中 文 信 息 学 报
第 16 卷 第 3 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 16 No. 3

一种基于向量空间模型的多层次文本分类方法*

刘少辉 董明楷 张海俊 李 蓉 史忠植

(中国科学院计算技术研究所智能信息处理重点实验室 北京 100080)

摘要:本文研究和改进了经典的向量空间模型(VSM)的词语权重计算方法,并在此基础上提出了一种基于向量空间模型的多层次文本分类方法。也就是把各类按照一定的层次关系组织成树状结构,并将一个类中的所有训练文档合并为一个类文档,在提取各类模型时只在同层同一结点下的类文档之间进行比较;而对文档进行自动分类时,首先从根结点开始找到对应的大类,然后递归往下直到找到对应的叶子子类。实验和实际系统表明,该方法具有较高的正确率和召回率。

关键词:文本分类;向量空间模型;信息增益;特征提取

中图分类号:TP391.1

An Approach of Multi-hierarchy Text Classification Based on Vector Space Model

LIU Shao-hui DONG Ming-kai ZHANG Hai-jun LI Rong SHI Zhong-zhi

(Laboratory of Intelligent Information Processing, Institute of Computing Technology,
Chinese Academy of Sciences Beijing 100080)

Abstract: This paper does research and improves on the classical approach of calculating the term weight in Vector Space Model. Furthermore, an approach of multi-hierarchy text classification based on Vector Space Model is proposed. In this approach, all classes are organized as a tree according to some given hierarchical relations, and all the training documents in a class are combined into a class-document. In order to construct the class models, it is just only to compare among the class-documents attached to the same node of the same layer. When it is going to classify the documents, one matching process is hierarchically performed from the root node to the leaf nodes until a corresponding subclass is found. The experiment and real systems indicate that the approach is of high classification Precision and Recall.

Keywords: Text Classification; Vector Space Model; Information Gain; Feature Selection

一、引言

随着信息技术的发展,特别是 Internet 应用的普及,人们已经从信息缺乏的时代过渡到信息极为丰富的时代。如何从大量信息中迅速有效地提取出所需信息也就成为一项重要的研究课题。由于分类可以在较大程度上解决目前网上信息杂乱的现象,方便用户准确地定位所需的信息,因此分类尤其是文本分类的研究变得越来越重要[1,11]。

文本分类的目标是在分析文本内容的基础上给文本分配一个或多个比较合适的类别。目

* 收稿日期:2001-11-8
本文得到国家自然科学基金(60173017)和北京自然科学基金(4011003)支持
作者刘少辉,男,1977 年生,博士研究生,主要研究方向为数据挖掘、信息检索.董明楷,男,1973 年生,博士研究生,主要研究方向为智能主体、描述逻辑.张海俊,男,1980 年生,硕士研究生,主要研究方向为智能主体、软件工程.李蓉,女,1973 年生,硕士研究生,主要研究方向为神经网络.史忠植,男,1941 年生,研究员,博士生导师,主要研究方向为人工智能、知识工程.

前已经有许多机器学习方法应用到该领域:Vapnik 提出的支持向量机(SVM)[2];在文本分类研究一开始就引起关注的 K 近邻(KNN)分类器[3];Yang 提出的一种线性最小二乘方拟合法(LLSF)[4];Apte 采用决策树方法进行分类[5]。另外,神经网络(Nnet)和贝叶斯[6]方法也被广泛地应用到文本分类中。

上述大多数方法都采用了经典的向量空间模型(VSM)。在该模型中,文档的内容被形式化为多维空间中的一个点,以向量的形式给出,然后通过计算向量间的距离决定向量类别的归属。而在向量空间模型中,经典的词语权重计算方法将词频和倒排文档频率结合起来(称为 tf.idf 方法)[7]。本文对 tf.idf 方法进行了分析并做了改进,使之更加合理。另外,在此改进的基础上,本文提出了一种多层次文本分类的方法。也就是把各类按照一定的层次关系组织成树状结构,并将一个文档类中的所有训练文档合并为一个类文档,在提取各类模型时只在同层同一结点下的类训练文档间进行比较;而对文档进行自动分类时,首先从根结点开始找到对应的大类,然后递归往下直到找到对应的叶子子类。实验和实际系统表明:该方法是行之有效的,具有较高的分类正确率与召回率。

