文章编号:1001 - 9081(2008)01 - 0159 - 03

一种面向专利文献数据的文本自动分类方法

蒋健安[1],陆介平[2],倪巍伟[1],孙志挥[1]
(1. 东南大学 计算机科学与工程学院,南京 210096; 2. 江苏省镇江市科技局,江苏 镇江 212001)
(jja65@163.com)

摘 要:中文专利文献自动分类目前尚无成熟适用的方法。分析了文本自动分类的关键技术,并结合专利数据的特点对无词典分词和权重计算进行了改进,提出了一种适用于专利数据分类的层次分类方法,给出了面向专利文献数据的文本自动分类系统的框架模型。实验表明,该系统具有较好的分类精度与效率。

关键词:文本分类; 专利文献; 国际专利分类码; K-近邻
中图分类号: TP391 **文献标志码**:A

Automatic text categorization for patent data

JIANG Jian-an[1], LU Jie-ping[2], NI Wei-wei[1], SUN Zhi-hui[1]
(1. *College of Computer Science and Technology, Southeast University, Nanjing Jiangsu* 210096, *China*;
2. *Science and Technology Commission of Zhenjiang, Zhenjiang Jiangsu,* 212001, *China*)

Abstract: At present, there are no practical and mature automatic text categorization methods for patent data. Therefore, this paper made a research on several key techniques about text categorization, improved the non-dictionary segment and weight calculation, and then proposed a hierarchical categorization method and an automatic text categorization framework for patent data. The experiment testifies that the system has a good classification accuracy and efficiency.

Key words: text categorization; patent; International Patent Classification(IPC); K-Nearest Neighbor(KNN)

0 引言

随着经济全球化的发展,专利知识产权越来越受到企业的重视,企业用于专利开发的力度不断加大。为了提高检索专利的效率,每一件被核准的专利都会依其技术内容被分类至某一个国际专利分类码(International Patent Classification, IPC)中。通过 IPC 分类,企业可以进行各类技术研发趋势与动向的预测,并能分析国家和竞争公司的整体技术动态,为技术部门跟踪、分析竞争对手的情况提供依据[1,2]。

随着近年来专利申请量的迅速增长,积累了大量的专利文献数据,目前专利分类仍采用传统的手工分类,工作人员根据专利信息内容,对照国际分类表,手工检索出相应的 IPC 分类类别[3]。而 IPC 分类表包含了与发明创造有关的全部知识领域,仅中国专利目前已使用的分类类别有近 5 万个,因此,手工分类方法效率较低,还存在着以下弊端: 1)周期长、费用高、效率低,而且往往需要具有专业知识的人员才能胜任;2)存在分类结果一致性不高的问题,对于相同的专利文献由不同的人来分类,其分类结果可能不相同,甚至是同一个人,在不同时间做相同的分类也可能会有不同的结果。

在文本分类方面,文献[4]提出了一种传统的文本分类算法——SVM 算法,文献[5]讨论了 Bayes 网络算法,在文献[6]中讨论了 Rocchio 算法。但由于专利数据涉及各个专业,生词多,更新快,很难构建适用于各专业领域的词典。已有的这些算法不能有效应用于专利文本数据。

针对这些不足,本文结合专利数据特点提出了一种对专利文献进行自动分类的方法。理论分析和试验结果表明,方法是有效可行的。

1 专利数据的预处理

1.1 基于后缀数组的领域词典

由于专利信息的特殊性,专利信息中包含了几乎所有科学领域的先进科学知识,并且包含各个领域的大量领域词汇。随着科技的发展,会不断产生新的专业词汇,所以预先设定一个领域词典并不是一种理想的方法,找到专利中涉及的所有领域的领域词典并不是一件简单的事情,将不断产生的新的专业词汇不断地加入到词典中也存在工作量大和实时性差的缺点。因此考虑采用基于后缀数组统计的方法来获得相应领域的词汇,并将抽取出的领域词汇构成下一阶段分词的领域词典。基于后缀数组统计的方法可以高效地抽取出文中重复的字符串,本文在此基础上加以改进,并通过对抽取的字符串自动过滤,得到有实际意义的专业词汇。

后缀数组是作为一种文本索引结构提出的,它记录了一个字符串中各后缀的字典序索引。通过对字符串的编码,可以用后缀数组进行字符串集序列的处理。近年来,在基因匹配、文本处理等领域中,后缀数组倍受关注。文献[7]给出了一个利用 $O(N)$ 的额外空间,在 $O(N\lg N)$ 时间内同时构造出后缀数组及最长公共前缀信息数组 LCP 的算法[7,8]。

