第43卷 第4期
2016年4月

计 算 机 科 学
Computer Science

Vol. 43 No. 4
Apr 2016

基于短语的贝叶斯中文垃圾邮件过滤方法

王青松　魏如玉

（辽宁大学信息学院　沈阳 110036）

摘　要　朴素贝叶斯算法在垃圾邮件过滤领域得到了广泛应用，该算法中，特征提取是一个必不可少的环节。过去针对中文的垃圾邮件过滤方法都以词作为文本的特征项单位进行提取，面对大规模的邮件训练样本，这种算法的时间效率会成为邮件过滤技术中的一个瓶颈。对此，提出一种基于短语的贝叶斯中文垃圾邮件过滤方法，在特征项提取阶段结合文本分类领域提出的新的短语分析方法，按照基本名词短语、基本动词短语、基本语义分析规则，以短语为单位进行提取。通过分别以词和短语为单位进行垃圾邮件过滤的对比测试实验证实了所提出方法的有效性。

关键词　垃圾邮件过滤，贝叶斯，特征项提取，基于短语，中文分词

中图法分类号　TP393.098　　**文献标识码**　A　　**DOI**　10.11896/j.issn.1002-137X.2016.4.052

Bayesian Chinese Spam Filtering Method Based on Phrases

WANG Qing-song　WEI Ru-yu

(College of Information, Liaoning University, Shenyang 110036, China)

Abstract　Naive Bayesian has been widely used in the field of spam filtering, in which the feature extraction is one of the essential links in the algorithm. In the past, only words were used as text features for the extraction in the method of Chinese spam filtering. In face of large-scale email training samples, time efficiency of this algorithm will become a bottleneck of spam filtering technology. A Bayesian spam filtering algorithm based on phrases was proposed here which combines a new phrase analysis method put forward in text classification field. Phrases are extracted as the unit according to the rules of basic noun phrases, verb phrases and semantic analysis. Through comparison test experiment of spam filtering based on words and phrases as unit, the effectiveness of the proposed method was confirmed.

Keywords　Spam filtering, Bayesian, Feature extraction, Phrased-based, Chinese word segmentation

随着 Internet 的全球化普及，电子邮件成为了常用的通讯交流方式，与此同时日益增多的垃圾邮件给人们的生产生活带来了极大的不便，造成了人力物力资源的浪费。据统计，我国已经成为全球第二大垃圾邮件受害国[1,2]。

目前主要的垃圾邮件过滤技术为黑名单技术[3]、基于规则的过滤技术[4]以及贝叶斯过滤技术[5]。其中对于黑名单技术，由于数据库需要不断更新，维护工作量大且不方便，容易导致正常邮件的丢失。基于规则的过滤技术虽然在过滤垃圾邮件初期表现良好，但其静态属性使得垃圾邮件发送者可以根据规律绕开检查。贝叶斯过滤技术[6]首先获得大量的训练邮件样本，从中提取出一组相关性极高的特征项，分别计算每个特征项在垃圾邮件和正常邮件中的比率，然后根据贝叶斯公式计算出新接收的邮件属于垃圾邮件和正常邮件的概率，最终判断该邮件是否为垃圾邮件。从目前的分类效果来看，这种方法以其运算速度快、易于实现等特点被广泛应用[7,8]。但由于我国主要的邮件语言为中文，因此所要关注的问题是面向中文的垃圾邮件过滤。贝叶斯方法虽然在英文环境效果显著，但在中文邮件[9]中，文本的原始特征空间维数比英文文本大，词性变化更加灵活，词语间也不像英文那样有明显的间隔，这些区别给分词带来了更大的难度，对贝叶斯方法也提出了更高的要求。

传统的贝叶斯中文垃圾邮件过滤[10]都是以词为单位进行分词，但单个单词所能涵盖的文本特征含义十分有限。词语的粗糙性以及不完备性会导致分类阶段的低效和不准确，并且给整个过滤系统带来负面影响。而短语作为由几个单词按一定的语序和语法规则组成的序列，包含更多的信息，可以有效弥补以词为单位的欠缺。本文即是根据中文邮件的特点，结合文本分类、文本挖掘等领域中赵军、詹卫东等提出的关于短语的分析方法[11-13]，提出了一种基于短语的贝叶斯中文垃圾邮件过滤方法，较为详细地给出了基本名词短语、基本动词短语的界定以及语义分析方面的约束。提出的基于短语的贝叶斯中文垃圾邮件过滤方法以短语作为特征项的优势有：