本文组织如下:第 2 节介绍对词语权重计算方法的研究和改进;第 3 节给出多层次文本分类的实现算法;第 4 节列出实验结果和分析;第 5 节给出结论。

二、词语权重的计算

在 VSM 中,每一篇文档都被映射成多维向量空间中的一个点,对于所有的文档类和未知文档,都可用此空间中的向量(T_1, W_1; T_2, W_2; …; T_m, W_m)来表示(其中 T_i 为词,W_i 为词对应的权值,用以刻画该词在描述此文档内容时的重要程度),从而将文档信息的表示和匹配问题转化为向量空间中向量的表示和匹配问题来处理。

对于词权重的计算,经典的 tf.idf 方法考虑两个因素:1) 词语频率 tf(term frequency):词语在文档中出现的次数;2) 词语倒排文档频率 idf(inverse document frequency):该词语在文档集合中分布情况的一种量化,常用的计算方法是 $\log_2(N/n_k + 0.01)$,其中 N 为文档集合中的文档数目,n_k 为出现该词语的文章数。

根据以上两个因素,可以得出公式:$W_{ik} = tf_{ik} \times \log_2(N/n_k + 0.01)$,其中 tf_{ik} 为词语 T_k 在文档 D_i 中出现的次数,W_{ik} 为词语 T_k 在文档 D_i 中的权值,$k = 1, 2, \cdots, m$(m 为词的个数)。

为了计算方便,通常要对向量进行归一化,最后有:

$$W_{ik} = \frac{tf_{ik} \times \log_2(N/n_k + 0.01)}{\sqrt{\sum_{k=1}^{m}(tf_{ik} \times \log_2(N/n_k + 0.01))^2}} \quad (1)$$

以上公式的提出是基于这样一个考虑:对区别文档最有意义的特征词应该是那些在文档中出现频率足够高而在文档集合中的其它文档中出现频率足够少的词语。

由于 W_{ik} 对特征词的选择起决定性的作用,有很多学者对此做了深入研究。Rocchio 提出了基于正例反例的批处理计算方法[13];Widrow 提出了即时的线性计算方法[14];李国臣等提出了基于对数似然比测试的词权重计算方法[9];张月杰等结合了词在文档中的位置,如标题、正文[15]。

我们所采用的方法是经典的 tf.idf 方法的改进,结合了词的信息量。也就是把文档集合 D 看成一个符合某种概率分布的信息源,依靠文档集合的信息熵和文档中词语的条件熵之间信息量的增益关系确定该词语在文本分类中所能提供的信息量,即词语在分类中的重要程度,然后将该信息量综合到权重计算公式中。这样就弥补了 tf.idf 方法虽然考虑了词语在文档集

合中的分布情况,但是并没有考虑分布的比例情况的缺陷。具体的计算公式如下所示:

$$W_{ik} = \frac{tf_{ik} \times \log_2(N/n_k + 0.01) \times IG_k}{\sqrt{\sum_{k=1}^{m}(tf_{ik})^2 \times [\log_2(N/n_k + 0.01) \times IG_k]^2}} \quad (2)$$

IG_k 为词语 T_k 的信息量,其计算公式为: $IG_k = H(D) - (H(D|T_k)$,其中文档集合 D 的信息熵, $H(D) = -\sum_{d_i \in D}(P(d_i) \times \log_2 P(d_i))$,词语 T_k 的条件熵 $H(D|T_k) = -\sum_{d_i \in D}(P(d_i|T_k) \times \log_2 P(d_i|T_k))$ 。至于如何定义文档 d_i 的概率,文[8]曾提到令 $P(d_i) = \frac{|wordset(d_i)|}{\sum_{d_j \in D}|wordset(d_j)|}$, $|wordset(d_i)|$ 表示文档 d_i 中不同词语的个数。但是若采用该公式,则在计算文档的概率时只考虑到词语的数目,两篇词语相同但是词语频率不相同的文档都会被认为概率相同。为了能进一步更加准确地反映文档分布的比例情况,而且兼顾到在tf.idf方法中词频是核心因素,我们定义 $P(d_i)$ 为:

$$P(d_i) = \frac{wordfreq(d_i)}{\sum_{d_j} wordfreq(d_j)} \quad (3)$$

$wordfreq(d_i)$ 表示文档 d_i 中所有词的词频之和。改进后的效果将在第 4 节详细给出。

三、多层次分类算法

一般的分类方法都采用全部类别共享一个分类器或者每个类别设置一个分类器的方法,又称为单分类器方法或者多分类器方法,而且这些方法中的类别都是在同一个层次,即处于同一个平面类空间上。当类别的个数较多时,提取模型的时间耗费巨大,而且对新文档进行分类的时候要和全部类模型进行比较,以便给该文档分配合适的类别。

针对以上不足,我们提出一种基于向量空间模型的多层次文本分类方法,也就是将全部类别按照一定的层次关系组织成一个树状结构。该方法的提出是基于这样一个考虑:属于同一结点下的各类肯定比不属于同一结点下的各类更有共性,比如足球、篮球、羽毛球之间的共性肯定比足球、软件、音乐之间的共性多。

正是基于以上的考虑,我们把分类任务划分成更小的与类层次结构对应的分类子问题。比如,存在一个区分体育和电脑网络的分类器,另外还存在一个仅用于体育类的分类器,用来区分足球、篮球、羽毛球。每一个子任务显然比原来的任务更加简单,因为在树结构中每个结点的分类器只需要在少部分类中区分,而且由于这部分类的共性较多,这样各类模型中所包含的特征词也比较少。

3.1 构建类模型

通过对给定的经过人工按照类层次结构进行分类后的文档集合进行训练,并经过特征选取(特征词和权值的选取)就可以构建对应的类模型,为自动分类提供基础。在构造各类模型的算法中,每个模型由向量表示,包括该类的特征词和对应的权值。

在特征词的选取中,我们综合考虑了频度和集中度两种因素。考虑频度因素的特征词选择方法认为,在某一类文档中出现次数越多的词越能够代表这类文档;考虑集中度因素的特征词选择方法认为,某类的特征词应该集中出现在该类的文档中,而不是均匀地分布在各类文档中[9]。另外,在实际应用中组成某个类的模型的特征项的个数也不易过多,可以只保留权值较高(超过某权值阈值)的项,否则会大大降低系统的处理速度[10]。

在我们的算法中,每一个文档类中的所有训练文档都合并为一个类文档以进行文档类的特征词的提取,至于权重的计算则采用第 2 节所提到的公式(2)。为了进一步提高模型的代表性,我们的算法在考虑权值的基础上,也考虑了以上提到的词频、词的集中度因素。详细算法 CCM(Create Class Model)如下:

输入:人工确定的各类之间的层次关系,实际上是一树状结构,每一结点(除了根结点)代表一个类,各训练文档都被人工分在叶子结点对应的子类中

输出:各类对应的类模型,以文本文件的方式存储

步骤:

对叶子结点到根结点每一层的所有结点自下而上进行处理:

1. 若该结点 Node 是叶子结点,则统计该结点对应的类文档中的词频信息,包括各词的词频统计、总词数和总词频的统计

2. 若该结点 Node 是非叶子结点,假设该结点有 $V_1, V_2, \cdots, V_t$ 共 t 个子结点,对应的文档类中有 $T_1, T_2, \cdots, T_s$ 共 s 个词

1) 根据公式(3)计算结点 V_i 对应的类文档 d_i 出现的概率 $p(d_i)$,其中 $i = 1, 2, \cdots, t$

2) 计算 $H(D)$ 和 $H(D|T_k)$,得到 IG_k,其中 $k = 1, 2, \cdots, s$

3) 提取结点 V_i 对应的类模型 C_i, $i = 1, 2, \cdots, t$

a) 初始化类模型 C_i 为空

b) 根据公式(2)计算词 T_k 的权值 W_{ik}, $k = 1, 2, \cdots, s$

c) 对各个词按照权值从大到小的顺序重新排列成 $T_1, T_2, \cdots, T_s$

d) 依次对从 T_1 到 T_s 的各词进行判断:若类模型 C_i 中的特征词个数已经达到阈值 NUM_T,则该类模型 C_i 提取结束;否则若词 T_k 的权值超过一定的阈值 α、词频超过一定的阈值 β、词的集中度超过一定的阈值 γ,而且该词不在事先已设定的停用词表中,则该词可以作为该类的特征词,和权值一起加入类模型中。