定义 1 后缀数组(suffix array) 令 $S[1..n]$ 为一有序字符集 $\sum$ 上的字符串,$|S| = n$,$|S|$ 表示 S 的长度。$S[i]$ 表示 S 中位置 i 上的字符。令 $S[i..n]$ 为 S 的第 i 个后缀,简记为 S_i。

收稿日期:2007 - 07 - 11;**修回日期**:2007 - 09 - 11。
基金项目:江苏省自然科学基金资助项目(BK2006095);教育部高等学校博士学科点科研基金资助项目(20040286009)。
作者简介:蒋健安(1982 -),男(回族),江苏南京人, 硕士研究生,主要研究方向:信息安全、数据库、知识发现; 陆介平(1959 -),男,江苏镇江人,教授,博士,主要研究方向:数据库、知识发现; 倪巍伟(1979 -),男,江苏淮安人,讲师,博士,主要研究方向:信息安全、数据库、知识发现; 孙志挥(1941 -),男,江苏南通人,教授,主要研究方向:数据库、知识发现。

如:S = 文本挖掘,其后缀数组为s_1 = 文本挖掘,s_2 = 本挖掘,s_3 = 挖掘,s_4 = 掘。

定义 2 设两个字符串 $s_1 = "a_1a_2\cdots a_m"$ 和 $s_2 = "b_1b_2\cdots b_n"$,称"$a_1a_2\cdots a_p$"为s_1和s_2的最长公共前缀(LCP),如果$p \leqslant \min(m, n)$,且$a_i = b_i(1 \leqslant i \leqslant p)$且$a_{p+1} ! = b_{p+1}$。记为$LCP(s_1, s_2)$。

基于后缀数组的领域词汇抽取算法:

第1步:读入一组某一类别的专利文献文本,每一篇专利文献包含专利名称、摘要、主权项。输入的文本集看成是一系列的字符串,这些字符串使用标点符号隔开。对于每一个字符串,根据此字符串生成一个后缀数组s,对于后缀数组s的每一个子字符串,将其放入HashMap中进行字符串的词频统计。HashMap为(Key, Value)结构,用于存放字符串和字符串在该类专利文献文本集中出现的次数,这里Key存放字符串,Value存放字符串出现的次数,Value值至少为1。

第2步:HashMap中的所有字符串按照字典顺序排序,并计算出相邻字符串的最长公共前缀LCP。

第3步:根据相邻字符串的最长公共前缀统计候选词及其词频。

第4步:去除低频候选词。设置候选次词频阈值,如20,当统计出的候选词次数小于阈值时,则将其去除。

第5步:去除含噪音字的候选词,中文里有一些字出现频率很高,但构词能力却很差,如"的","很"等。从抽词角度看,它们很难构成词的一部分,所以被称为噪音字[9]。噪音字对抽词效率和结果的准确性有较大的负面影响,需要将其过滤掉。我们参照《实用现代汉语语法》[10]和《现代汉语语料库文本分词规范》等文献资料[9],将这些噪音字归纳出来,主要有:

[我你他她们某该各每这那什哪么谁年月日时分秒几多来在就又很的呢吧吗了么嘛哇儿哼啊嗯是着都不和说也看把还个有小到一得地为中于对会之第此或]。当候选词条中含有以上噪音字时,则将该词条去除。

第6步:对统计出的候选词进行筛选,在候选词集中,假如有两个字符串W_1和W_2,W_1是W_2的字串,这里W_1和W_2可能都是词,也可能W_1是词而W_2不是词,还可能W_1不是词而W_2是词,令$con(w_1, w_2) = TF(w_2)/TF(w_1)$表示$W_2$相对于$W_1$的置信度,其中$TF(W_1)$和$TF(W_2)$分别为$W_1$和$W_2$的词频,设置信度上限和下限,若$con$高于上限,则$W_1$可以去除,若低于下限,则$W_2$可以去除。