1）从提取的特征项数量上来看，以短语为单位的规模比以词为单位的小很多，大大降低了特征向量空间的维数，从而为基于向量空间的贝叶斯垃圾邮件过滤减少了很多计算量，

到稿日期：2015-03-31　返修日期：2015-07-13　本文受国家社科青年基金项目：基于空间计量分析的人口规模、结构对资源环境的影响效应研究(13CRK027)资助。

王青松(1974－)，男，硕士，副教授，主要研究方向为数据挖掘、数据库及其应用，E-mail：1301833668@qq.com；**魏如玉**(1990－)，女，硕士生，主要研究方向为机器学习、数据挖掘，E-mail：704254105@qq.com。

提高了系统效率。

2)从对文本的表达能力上来说，短语的结构相对稳定，表达的内容相对于同类词语来说更为精确，避免了同义词和多义词，所以短语更能准确地表示邮件文本的内容特点，有效提高了分类的准确率。

1 邮件预处理

1.1 邮件向量化

电子邮件本身是一种无结构的文本，为了使计算机能够对邮件进行学习和处理，一般要采用向量空间模型(Vector Space Model)，将训练集文档都转化为向量表示之后所有的运算都基于这些向量进行，不再使用原始的文本形式的文档。

对于任意一封电子邮件 d，对应的特征向量为：$x=\{x_1, x_2, \cdots, x_n\}$，$x_{i=0,1\cdots n}\in(0,1)$，其中 $x_1, x_2, \cdots, x_n$ 分别为邮件 d 对应的特征项 $X_1, X_2, \cdots, X_n$ 的特征值。若特征项 X_i 在邮件 d 中出现，则 x_i 取 1，否则 x_i 取 0[14]。

1.2 分词和词性标注

中文文本自动分词技术经过十几年的发展取得了很大的进步，出现了一些实用的分词系统[15-17]，这些系统在分词速度和精确度上都达到了相当高的水平。本文研究方法采用中国科学院计算技术研究所研发的 ICTCLAS 汉语分词系统实现自动分词。另外，词性标注是指为文本中的每一个词都标记一个合适的标记，也就是说要确定每个词是名词、动词、形容词或其他词性。例如：

中国科学院/*nt* 计算技术/*n* 研究所/*n* 汉语分词/*n* 系统/*n*

其中 *nt* 表示机构团体名，*n* 表示名词，并在集合中使用停用词表删除助词、虚词等无意义或者贡献能力不大的词语。

2 邮件特征项的提取

相比于分类算法，对特征项的选择(即使用哪些特征项来代表一篇文本)往往更能影响分类的效果。本文提出的方法与传统方法的不同之处在于特征项的提取采用基于基本短语的模式。虽然构成短语这一阶段会增加计算量，但由短语代替单个词语所减少的空间维度以及由短语稳定性所带来的更强的文本表达力则会大大减少后续过滤过程中的计算复杂度。基本短语可定义为以下几种结构[18]：粘合式定中结构、粘合式状中结构、粘合式述补结构、粘合式联合结构。本文主要介绍能代表文本主要特征的基本名词短语和基本动词短语的构成和语义分析。下面介绍基于短语模式的特征项提取应遵循的几个规则。

2.1 基本名词短语的界定

短语的界定即为确定不同类型的短语的边界位置的过程，是单词构成短语的主要步骤，作为能代表文本主要特征的基本名词短语和基本动词短语，它们的界定规则对特征项空间的维度和准确性来说非常重要。Church 首次定义英语中的 baseNP[19] 为“简单的非嵌套的名词短语”，即一个基本名词短语内部不能再嵌套更小的名词短语，但其并不符合汉语中几个名词也可以组成名词短语的特点，我国赵军、黄昌宁[20] 等人根据词语的潜在依存关系分析了汉语 baseNP 的结构，并给出了 baseNP 的相关定义。本文在该定义的基础上做了相应的扩展来界定 baseNP。在这里只给出汉语 baseNP 的形式化描述：