不难看出,此算法提取各类模型时只在同层同一结点下的类训练文档间进行比较,这样不仅提取类模型的时间相对减少,而且由于同层同一结点下的类之间的共性相对较多,则各类模型中的特征词个数比较少,使得文档分类的耗时也相应减少。

3.2 文档自动分类

自动分类就是通过计算机对大量的新文档进行分类。我们首先要把这些文档用归一化后的向量来表示,包括该文档中的词和该词在文档中的权重,文档中词权重的计算主要考虑词频、词的位置等因素;然后对该文档向量根据类层次结构从上往下逐层和各类模型匹配,即计算它们的相似程度,直到找到与叶子结点对应的合适的子类。详细算法 TC(Text Classification)如下:

输入:人工确定的各类之间的层次关系,实际上是一树状结构(与算法 CCM 的相同),每一结点(除了根结点)代表一个类;根据算法 CCM 得到的各类模型;待分类的文档

输出:分配好合适类别的文档

步骤 1:对待分类的文档进行向量化,设对应的向量为 $d = (T_1, W_1; T_2, W_2; \cdots; T_m, W_m)$

步骤 2:结点 $X \leftarrow$ 根结点,若结点 X 是非叶子结点,反复执行:

1) 设结点 X 有 $V_1, V_2, \cdots, V_t$ 共 t 个子结点,结点 V_i 对应的类模型为 $C_i = (t_1, w_1; t_2, w_2; \cdots, t_n, w_n)$, $i = 1, 2, \cdots, t$

2) 依次计算文档和各类之间的相似度,其中相似度 $Sim(C_i, d)$ 用向量 d 和 C_i 之间

的夹角来度量:

a) $Sim(C_i, d) = 0$

b) 若 $m \geq n$,

i) 则对 C_i 中的每个词 $t_u(u = 1, 2, \cdots, n)$ 在向量 d 中查找是否存在,若存在,对应的词为 T_v,则 $Sim(C_i, d) = Sim(C_i, d) + W_v \times w_u$

ii) 否则对 d 中的每个词 $T_v(v = 1, 2, \cdots, m)$ 在向量 C_i 中查找是否存在,若存在,对应的词为 t_u,则 $Sim(C_i, d) = Sim(C_i, d) + W_v \times w_u$

c) 找到和向量 d 相似度最大的类 C_{max},即 $Sim(C_{max}, d) = Max_{i \in [1, t]}(Sim(C_i, d))$,结点 $X \leftarrow$ 结点 V_{max}

步骤 3:结点 X 对应的类就是自动分给文档的类

四、实验结果及分析

我们选取的实验数据有两种,首先是国际通用的文本分类标准测试集 Reuters - 21578,它包含了路透社 1987 年的新闻,全部文档使用英语,并经路透社人工汇集和分类。该测试集本身没有预先定义类层次结构,我们选取了 8 个类别共 471 篇文档作为数据集一,其中 362 篇作为训练集,109 篇作为测试集,并根据这些类别的特点人为设定了类层次结构,如图 1 所示。

此外,我们实验室为某著名 IT 公司开发了关于文本分类方面的系统,主要思路是:按照公司的需要构建具有树状结构的分类体系,通过对给定的经过人工按照类层次结构进行分类后的文档集合进行训练,得到各个类对应的类模型;然后对网络蜘蛛(Spider)抓回来的网页信息提取相应的文档内容,并且对该内容进行自动分词(包括正反向最大匹配切词、歧义处理、姓名识别,音译名识别、地名识别),再根据自动分类算法将该网页分类。经过近半年的试运行和多次的修正,已取得了较为理想的分类效果。在测试阶段,我们从新浪、FM365、网易等网站下载的网页中整理出 11000 篇文档作为数据集二,其中的 9360 篇作为训练集,1740 篇作为测试集。这些文档对应的类结构如图 2 所示。