1.2 文本的表示

为了方便计算机处理,文本文档必须要有一种有效的表示方法。目前,在信息处理领域,向量空间模型(VSM)是应用较多且效果较好的表示方法之一[11]。采用向量空间模型,每篇文档被表示成形如$\boldsymbol{d} = \langle t_1, w_1; t_2, w_2; \cdots; t_n, w_n\rangle$的向量,其中$t_i$是词条项,$w_i$是$t_i$在文档$\boldsymbol{d}$中的权值。因此,所有的$n$维词条向量组成一个文档向量空间。$T_i$可以是一个词、词组或短语,权值$w_i$表示体在文档$\boldsymbol{d}$中的重要程度,目前普遍采用的权重计算方法是TFIDF公式[12]。TFIDF虽然在一定程度上体现了词的区分程度,但是该方法并没有考虑词的位置对文档的区分度。我们知道对于文本,不同位置的词对文档的区分度是不同的。结合专利文献数据,有意义的文本内容是标题,摘要,主权项,而标题中出现的词往往更具代表性。经典的权重计算公式,把处于文档中不同部分的词同等对待,而没有考虑到文档位置信息对该文本的区分度的差异。因此,本文提出了一种考虑位置信息进行加权来计算特征词权重的方法,这样可以更加突出重要的词汇,能更好地代表文本实际包含的内容。相关定义如下:

定义 3 位置权重。给定文档$D(P_1, P_2, \cdots, P_m)$,其中$P_i(i = 1, \cdots, m)$表示组成文档D的子部分,为每个子部分赋予一个权值pw_i,用于体现该部分的词对文档的贡献程度,我们称w_i为文档D的P_i部分的位置权重,这里$0 < pw_i < 1(i = 1, \cdots, m)$,$\sum_{i=1}^{m} pw_i = 1$

定义 4 带位置权重的特征词词频。给定文档$D(P_1, P_2, \cdots, P_m)$中的任意一个词t,则该词在文档D中的带位置权重的词频为$ptf(t, D) = \sum_{i=1}^{m} tf_i pw_i = 1$,这里$pw_i$表示$D$的$P_i$子部分的位置权重,$tf_i$表示词$t$在$P_i$子部分出现的次数,若$t$在子部分$P_i$没有出现,则$tf_i = 0$。

本文中将专利文献信息定义为三元组$D = D(P_1, P_2, P_3)$,其中P_1表示标题,P_2表示摘要,P_3表示主权项,且$pw_1 > pw_2 > pw_3$。

引入位置权重后,对于文档D中的任意一个词,则该词的权重计算公式为:

$$w_i = \frac{ptf(t_i, D) \times \ln(N/n_i + 0.01)}{\sqrt{\sum_{j=1}^{n} [ptf(t_{ij}, D)]^2 [\ln(N/n_{ij} + 0.01)]^2}}$$

词频tf:词在文档中出现的出线的频率。

词的倒排文档频率idf:该词在文档集合中分布情况的量化,通常的计算方法是$\ln[(N/n_i) + 0.01]$,其中N为文档集中的文档总数,n_i为出现该词的文档数,称为该词的文献频数。

归一化因子:为降低个别高频特征词对其他低频词的抑制作用,对各分量进行标准化。

1.3 特征提取

文本分类中文档向量的维数往往很多,造成大多数分类器的学习算法非常低效,而且一些通用的、各个类别都包含的词条对分类的贡献很小。因此,迫切需要能够在不影响分类精度的情况下降低维度,特征提取的功能便是取出那些表现力不强的词条,筛选出具有代表性的特征项词条。本文采用词和类别的互信息量(MI)作为特征提取的标准。

其表达式如下:

$$MI(t, c) = \ln \frac{P(t,c)}{P(c)P(t)}$$

$$MI(t) = \sum_{c \in C} P(c) MI(t,c)$$

其中:

$P(c)$为从文本集合中随机选取的文本属于类c的概率;

$P(t)$为包含特征项t的文本在整个类别集合中出现的概率;

$P(t, c)$为类c中包含特征项t的文本在整个类别集合中出现的概率。

2 面向专利数据的层次分类方法

不同的分类方法各有其优点和局限性,需要根据具体的应用进行选择。本文研究的分类系统面向大规模跨领域的专利文献集,对实时性要求较高,因此在保证较高的分类准确率同时,要尽可能提高分类的效率。专利文献数据有其自身的

特点,在专利文本的分类中,各大类属于不同科学领域,具有较大的区分度,简单方法即可加以区分,而专利文本的小类,属于同一个科学领域的不同研究方向,具有比较高的相似度,需要较高精度的分类方法。