1)baseNP→baseNP＋baseNP

2)baseNP→baseNP＋名词|名动词

3)baseNP→限定性定语＋baseNP

4)baseNP→限定性定语＋名词|名动词

其中限定性定语为形容词、区别词、动词、名词、处所词、量词等。例如，“操作系统”、“公司员工”等短语由表达式 1)来表示，而“火车提速”、“出国旅游”等短语可由表达式 2)来定义，表达式 3)定义了“一双鞋”、“三好学生”等短语，表达式 4)定义了类似“自然语言处理”和“第三世界国家”等短语。

2.2 基本动词短语的界定

与基本名词短语不同，至今没有统一的定义来规范基本动词短语，所以用它们经常出现的形式来界定，主要有以下几种形式[21]：

1)述宾结构：作思想报告、交换角色；

2)述补结构：拿到、不想睡、逛商场；

3)状中结构：马上回答、去北京工作；

4)连动结构：来工作、去劳动；

5)联合结构：工作和生活、边走边唱；

6)除此之外，还包括一些常用的动词短语，如：想了想、听说过、睡了、坐着等。把这些附加了“着”、“了”、“过”的词记录为具有“着了过”属性的动词。

2.3 基本短语的语义分析

短语比词涵盖更多的信息，但与此同时在短语的边界界定方面难度也更大，除了根据不同短语类型进行界定外，语义分析也是十分普遍和重要的方法。在任意两个或者两个以上的成分组成的结构中，如果不只有一种结构可选，而是几种结构关系都成立，那么就存在歧义。比如：“几个学校的老师出席了这次大会”，可以理解为“几个学校的/老师/出席了这次大会”；也可以理解为“几个/学校的老师/出席了这次大会”。语义分析的目的是要正确划分和识别短语所描述的概念内涵和外延，同时避免了以短语为单位会扩展分类的特征空间的问题。

常见关于短语的分析方法主要有基于规则、基于统计以及规则和统计相结合的方法。本文采用规则和统计相结合的方法，其基本思想为：从文本的第一个分词词语开始分别与后面的词语进行组合，经过上述过程检验且符合上述规则的即为有效短语，若不符合则取下一个词语重复上述过程，直到文本的最后一个词语完成上述操作为止。通常组成短语的词语数量是不固定的，设定的范围大一些才能尽可能多地涵盖更多的短语，这里参照赵蕾蕾[22] 等人所提方法中设置的 5 个为一组，为了避免词语的重复组合，第一词语是每次组合都要包含在内的。由于生成了大量的短语，其特征项也相应得到了增加，但是其中也存在一些意思相近或相同的短语，所以组合后的短语仍要保留其词性标注，以便之后进行语义分析和词频的统计。基于统计的思想采用互信息的方法来划分短语的边界。互信息可以度量一个消息中两个信号之间的相互依赖程度，而在这里互信息表示的是词语之间结合的紧密程度，可通过句中相邻的词性标记互信息值的大小来判断，一般互信息值极小的位置为短语的边界。互信息方法的计算方法如式(1)所示：

$$MI(F)=\sum_i P(C_i)\log\frac{P(W|C_i)}{P(W)} \tag{1}$$

其中,C_i 是类别,$P(W|C_i)$表示文本中出现词 W 时文本属于 C_i 的概率,$P(C_i)$是类别出现的概率。

2.4 特征项提取算法

本算法基于邮件向量空间,前提是已经将搜集的邮件按第 1 节的邮件预处理步骤进行了转化。其中 w_0 为原始的向量空间,w_1 为经过字符串匹配之后的向量空间,w_2 为利用互信息方法确定边界后的向量空间,w_3 为新生成的由词和短语组成的向量空间,s 为停用词表,X_i 为样本中的所有单词。

算法 1　基于短语的特征项提取算法

输入:从邮件样本中提取所有单词 X_i 组成的向量空间 w_0
输出:由词和短语组成的新的向量空间 w_3
If(w_0 非空)
　用分词系统对 w_0 中的 X_i 进行分词和词性标注
　取标注后的向量空间中的每个单词 X_i
　If(X_i 为语气助词、介词、连词等虚词)
　Do 从 w_0 中删去 X_i
　　将 X_i 添至停用词表 s 中
　For(对于处理过的 w_0 中的每个单词,取 5 个为一组)
　　用字符串匹配的方法实现其与相邻词语的随机搭配
　　得到向量空间 w_1
　End For
　For(对于包含在 w_1 中的每个词语)
　　利用互信息方法计算词语之间结合的紧密程度
　　选择值极小的位置为短语的边界得到向量空间 w_2
　End For
　If (w_2 中短语符合基本短语规则 && 符合基本短语语义约束)
　Do 保留其短语属性并加至向量空间 w_3
　Else 保留其词的属性并加至向量空间 w_3
End If
输出由词和短语组成的新的向量空间 w_3

