $C=\{c_P,c_N\}$,在特征相互独立的情况下,考虑特征词的权值,其分类算法如式(4)。

$$c_{NB}=\underset{c_j\in C}{\operatorname{argmax}}\{P(c_j)\prod_{i=1}^{n}P(w_i,c_j)^{wt(w_i)}\} \tag{4}$$

其中:$P(c_j)$ 是类别 c_j 的先验概率;$P(w_i,c_j)$ 是特征词 w_i 在类别 c_j 中的后验概率;$w_t(w_i)$ 是测试语料中词 w_i 的权值,当采用 bool 型权值时,$wt(w_i)=1$。对于先验概率 $P(c_j)$,可以视每类是相同的,也可以预先估计,本文采用的是根据已正确标注的训练语料预先估计,计算方式如式(5)。

$$P(c_j)=\frac{\text{doc}(c_j)}{\sum_{c_j\in C}\text{doc}(c_j)} \tag{5}$$

其中:$\text{doc}(c_j)$ 是属于类别 c_j 的文本数。

对于后验概率 $P(w_i,c_j)$,即特征词 w_i 出现在类别 c_j 中的概率,一般是从训练语料中通过计算进行估计。本文采用词 w_i 在属于类别 c_j 的文本中的权值之和除以类别 c_j 的文本中所有词的权值之和。为避免 $P(w_i,c_j)$ 等于 0,可采用 Laplace 转换,由此后验概率计算方法如式(6)。其中 $\text{weight}(w_i,c_j)$ 是词 w_i 在属于类别 c_j 的文本中的权值之和。

$$P(w_i,c_j)=\frac{\text{weight}(w_i,c_j)+\delta}{\sum_{i=1}^{n}\text{weight}(w_i,c_j)+\delta|V|}$$
$$\left(V=\sum_{c_j\in C}\sum_{i=1}^{n}\text{weight}(w_i,c_j),\delta=\frac{1}{|V|}\right) \tag{6}$$

采用 Laplace 转换,一般情况常数 V 取所有词的权值总和,δ 取 1。当 $\delta=1$ 时存在一些问题,增大了训练语料中未出现的特征词的存在概率,并且缩小了出现词的概率。为了解决这个

问题,本文取 $\delta = 1/|V|$,等效于当特征词不存在时,后验概率为一个极小的存在概率。当特征词存在时,也不影响原有的概率。

朴素贝叶斯分类器又被称为最优分类器,其分类算法实现比较简单,分类效率也比较高,在文本分类方面表现比较好。在利用朴素贝叶斯分类器进行文本分类时,需要先进行训练,估计类别的先验概率和特征的后验概率,再进行分类。

3　分类实验及结果分析

根据第 1、2 章介绍的文本处理和分类器构建方法,对采集来的宾馆评论语料进行了文本情感分类实验。本章主要是对使用 CHI 统计和情感词典两种不同的特征选择方法,以及词频和 Bool 两种权值进行对比实验研究,结果显示使用情感词典优于使用 CHI 统计进行特征选择。

3.1　中文情感语料采集及处理介绍

笔者通过自己开发的网页自动采集程序,从携程网(www.ctrip.com)下载了 2008—2009 年北京(BJ)、上海(SH)、广州(GZ)三个城市的宾馆评论,作为本文进行研究所需的语料库。在该语料库中的每条评论,用户都从房间卫生、酒店服务、周边环境和设施设备四个方面进行了给分,并有综合得分,分值为 1~5 分。将综合得分大于等于 3 分的评论标注为正向评论,小于 3 分的评论标注为负向评论。虽然该语料库的质量和标注不规范,但也没有对其进行专业的整理,笔者认为该语料库符合当前互联网的实际情况,对研究互联网中文网页情感分类具有较好的意义。语料库组成结构如表 1 所示。

表 1　测试语料库组成结构

语料	数量	正面	负面	中文特征词	中文情感词
BJ	43 006	39 211	3 795	19 084	2 209
SH	55 786	50 642	5 144	19 028	2 232
GZ	16 852	15 108	1 744	13 500	1 642