研究者普遍认为 KNN 方法是分类准确性最好的方法之一,但 KNN 方法存在时间复杂度过高的问题,特别是在训练样本规模较大时,其分类速度很慢。对属于不同科学领域的 IPC 大类,属于不同科学领域,专利具有很好的区分度,可以采用简单的方法加以区分;对于一个大类中的不同小类属于同一个科学领域的不同研究方向的专利,具有比较高的相似度,因此需要使用高精度的分类算法 KNN 加以区分。本系统结合专利文献数据的类别特点,设计了层次化的分类方法:将专利的类别分为两部分,即大类和小类。首先用平均向量法将带分类文本划分到某一个大类中去,再在这个大类中使用 KNN 算法将该文本划分为该大类中的某一个小类,这样就完成了专利文本的分类工作。

平均向量算法需要计算各大类的向量权值,首先使用特征词抽取算法获得一组特征词,对于所抽取的每一个特征词,取训练集中属于该大类的所有文档的这一特征词向量权值的算术平均值作为该特征词在类别特征向量中的权值。这样对于每一个大类都可以构造出一个类别特征向量 $\boldsymbol{c}_i(t_1,\overline{w_1};t_2,\overline{w_2};\cdots;t_m,\overline{w_m})$ 来代表这一个大类。同样用这些特征词构造待分类专利文本的特征向量 $\boldsymbol{d}(t_1,w_1;t_2,w_2;\cdots;t_n,w_n)$。计算 $\boldsymbol{d}$ 与 $\boldsymbol{c}_i$ 的相似度,选择相似度最大的分类作为该待分类专利的大类,然后使用 KNN 方法确定该专利的小类。

算法采用向量间余弦夹角来度量文档间的相似程度,假设待分类文档向量为 $\boldsymbol{d}_i$,类别向量为 $\boldsymbol{c}_j$,其相似度为:

$$\mathrm{sim}(\boldsymbol{d}_i,\boldsymbol{c}_j)=\cos(\boldsymbol{d}_i,\boldsymbol{c}_j)=\frac{\sum_{k=1}^{m}w_{ik}\times w_{jk}}{\sqrt{\left(\sum_{k=1}^{m}w_{ik}^2\right)\cdot\left(\sum_{k=1}^{m}w_{jk}^2\right)}}$$

分类算法流程见图 1,算法描述如下:

阶段一　训练

输入:训练文本集

输出:代表大类的特征向量

第 1 步:对训练文本集使用后缀数组算法抽取领域词汇,并和事先准备的通用词典合并生成分词词典。

第 2 步:对训练文本集中的文档使用分词词典分词,然后统计词频,每篇文档生成一个向量 $\boldsymbol{d}$。

第 3 步:计算向量 $\boldsymbol{d}$ 中每个词条的互信息量(mutual information),进行排序,按照排序结果进行特征词的降维处理,这里设定一个阈值,例如 500,选择排序结果前 500 位的词作为降维后的特征词。

第 4 步:根据改进的 TFIDF 公式计算每个降维后的特征词的权重 w_i,生成特征向量表,每篇文档表示为向量$(t_1,w_1;t_2,w_2;\cdots;t_n,w_n)$,$t_i$ 为特征项词条,w_i 为对应的权值。

第 5 步:对于每一大类中的特征项词条 t_i,计算其在训练集中所有文档特征向量中权值的算术平均值$\overline{w_i}$,作为该词条在类别特征向量中的权值,构造大类特征向量 $\boldsymbol{c}_i(t_1,\overline{w_1};t_2,\overline{w_2};\cdots;t_m,\overline{w_m})$。

阶段二　分类

输入:待分类文档(测试文本集)

输出:文档所属类别(小类)

第 1 步:对测试文本集使用分词词典分词,并使用阶段一中降维后的特征词进行特征表示,生成特征向量表,每篇文档表示为向量$(t_1,w_1;t_2,w_2;\cdots;t_n,w_n)$,$t_i$ 为特征项词条,w_i 为对应的权值。