在上述特征项提取算法中采用统计和规则相结合的基本短语构成算法,并根据基本短语的定义实现了由词到基本短语的转换。主要以基本短语作为文本的特征项,对于没有组成基本短语的词语,经过筛选,部分词语也作为特征项保留。弥补了传统的贝叶斯算法中只采用词语作为文本的特征项所带来的局限性。

3 基于短语的贝叶斯垃圾邮件过滤

贝叶斯垃圾邮件过滤器是一种贝叶斯分类器,即将邮件分成正常和垃圾两类。我们在第 1 节中已经提取了反映邮件是否正常的特征向量$(x_1,x_2\cdots,x_n)$,并分别用 $x_i=1$ 和 $x_i=0$ 来表示此特征在本邮件中存在和不存在两种情况。但贝叶斯方法计算量很大,较难实现,本文采用简化过的朴素贝叶斯分类。这种算法是在假设文本中词的分类是相互独立的前提下进行的,选取概率最大的类别作为待分类文本的类别。实际应用虽然不能完全满足其独立性要求,但大量实验[23,24]表明用此假设构造的分类器确实十分有效。数学模型表达式为:

$$P(C_i|D_j)=\frac{P(C_i)P(D_j|C_i)}{P(D_j)} \tag{2}$$

其中,C_i 为某一类别,D_j 为待分类的文本;$P(C_i)$从训练集中估计。对于不同的类别,式中的分母是不变的,选择分子最大的类别作为待分类文本的类别。图 1 即为完整的基于短语的朴素贝叶斯垃圾邮件过滤系统流程。

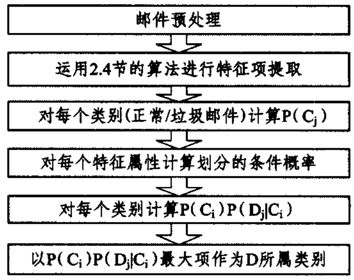

图 1　训练阶段和分类阶段的流程

相较于传统的朴素贝叶斯垃圾邮件过滤,本文所提的基于短语的贝叶斯中文垃圾邮件过滤方法的主要改进在于第 2 步中的特征项提取算法,即使用第 2 节中所提出的以短语为单位进行特征提取。虽然生成短语时进行必要的句法分析会在一定程度上增加方法的计算量,但短语同时也大大减少了以词为单位时大量的同义词和多义词的现象,不仅大大缩减了向量空间,而且提高了对文本的表达能力。之后的先验概率和条件概率都是在提取的特征向量集的基础上进行计算和统计的,改进之后的算法可以大大减少特征向量集的规模,进而减少之后的计算量,使系统效率和性能有较大的提升。

4 实验与结果分析

实验的目的是比较传统的朴素贝叶斯中文垃圾邮件过滤算法与基于短语模式的朴素贝叶斯垃圾邮件过滤算法的性能。为了能够清晰地表达比较的结果,我们引入了几个参数:查准率 SP(Spam Precision)、查全率 SR(Spam Recall)和综合指标 F1。分别定义为:

$$SP=\frac{\text{正确识别出的垃圾邮件数}}{\text{识别出的垃圾邮件总数}} \tag{3}$$

$$SR=\frac{\text{正确识别出的垃圾邮件数}}{\text{样本中垃圾邮件总数}} \tag{4}$$

$$F1=\frac{SP*SR*2}{SP+SR} \tag{5}$$

其中,SP 侧重安全性,反映过滤的正常邮件和垃圾邮件的关系;SR 侧重有效性,即被过滤掉的垃圾邮件数占总垃圾邮件数的百分比;F1 则综合考虑了查准率和查全率两方面的指标。