整个实验分成两部分:第一部分比较所有的类在同一层次和在不同层次下分类的准确率与召回率;第二部分比较多层次分类方法下不同的权重计算方法对应的分类的准确率和召回率。数据集一对应的实验结果分别见表 1 和表 2、数据集二对应的结果分别见表 3 和表 4。

由表 1 和表 3 可以看出,在对少数某些类别如 hog、冰雪运动等类的文档进行分类时多层结构下的召回率或者准确率比单层结构下的要低,这是因为在每一层进行模型匹配以选择最适合的类别时都会有一定的误差,层数越往下总误差就会越大。但是从总体看,多层结构下分类的召回率和准确率都要优于单层结构的,尤其是对于那些相对全部类来说特征比较模糊的类别提高的效果就比较明显,如篮球其它、体育其它等类。另外,由表 1 和表 3 的对比分析,文档类别的数目越多,定义的类层次结构越合理,使用多层次分类方法获得的效果就越好。

分析表 2 和表 4 的实验结果,可以看到,改进后的 tf.idf 方法总体上优于 tf.idf 方法。然而应该指出,尽管改进后的权重计算方法使得分类的召回率和准确率有所改善,但是针对整个文本分类问题的效果仍未出现明显提高。正如 Yang 在文[3]中所述:文本分类问题是涉及到文本表示、相似度计算和算法决策等多种复杂技术的综合应用。

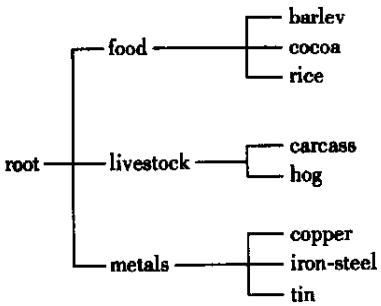

图 1　数据集一的类层次结构

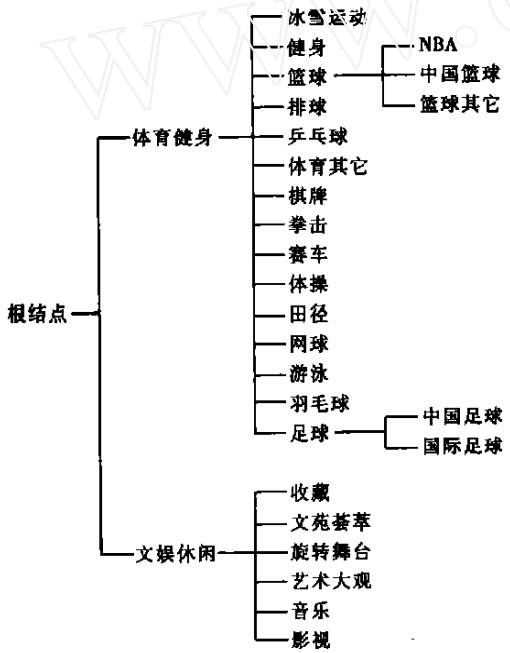

图 2　数据集二的类层次结构

表 1　数据集一对应的单层和多层类结构下的分类的召回率和准确率

类别	单层		多层	
	召回率 %	准确率 %	召回率 %	准确率 %
barley	92.6	76.9	92.6	80.3
cocoa	88.2	87.7	89.4	92.6
rice	95.5	91.8	97.0	95.6
carcass	89.3	91.5	94.7	94.7
hog	85.2	95.8	73.0	98.1
copper	98.7	100	98.7	98.7
iron-steel	92.5	95.1	95.5	96.2
tin	90.9	90.9	90.9	93.8

表 2　数据集一对应的多层类结构下不同的权重计算方法得到的分类的召回率和准确率

类别	tf.idf 方法		改进后的 tf.idf 方法	
	召回率 %	准确率 %	召回率 %	准确率 %
barley	88.9	80.0	92.6	80.3
cocoa	89.4	92.6	89.4	92.6
rice	92.5	91.1	97.0	95.6
carcass	90.7	91.9	94.7	94.7
hog	73.0	98.1	73.0	98.1
copper	94.9	94.9	98.7	98.7
iron-steel	92.5	93.1	95.5	96.2
tin	87.9	90.6	90.9	93.8