第 2 步:计算测试文本与每个大类特征向量的相似度,确定该专利文本所属大类。

第 3 步:对确定大类的文本与属于该大类的训练文本计算相似度,确定该专利文本所属小类。

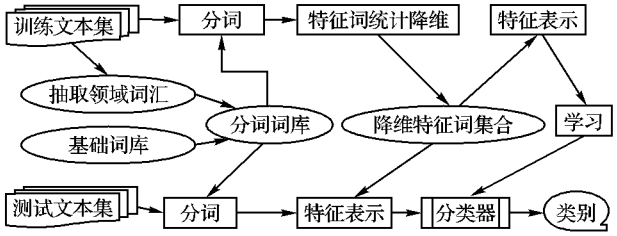

图 1　专利文献自动分类过程

3　评估方法与实验结果

分类系统的评价通常有两个重要指标:查全率(recall)和查准率(precision)。

其中:

$$\text{查准率}=\frac{\text{程序正确分到某类的文档数}}{\text{程序实际分到某类的文档数}}$$

$$\text{查全率}=\frac{\text{程序正确分到某类的文档数}}{\text{实际应分到某类的文档数}}$$

本系统针对中国专利数据库的专利文档进行分类,选取了 3 大类 6 个小类共 1 500 篇专利文档进行训练和测试。对于每个分类分别取出 200 篇作为训练集,50 篇作为测试集。测试结果见表 1 和表 2。

表 1　测试结果表(大类)

大类名称	查准率/%	查全率/%
A61	100.0	90.0
F01	92.0	100.0
H04	96.2	97.5

表 2 测试结果表(小类)

小类名称	查准率/%	查全率/%
A61L	83.3	77.5
A61K	91.4	80.0
F01C	77.8	87.5
F01M	78.8	82.5
H04Q	95.1	97.5
H04M	90.0	90.0

由测试结果可知,本系统无论是大类或小类均达到了较好的分类效果,同时相比于直接使用 KNN 方法,缩小了比较的范围,大幅提高了效率。

4　结语

本文讨论了面向专利数据的文本自动分类系统的关键技术,并结合专利数据的特点对无词典分词和权重计算进行了改进,重点描述了一个面向专利文献数据的文本分类系统的实现框架及改进算法。理论分析和测试结果表明系统具有较好的分类效果,所提改进方法是有效可行的。下一步的工作将在本系统的基础上,深入地结合机器学习、自然语言处理等理论知识,尝试其他分类算法,进一步提高分类效率和分类精度。

(下转第 167 页)

50%, K = 70% 时的 C、M、Y 在各个阶调的 δ 的取值,从而得到一般规律。方程加入修正系数 δ 后检验方程精度,发现检验色块色差小于 6 的比例达到 94.4%,且平均色差小于 4,较好地达到了印刷忠实复制的要求。

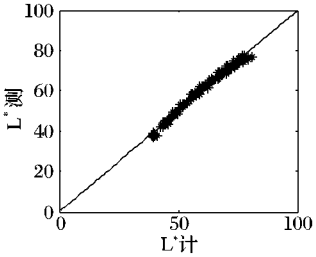

图 3　系数加入 δ 修正后实测 L* 值与计算 L* 值的对比

4　CMYK 到 L*a*b* 方程的建立

通过对实验样本的色度值进行处理与分析,可以得到 K 等于 0,10%, 30%, 50%, 70% 时 C、M、Y 从 0 到 100% 各阶调对应的平面系数,从而回归出各典型 K 值下所对应的各平面方程系数与相应色网点面积值的二次表达式。如:$D_{ci} = P_{ac} \times C^2 + Q_{ac} \times C + R_{ac}$。采取三次样条曲线[5] 拟和可以确定四色平面方程系数的二次项 P_{ac}、一次项 Q_{ac}、常数项 R_{ac} 与 K 网点面积率之间的关系,C 平面方程系数 D_{ci} 的各项系数与 K 关系图如图 4 所示。

由此,任意颜色网点组合 C_i、M_i、Y_i、K_i 所对应的平面方程系数 D、E、F 的各项系数均可表示为 K 的函数。将各系数表达式带入式(2) 即可得到 CMYK 到 L*a*b* 的转换方程。这样,根据已知的四色网点百分比即可得到对应的 L*、a*、b* 值。

5　结语

颜色建模是建立颜色或染料与它们的混合色之间的光谱关系或色度关系,印刷品、显示图像和相片等颜色复制的结果都与原色之间有一定的数学关系。基于四色印刷网点呈色平面规律研究 CMYK 与 L*a*b* 的转换算法并建模,其研究基础来源于彩色网点呈色的光学现象,有很好的应用基础,该模型研究的思考路径有可能成为色彩空间转换技术研究的新的生长点。应用本文研究思路采用国际标准数据建模并进一步修正模型精度,再利用印刷复制的灰平衡理论或者遗传神经算法等其他方法完成逆向转换,结合颜色视觉、色彩学理论探讨彩色网点平面呈色机理,由此完善彩色印刷网点呈色平面理论并创建网点面积率与颜色色度值之间的转换算法,可为彩色图像分色、色彩管理、计算机配色、图像画面直接检测和控制开创新的技术路线和研究方法。