对于两项实验数据比较,首先比较提取特征项阶段传统方法中以词为单位和本文提出的以短语为单位提取的特征向量的维数;然后比较词模式下 KNN 垃圾邮件过滤方法、词模式下贝叶斯垃圾邮件过滤方法以及短语模式下贝叶斯垃圾邮件过滤方法的查全率、查准率以及综合指标。鉴于 KNN 也是当前比较流行的基于内容的反垃圾邮件方法,前两者的比较可以验证贝叶斯方法的效率,而后两者的比较可以验证以词为单位是否会提高系统的整体性能。所用的语料选自中国教育和科研计算机网紧急响应组提供的电子邮件数据集 CCERT 2005-Jul,这个数据集中包括一个正常邮件集和一个垃圾邮件集,我们从中选择了 3000 封正常邮件和 1200 封垃圾邮件,其中 html 标记和附件都已被除去。实验环境为 Intel Core i3-2330M CPU、2. 20GHz 主频、2GB 内存、Windows7 Professional。使用 JAVA 语言在 Eclipse 平台上对算法进行

测试。所做实验重在比较这两种方式的差异，因此并没有结合白名单、黑名单技术，在数据上可能与侧重点不同的其他实验得出的结果有些差别。图 2 为在特征提取阶段两种不同的方式下提取的特征项数量的比较。

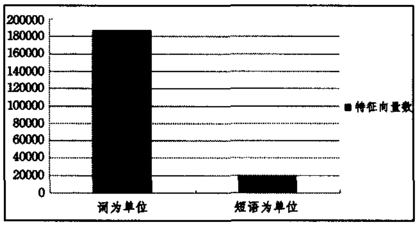

图 2　两种方式下特征项数量的比较

其中以短语为单位所得特征项数目为 17976，包括基本名词短语（BaseNP）12000 个，基本动词短语（BaseVP）5976 个，远远少于以词为单位所提取到的数目 186666。由实验数据可知使用本文提出的方法可以大大缩小空间向量维数，为之后使用朴素贝叶斯算法进行垃圾邮件过滤节省了很多计算量，提高了系统的效率。

短语能够比词涵盖更多的信息，在邮件分类方面也更能提高准确度。下面测试在 3 种不同方式下垃圾邮件过滤在查准率、查全率和综合指标这些系统性能上的表现。结果比较结果如图 3 所示。

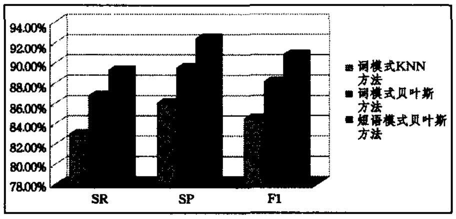

图 3　3 种方式下系统性能的比较

从所得的实验结果可以看出：同样是以词为单位进行特征提取、垃圾邮件过滤，贝叶斯方法在查全率、查准率上达到的 86.74%和 89.50%均高于 KNN 方法的 82.96%和 86.00%；而在贝叶斯方法中，以短语为单位进行特征提取然后再进行分类过滤所得的结果 89.24%、92.36%则更优于以词为单位的贝叶斯过滤方法。

结束语　本文结合文本分类方法中的短语分析规则，提出了一种基于短语的中文贝叶斯垃圾邮件过滤方法。本方法中，特征项提取阶段以短语为单位，生成维度更小精确度更高的向量空间，再以此为基础进行垃圾邮件过滤。实验测试表明，与其他先进方法（如 KNN）和以词为单位提取特征项的贝叶斯中文垃圾邮件过滤方法相比，本方法在查全率和查准率方面都有一定的提高。

但本研究还存在一些不足。尽管词和短语已经能较完善地反映文本内容，但是在特征项提取阶段仍然有进一步改善的可能。另外本文在实验测试前已将邮件的附件除去，若在附件中携带病毒，则它很容易入侵到计算机，这方面研究还需要很大的努力。

参考文献

[1] China Internet Network Information Center. China Internet network development state statistic report[R]. Beijing: China Internet Network Information Center, 2004(in Chinese)
中国互联网络信息中心. 中国互联网络发展状况统计报告[R]. 北京：中国互联网信息中心，2004

[2] Zhai Jun-chang, Qin Yu-ping, Che Wei-wei. Improvement of Information Gain in Spam Filtering[J]. Computer Science, 2014, 41(6): 214-224(in Chinese)
翟军昌，秦玉平，车伟伟. 垃圾邮件过滤中信息增益的改进研究[J]. 计算机科学，2014，41(6)：214-224

[3] Xu Ji, Gong Jian. An Integrated Way to Filter Spam[J]. Computer Science, 2005, 32(2): 69-72, 86(in Chinese)
徐激，龚俭. 垃圾邮件的综合过滤方法[J]. 计算机科学，2005，32(2)：69-72，86