本文从三个语料库并集中选择了综合得分为 1 分的评论 1 000 篇和为 5 分的评论 10 000 篇作为训练语料,按照正面和负面文档的比例,共组成了三种训练语料库。其中 T1 为平衡语料、T5 中正面与负面的文本数量为 5∶1,T10 中正面与负面的文本比例与测试语料基本相等,正面与负面的文本数量为 10∶1,其组成结构如表 2 所示。

表 2　训练语料库组成结构

语料	数量	正面	负面	中文特征词	中文情感词
T1	2 000	1 000	1 000	7 776	1 062
T5	6 000	5 000	1 000	10 597	1 352
T10	11 000	10 000	1 000	12 679	1 560

3.2　评价指标

本文对分类器的性能进行评测时,采用了微平均(F_1)作为评价分类结果的指标,需先计算查准率(precision)和召回率(recall),计算方法如式(7)~(9)。

$$\text{precision} = \frac{\sum_{c_j \in C} \text{true}(c_j)}{\sum_{c_j \in C} \text{doc}(c_j)} \tag{7}$$

$$\text{recall} = \frac{\sum_{c_j \in C} \text{true}(c_j)}{\sum_{c_j \in C} \text{response}(c_j)} \tag{8}$$

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \times 100\% \tag{9}$$

式(7)(8)中的 $\text{true}(c_j)$ 是分类为 c_j 并且正确的文本数,式(8)中的 $\text{response}(c_j)$ 是分类为 c_j 的文本数。对于正向和负向综合概率相等或者没有特征的文本,将其作为没有情感倾向的客观描述,不作分类。如果出现无法分类的文本,文本集的查全率、查准率和微平均不相等。

3.3　分类实验及分析

实验 1　使用 CHI 统计方法选择特征(维度设置为 2 000)和情感词典选择特征,采用词频和 Bool 型权值,分别使用 T1、T5、T10 语料进行训练,对 BJ、SH、GZ 语料库进行了分类测试,实验结果如表 3 和 4 所示。

表 3　使用 CHI 统计方法的分类实验结果(F_1:100%)

语料	T1		T5		T10	
	词频	Bool	词频	Bool	词频	Bool
BJ	75.36	74.62	83.50	83.56	85.54	85.84
SH	75.69	75.13	83.05	83.33	84.89	85.48
GZ	73.04	72.72	81.52	81.87	83.44	84.20

表 4　使用情感词典的分类实验结果(F_1:100%)

语料	T1		T5		T10	
	词频	Bool	词频	Bool	词频	Bool
BJ	77.07	75.92	86.76	86.79	88.70	88.89
SH	77.03	75.97	85.99	85.95	87.80	87.97
GZ	75.23	73.90	84.50	84.63	86.38	86.44

对实验结果进行比较可以看出,只选用情感词作为特征选择,在所有语料上的分类结果微平均都得到了提升,这说明将情感词典作为特征选择,可以提高文本情感分类的效果。使用情感词典还有几个优点:使用情感词典不用考虑训练和测试语料之间特征词索引对照的问题,直接使用情感词典中词的索引,不需要对语料进行重新处理,节省了大量的文本处理时间;使用降维方法,选择维度多少才是最佳的结果是个难题,实际应用中不可能多次进行试验分析,而使用情感词典不用考虑这个问题。

3.4　对大量中文文本进行情感分类应用的讨论

互联网上有海量的信息,并且每天都在不断增加,如对产品的评论,人们在使用后才可能发表到网上。所以说互联网上的评论有两个特点,一是海量的,二是不断增多的。如果想在互联网上运用情感分类,就需要一个快速稳定,并且可以持续运行的分类器。