五、结束语

为了能进一步更加准确地反映文档分布的比例情况,本文对经典的词语权重的计算方法进一步做了改进,经过实验验证,其性能总体上优于传统的方法。

考虑到各类别之间的关系,我们提出了一种基于向量空间模型的多层次文本分类方法,把分类任务划分成更小的与类层次结构对应的分类子问题,在提取各类模型时也只在同层同一结点下的类训练文档间进行比较,以减少计算量,使类模型更加正确,这无疑是对分类方法一种有益的尝试。在此基础上我们开发的有关文本分类的系统运行状况良好,具有速度快、准确度较高等特点。由于类层次结构都是事先人为全部指定,能否通过计算机自动提取这些层次结构供人参考正是我们现在进行的工作。此外,考虑到向量空间模型中的文档表示方法丢失了大量的词语关联信息,如何在文档表示、分类模型提取、分类算法中弥补这些损失也将是我们今后研究的重点。

表 3　数据集二对应的单层和多层类结构下的分类的召回率和准确率

类别	单层		多层	
	召回率 %	准确率 %	召回率 %	准确率 %
冰雪运动	92.1	86.7	95.4	85.3
健身	91.1	90.2	92.3	91.8
NBA	93.9	82.8	94.6	92.2
中国篮球	84.1	79.9	92.5	87.1
篮球其它	11.5	81.8	46.2	92.7
排球	91.5	92.6	90.9	92.8
乒乓球	96.2	96.2	96.2	96.2
体育其它	66.9	81.1	74.5	90.9
棋牌	93.3	92.3	93.3	92.3
拳击	91.7	93.3	94.0	91.8
赛车	88.9	86.7	92.9	92.9
体操	93.6	92.3	91.5	95.7
田径	94.5	92.4	88.5	94.4
网球	83.8	92.9	87.1	87.1
游泳	77.8	89.4	90.6	88.3
羽毛球	93.8	91.4	96.2	94.7
中国足球	92.0	87.5	93.7	87.3
国际足球	90.9	88.7	92.1	88.9
收藏	88.3	75.8	92.6	79.4
文苑荟萃	87.1	81.6	97.1	89.8
旋转舞台	61.8	83.5	77.9	80.3
艺术大观	59.3	81.8	65.8	89.6
音乐	86.7	76.3	92.0	85.9
影视	87.2	85.8	90.5	90.7

表 4　数据集二对应的多层类结构下不同的权重计算方法得到的分类的召回率和准确率

类别	tf.idf 方法		改进后的 tf.idf 方法	
	召回率 %	准确率 %	召回率 %	准确率 %
冰雪运动	95.4	85.3	95.4	85.3
健身	89.8	90.9	92.3	91.8
NBA	95.9	91.7	94.6	92.2
中国篮球	92.5	89.6	92.5	87.1
篮球其它	41.6	89.7	46.2	92.7
排球	90.9	92.8	90.9	92.8
乒乓球	96.2	92.3	96.2	96.2
体育其它	70.2	90.4	74.5	90.9
棋牌	93.3	92.3	93.3	92.3
拳击	94.0	91.8	94.0	91.8
赛车	88.9	92.9	88.9	92.9
体操	91.5	95.7	91.5	95.7
田径	88.5	94.4	88.5	94.4
网球	87.1	87.1	87.1	87.1
游泳	90.6	88.3	90.6	88.3
羽毛球	96.2	94.7	96.2	94.7
中国足球	92.4	88.6	93.7	87.3
国际足球	91.3	87.2	92.1	88.9
收藏	93.6	78.5	93.6	79.4
文苑荟萃	89.5	87.6	90.1	89.8
旋转舞台	75.6	72.5	77.9	80.3
艺术大观	64.1	92.1	65.8	89.6
音乐	89.7	83.1	92.0	85.9
影视	89.3	88.8	90.5	90.7