[4] Li Yu-feng, Gao Xiao-jing. Comprehensive Approach For Chinese Spam Email Filtering[J]. Computer Applications and Software, 2011, 28(8): 220-226(in Chinese)
李玉峰，郜晓晶. 中文垃圾邮件过滤综合方法[J]. 计算机应用与软件，2011，28(8)：220-226

[5] Androutsopoulos I, Sakkis G, Paliouras G, et al. Learning to Filter Spam E-Mail[C]// European Conference on Principles and Practice of Knowledge Discovery in Databases. Lyon, France, 2000: 1-13

[6] Zhang Ming-feng, Li Yun-chun, Li Wei. Survey of Application of Bayesian Classifying Method to Spam Filtering[J]. Application Research of Computers, 2005(8): 14-19(in Chinese)
张铭锋，李云春，李巍. 垃圾邮件过滤的贝叶斯方法综述[J]. 计算机应用研究，2005(8)：14-19

[7] Androutsopoulos I, Koutsias J, Chandrinos K V, et al. An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Encrypted Personal E-mail Messages[C]//Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Athens, Greece, 2000: 160-167

[8] Etzold D. Improving spam filtering by combining Naive Bayes with simple k-nearest neighbor searches [OL]. http://cds.cern.ch/record/688245

[9] Lu Jian-jiang, Zhang Wen-xian. Design for Chinese Text Classier[J]. Computer Engineering and Applications, 2002, 15: 49-51(in Chinese)
陆建江，张文献. 中文文本分类器的设计[J]. 计算机工程与应用，2002，15：49-51

[10] Huang Zhi-gang. Chinese Spam Filtering System Design and Implementation Based on Bayesian[D]. Chengdu: University of Electronic Science and Technology of China, 2007(in Chinese)
黄志刚. 基于贝叶斯的中文垃圾邮件过滤系统的设计与实现[D]. 成都：电子科技大学，2007

[11] Zhao Jun, Huang Chang-ning. The Model for Chinese Basenp Structure Analysis[J]. Chinese Journal of Computers, 1999, 22(2): 141-146(in Chinese)
赵军，黄昌宁. 汉语基本名词短语结构分析模型[J]. 计算机学报，1999，22(2)：141-146

[12] Zhan Wei-dong. A Study of Constructing Rules of Phrases in Contemporary Chinese for Chinese Information Processing [M]. Beijing: Tsinghua University Press, 2000: 106-115(in Chinese)

（下转第 269 页）

下数据稀疏的区域可能造成一定的不良影响，但并不影响数据挖掘模型整体挖掘精确度的提升，实验结果也说明了这一点。而如何消除该方法可能带来的些许不良影响，则可作为非线性标准化方法未来的一个研究方向。

参考文献

[1] Kamiran F, Calders T. Data preprocessing techniques for classification without discrimination[J]. Knowledge and Information Systems, 2012, 33(1): 1-33

[2] Guo Xi-yue, He Ting-ting. Survey about Research on Information Extraction[J]. Computer Science, 2015, 42(2): 14-17(in Chinese)
郭喜跃，何婷婷. 信息抽取研究综述[J]. 计算机科学，2015, 42(2): 14-17

[3] Wang R Y, Storey V C, Firth C P. A framework for analysis of data quality research[J]. IEEE Transactions on Knowledge and Data Engineering, 1995, 7(4): 623-640

[4] Jiawei H, Kamber M. Data mining: concepts and techniques [M]. San Francisco, CA, ltd: Morgan Kaufmann, 2001

[5] Weigend A S. Time series prediction: forecasting the future and understanding the past[R]. Santa Fe Institute Studies in the Sciences of Complexity, 1994

[6] Mendelsohn L. Preprocessing data for neural networks [OL]. https://www.tradertech.com/mendelsohn/library/neural-networks/preprocessing-data

[7] Yu L, Wang S, Lai K K. An integrated data preparation scheme for neural network data analysis[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(2): 217-230

[8] Liping Y, Yuntao P, Yishan W. Research on data normalization methods in multi-attribute evaluation[C]// International Conference on Computational Intelligence and Software Engineering, 2009(CiSE 2009). IEEE, 2009: 1-5

[9] Pyle D. Data preparation for data mining [M]. Morgan Kaufmann, 1999

[10] Uragun B, Rajan R. Developing an appropriate data normalization method[C]// 2011 10th International Conference on Machine Learning and Applications and Workshops (ICMLA). IEEE, 2011, 2: 195-199