使用朴素贝叶斯分类器,可以将类别的先验概率和特征的后验概率存储起来,以此可以多次分类,但是如果想对先验和后验概率进行修正,就必须重新计算。从式(5)(6)中可以看出,只要统计 $\text{doc}(c_j)$ 和 $\text{weight}(w_i, c_j)$ 就可以计算出先验和后验概率,如果改为直接存储 $\text{doc}(c_j)$ 和 $\text{weight}(w_i, c_j)$ 的值,就可以实现随时对先验和后验概率进行修正。并且选用情感词典作为特征,固定的特征可以使分类器的训练和测试能够持续进行,而不用担心特征对应的问题。这样来,当一个或一组文本处理完成后,就可以把新增的文本数和特征权值累加到存储结果中,从而达到修正先验和后验概率的目的。

实验 2　先取小部分语料(T1)进行训练,构建一个基础的概率模型,然后将语料库(BJ)定量 2 000 条不断输入到分类器测试,每次输入先对已经输入的所有语料进行测试,得到测试的评价指标,之后使用最新一组的输入语料修正概率模型。本文分别采用了两种方式修正,一种是使用语料的标注类别,一种是根据最大期望算法(EM)的思想[8],模拟对语料标注不明情况下的情感分类,是利用分类的结果作为修正概率模型。实验结果如图 2 所示。

(下转第 3743 页)

加少量节点的情况下就可以获得比较大的性能提升，而节点数大于 6 个以后所获得的性能提升没有第 4、5 个节点显著，但仍可以设想如果在更大数据量情况下，越多节点应该会获得更好的效率。

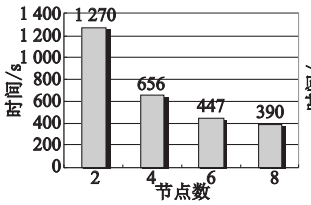

图4　节点数效率对比

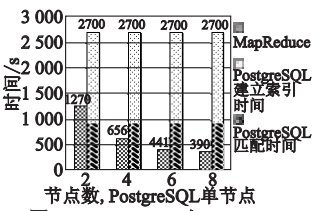

图5　MapReduce与PostgreSQL性能对比

天文星表数据一般只需要匹配一次就可以作为结果一直使用，所以可以认为数据库建立索引时间和匹配时间的和需要一起与 MapReduce 进行比较。可以看到，MapReduce 有非常大的优势，但是在非大规模两个星表匹配查询中，如单个点的匹配查询，建立了索引的 DBMS 的性能较好。

4　结束语

本文将 MapReduce 引入天文星表交叉认证领域中，给出了详细的算法设计与实现，介绍了利用 MapReduce 框架实现并行星表交叉认证。与传统的 PostgreSQL 相比，使用 MapReduce 无须建立索引，不仅可以获得更好的性能，而且支持使用者自由地调节误差半径等匹配参数；此外 MapReduce 可以在分布式计算环境中使用，从而有效地支持批量查询。

在今后工作中，将测试更大的数据集，并对星表交叉认证算法进行优化，扩展其功能，以实现更快速、准确的匹配。此外，随着 MapReduce 和并行数据库技术的不断发展，两者正在呈现融合的趋势。类 SQL 的 HBase、Pig、Hive 等 MapReduce 新项目在不断涌现。传统的 DBMS 也在尝试利用 MapReduce 来解决一些如建立 R-tree 等问题。未来的工作中也将尝试混合 MapReduce 与并行数据库两种架构来实现星表交叉认证。

参考文献：

[1] DJORGOVSKI S G, BRUNNER R J. Astronomical archives of the future: a virtual observatory[J]. Future Generation Computer Systems, 1999, 16(1):63-72.

[2] CUI Chen-zhou, ZHAO Yong-heng. Worldwide R&D of virtual observatory[J]. Proceedings of the International Astronomical Union, 2007, 3:563-564.

[3] Viewing the heavens through the cloud [EB/OL]. [2009-12-14]. http://ssg.astro.washington.edu/research.shtml? research/CluE1.

[4] ZHAO Qing, SUN Ji-zhou, YU Ce, *et al*. A paralleled large-scale astronomical cross-matching function[C]//Proc of Lecture Notes in Computer Science, vol 5574. 2009:604-614.