参 考 文 献

[1] 李晓黎,刘继敏,史忠植. 概念推理网及其在文本分类中的应用. 计算机研究与发展,2000,37(9):1032 - 1038

[2] Vapnik V. The Nature of Statistical Learning Theory. New York, ,Springer-Verlag,1995

[3] Yang Y. Expert network: effective and efficient learning from human decisions in text categorization and retrieval. In Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SIGIR '94),1994,13 - 22

[4] Yang Y. Chute C G. An example-based mapping method for text categorization and retrieval. ACM Transaction on Information Systems (TOIS),1994,12(3):252 - 277

[5] Apte C. Damerau F,and Weiss S. Text mining with decision rules and decision trees. In Proceedings of the Conference on Automated Learning and Discovery,Workshop 6: Learning from Text and the Web,1998

(下转第 26 页)

参 考 文 献

[1] Salton. G, Automatic Text Processing : The Transformation : Analysis and Retrieval of Information by Computer , Addison - Wesley , Reading , Mass , 1989

[2] 李晓黎,刘继敏,史忠植. 概念推理网及其在文本分类中的应用. 计算机研究与发展,37(9) : 1032 - 1038,2000 年 9 月

[3] John M. Picrrc , On the automated classification of web sites , Linkoping Electronic Articles in Computer and Information Science , Vol. 6 , 2001

[4] Thomas Bayer , Ingrid Renz ,Michael Stein , Ulrich Kressel , Domain and language independent feature extraction for statistical text categorization , proceedings of workshop on language engineering for document analysis and recognition - ed. by L. Evett and T. Rose , part of the AISB 1996 Workshop Series , Sussex University ,England ,April 96 ,21 - 32

[5] Antal van den Bosch , Walter Daelemans , Ton Weijters , Morphological analysis as classification: an inductive - learning approach , Proceedings of NEMLAP - 2 , 2 , July , 1996

[6] Manuel de Buenaga Rodriguez , Jose Maria Gomez - llidalgo , Belen Diaz - agudo , Using WORDNET to complement training information in text categorization , Second International Conference on Recent Advances in Natural Language Processing , 1997

[7] Ellen Riloff and Wendy Lehnert , Information Extraction as Basis for High - precision Text Classification , ACM Transactions on Information System ,July 1994 ,12(3)

[8] Zhu Jingbo , Yao Tianshun , FIFA : a simple and effective approach to text topic automatic identification , In: Proceedings of International Conference On Multilingual Information Processing 2002 , Shenyang , China , Feb. 2002 ,207 - 215

[9] 姚天顺等,自然语言理解. 清华大学出版社,1995

(上接第 14 页)

[6] Mitchell T. Machine Learning. McGraw : Hill ,1996

[7] Salton G,Buckley B. Term weighting approaches in automatic text retrieval. Information Processing and Management ,1998 ,24(5) :513-523

[8] 鲁松,李晓黎,白硕等. 文档中词语权重计算方法的改进,中文信息学报,2000 ,14(6) :8 - 13

[9] 李国臣. 文本分类中基于对数似然比测试的特征词选择方法,中文信息学报,1999 ,13(4) :16 - 21

[10] 邹涛,王继成,黄源等. 中文文档自动分类系统的设计与实现,中文信息学报,1999 ,13(3) :26 - 32

[11] 黄萱菁. 大规模中文文本的检索、分类与摘要研究,复旦大学博士学位论文,1998

[12] Yang Y. and Liu X. . A re-examination of text categorization methods. In Proceedings of the 22nd Annual ACM SIGIR Conference on Research and Development in Information Retrieval ,42 - 49 ,1999

[13] Rocchio Jr. . ,J. J. . Relevance feedback in information retrieval. In Salton , G. ,editor , The SMART Retrieval System: Experiments in Automatic Document Processing ,pp. 313-323. Prentice-Hall , Inc. , Englewood Cliffs ,New Jersey ,1971

[14] Widrow B. ,Stearns S. D. . Adaptive Signal Processing. Prentice-Hall , Englewood Cliffs ,NJ ,1979

[15] 张月杰,姚天顺. 基于特征相关性的汉语文本自动分类模型的研究,小型微型计算机系统,1998 ,19(8) :49 - 55