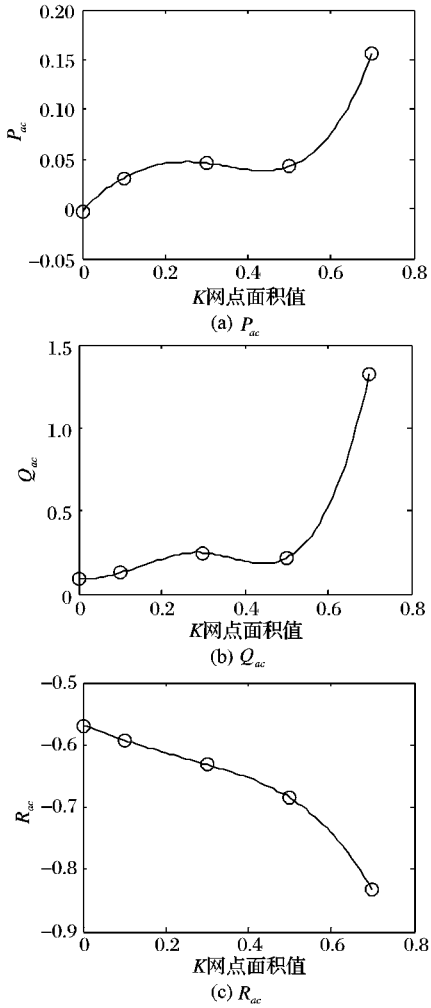

图 4　C 平面方程系数 D_{ci} 的各项系数与 K 关系

参考文献:

[1] KAJI M, YAZUMA. Y, NONAKA. M. Some colorimetric properties included in the color characterization data of process prints[C]// TAGA Proceedings. New York: [s. n.], 1998: 226 – 241.

[2] 徐敏, 徐锦林. 彩色印刷图像平面呈色模型的验证与应用研究[J]. 包装工程, 2004(2): 34 – 38.

[3] 曹从军, 周明全, 徐锦林. L*a*b* 到 CMY 转换方程的研究[J]. 仪器仪表学报, 2004, 25(4): 129 – 132.

[4] 许波, 刘征. MatLab 工程数学应用[M]. 北京: 清华大学出版社, 2000.

[5] 闵涛, 秦新强, 赵凤群. 数值分析[M]. 北京: 中国科学文化出版社, 2003: 40 – 42, 60 – 69.

(上接第 161 页)

参考文献:

[1] CAMUS C, BRANCALEON R. Intellectual assets management: from patents to knowledge[J]. World Patent Information, 2003(25): 155 – 159.

[2] FATTORI M, PEDRAZZI G, TURRA R. Text mining applied to patent mapping: a practical business case[J]. World Patent Information, 2003(25): 355 – 342.

[3] 暴海龙, 李金林. 专利检索中的 IPC 和主题词识别方法研究[J]. 北京理工大学学报: 社会科学版, 2004, 5(5): 74 – 76.

[4] JOACHIM T. Text categorization with support vector machines: learning with many relevant features[R]. Dortmund: Technical Report 23, 1997.

[5] FRIEDMAN N, GEIGER D. Bayesian network classifiers[J]. Machine Learning, 1997, 29(2/3): 131 – 163.

[6] JOACHIMS T A. Probabilistic analysis of the rocchio algorithm with TFIDF for text categorization[C]// Proceedings of the 14th International Conference on Machine Learning ICML97. Nashville: [s. n.], 1997: 143 – 151.

[7] UDI M, MYERS G. Suffix arrays: a new method for on-line string searches[J]. SIAM Journal on Computing, 1993, 22(5): 935 – 948.

[8] YAMAMOTO M, CHURCH K. Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus[J]. Association for Computational Linguistics, 2000, 27(1): 1 – 30.

[9] 韩洁, 周勇, 刘少辉. 基于 WWW 的未登录词识别研究[J]. 计算机科学, 2002, 29(12): 155 – 156.

[10] 刘月华, 潘文娱. 实用现代汉语语法[M]. 北京: 外语教学与研究出版社, 1983.

[11] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing[J]. Communication of the ACM, 1995(1): 2 – 8.

[12] SEBASTIANI F. Machine learning in automated text categorization[J]. ACM computing surveys, 2002, 34(1): 11 – 12, 32 – 33.