[5] 高丹，张彦霞，赵永恒. 中国虚拟天文台交叉认证工具的开发和应用[J]. 天文学报，2008，49(3):348-358.

[6] CGP. Report on cross matching catalogues [EB/OL]. (2003-09-29) [2009-12-14]. http://wiki.astrogrid.org/pub/Astrogrid/DataFederationandDataMining/cross.htm.

[7] POWER R. Cross match simulation[CP/OL]. (2007-04-23)[2009-12-14]. http://www.ict.csiro.au/staff/robert.power/projects/CM/ps/cm.htm.

[8] O'MALLEY O. TeraByte sort on Apache Hadoop[EB/OL]. (2008-05)[2009-12-14]. http://sortbenchmark.org/YahooHadoop.pdf.

[9] DEAN J, GHEMAWAT S. MapReduce: simplified data processing on large clusters[J]. Communications of the ACM, 2008, 51(1): 107-113.

[10] CUTRI R M, SKRUTSKIE M F, VAN DYK S, *et al*. 2MASS all sky catalog of point sources, the IRSA 2MASS all-sky point source catalog, NASA/IPAC infrared science archive [EB/OL]. (2003) [2009-12-14]. http://irsa.ipac.caltech.edu/applications/Gator/.

[11] CHURCHWELL E, BABLER B L, MEADE M R, *et al*. The Spitzer/GLIMPSE surveys: a new view of the milky way[J]. Publications of the Astronomical Society of the Pacific, 2009, 121: 213-230.

（上接第 3739 页）

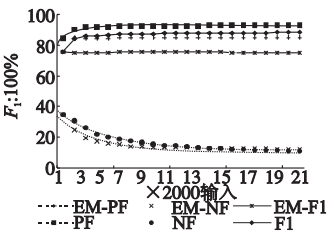

图2　实验2的测试结果

从结果可以看出，使用正确的标注修正概率模型，正向和总体的微平均值是上升的，但对于负向是处于下降的，这是因为测试语料不对称造成的。采用 EM 算法的效果不够好，这是因为分类结果的错误影响了概率模型的修正。如果能构建一个较好的概率模型，以及合理的概率修正机制，使用情感词典和朴素贝叶斯方法可以持续对大量的文本进行快速的情感分类。

4　结束语

通过一系列的实验可以看出，在使用朴素贝叶斯方法进行中文文本情感分类时，其特征选择和构建概率模型是比较重要的。实验表明，特征选择用情感词典效果比较好，但目前中文方面还没有一个比较权威的基础情感词典，并且在不同领域，情感词汇又有区别，所以建立一个全面的中文基础情感词典以及建立不同领域的辅助情感词典是很有必要的。在情感词典的基础上，构建不同领域的朴素贝叶斯概率模型，不断进行修正，可以实现对互联网上大规模网页情感分类、舆论观点分析等工作。但是如何在人工标注较少的情况下，构建较好的概率模型还值得深入研究。

参考文献：

[1] 李军. 中文评论的褒贬义分类试验研究[D]. 北京：清华大学，2008.

[2] 王素格，魏英杰. 停用词表对中文文本情感分类的影响[J]. 情报学报，2008，27(2):175-179.

[3] 《现代汉语常用词表》课题组. 现代汉语常用词表(草案)[K]. 北京：商务印书馆，2008.

[4] 王素格，杨安娜，李德玉. 基于汉语情感词表的句子情感倾向分类研究[J]. 计算机工程与应用，2009，45(24):153-155.

[5] 柳位平，朱艳辉，栗春亮，等. 中文基础情感词词典构建方法研究[J]. 计算机应用，2009，29(11)：2882-2884.

[6] YANG Yi-ming, PEDERSEN J O. A comparative study on feature selection in text categorization[C]//Proc of the 14th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann, 1997: 412-420.

[7] MITCHELL T M. 机器学习[M]. 北京：机械工业出版社，2003.

[8] 李静梅，孙丽华，张巧荣，等. 一种文本处理中的朴素贝叶斯分类器[J]. 哈尔滨工程大学学报，2003，24(1):71-74.