[11] Zhang Yu-nong, Li Ming-ming, Chen Jin-hao, et al. Solving the problem of Runge phenomenon by coefficients and order determination method[J]. Computer Engineering and Applications, 2013, 49(3): 44-49(in Chinese)
张雨浓，李名鸣，陈锦浩，等. 龙格现象难题被解之系数与阶次双确定方法[J]. 计算机工程与应用，2013, 49(3): 44-49

(上接第 259 页)

詹卫东. 面向中文信息处理的现代汉语短语结构规则研究[M]. 北京：清华大学出版社，2000: 106-115

[13] Zhou Ming, Huang Chang-ning. Approach to the Chinese Dependency Formalism[J]. Journal of Chinese Information Processing, 1994, 3: 35-52 (in Chinese)
周明，黄昌宁. 面向语料库标注的汉语依存体系的探讨[J]. 中文信息学报，1994, 3: 35-52

[14] Androutsopoulos I, Koutsias J, Chandrinos K V, et al. An evaluation of Naive Bayesian anti-spam filtering[C]// Proc of the Workshop on Machine Learning in the New Information Age Joint 11th European Conference on Machine Learning. Barcelona, Spain, 2000: 9-17

[15] Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules in Large Databases[C]// Proceedings of 20th International Conference on Very Large Data Bases(VLDB 1994). Santiagode Chile, Morgan Kaufmann, 1994: 487-499

[16] Park J S, Chen M S, Yu P S. An Effective Hash-Based Algorithm for Mining Association Rules[C]// Proceedings of the ACM SIGMOD International Conference on Management of Data(SIGMOD'95). San Jose, 1995: 175-186

[17] Savasere A, Omiecinski E, Navathe S. An Efficient Algorithm for Mining Association Rules in Large Databases[C]// 21st VLDB Conf. Zurich, Switzerland, 1995: 432-444

[18] Zhang Yu-qi, Zhou Qiang. Automatic Identification of Chinese Base Phrases [J]. Journal of Chinese Information Processing, 2002, 16(6): 1-8(in Chinese)
张昱琪，周强. 汉语基本短语的自动识别[J]. 中文信息学报，2002, 16(6): 1-8

[19] Croft W. Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information [M]. Chicago and London: The University of Chicago Press, 1991: 66-78

[20] Zhao Jun, Huang Chang-ning. A Probabilistic Chinese BaseNP Recognition Model Combined with Syntactic Composition Templates[J]. Journal of Computer Research and Development, 1999, 36(11): 1384-1390(in Chinese)
赵军，黄昌宁. 结合句法组成模板识别汉语基本名词短语的概率模型[J]. 计算机研究与发展，1999, 36(11): 1384-1390

[21] Li Mu, Gao Jian-feng, Huang Chang-ning, et al. Unsupervised Training for Overlapping Ambiguity Resolution in Chinese Word Segmentation[J]. Proceedings of the Second SIGHAN Workshop on Chinese Language Processing(SIGHAN'03). 2003: 1-7

[22] Zhao Lei-lei. Feature extraction method based on the pattern of words and basic phrases[D]. Baoding: Hebei University, 2009(in Chinese)
赵蕾蕾. 基于词和基本短语模式的特征提取方法[D]. 保定：河北大学，2009

[23] Langley P, Wayne L, Thompson K. An analysis of Bayesian classifiers[C]// Proc of the 10th National Conf on Artificial Intelligence. San Jose, California, 1992: 223-227

[24] Dominggos P, Pazzani M. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss[J]. Machine Learning, 1997, 29: 103-130

[25] Wang Guo-yin, Zheng Zheng, Zhang Yi. RIDAS-A Rough Set Based Intelligent Data Analysis System [C]// Proc of First IEEE International Conference on Machine Learning and Cybernitics (ICMLC2002). Beijing, 2002: 646-649

[26] Gu Yi-jun, Fan Xiao-zhong, Wang Jian-hua, et al. Automatic Selection of Chinese Stoplist [J]. Transactions of Beijing Institute of Technology, 2005, 25(4): 337-340(in Chinese)
顾益军，樊孝忠，王建华，等. 中文停用词表的自动选取[J]. 北京理工大学学报，2005, 25(4): 337-340