一种基于语义特征的逻辑段落划分方法及应用

朱振方 刘培玉 王金龙
(山东师范大学信息科学与工程学院 济南 250014)

摘 要 引入了一种以逻辑概念为中心的段落化匹配方式。该方法建立在概念词典之上,通过分析待分类文本中所包含的逻辑概念,将待分类文本中表达相同意义的段落进行聚类分析以得到一个逻辑层次,并建立以此逻辑层次划分方法为基础的逻辑段落概念,然后以该逻辑段落作为依据来衡量不同的段落对于文本主题表示的贡献程度。同时,针对匹配过程中存在的多义词和同义词现象,引入了同义词概念扩充和关联词语扩充。实验证明,该方法能够获得更高的内容过滤准确率,有效提高分类效果。

关键词 概念,段落化,文本分类,概念词典
中图法分类号 TP301 **文献标识码** A

Logical Paragraph Division Based on Semantic Characteristics and its Application

ZHU Zhen-fang LIU Pei-yu WANG Jin-long
(School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China)

Abstract A new matching method based on logic-centered paragraphs was introduced. The method built on the basis of the concept dictionary carried out the cluster analysis of the paragraphs which have the same meaning in the text by analyzing the logical concept of the text to be classified so as to get a logical level, and established the logical paragraph concept on the basis of the division method of the logical level, then measured the contribution of different paragraphs to the text theme according to the logical paragraph. At the same time, in order to solve problem of synonyms and polysemy in the matching process, the expansion of the synonyms concept and related words were introduced. Experimental results show that this method can obtain a higher accuracy rate in content flitting, improving the effectiveness of classification effectively.

Keywords Concept, Paragraphs, Text classification, Concept dictionary

1 引言

随着信息技术的发展,丰富的网络信息给用户发现信息、利用信息带来了方便。但是,人们在享受网络所带来的便利的同时,又不可避免地接触到大量的不良信息。信息分类是一种用来分类大量信息流,为用户提供相关信息子集的技术[1],即根据用户的信息需求,利用一定的工具从大规模的动态信息流中自动筛选出满足用户需求的信息,剔除无用信息的过程。

目前国内外关于文本信息过滤的研究基本上可以概括为两个方面。其一是关于用户模型研究,即用户模板(user profiles)的构建及其算法;其二是基于文档与用户需求的匹配技术研究,即用户模板与文本的匹配技术(filtering methods)。这两个方面是文本信息过滤的两大关键技术。

而在匹配技术中,不管是应用什么样的算法,最终都是对文本进行分类,都是对于事先构建模板的匹配。当前常用的匹配模型是向量空间模型(VSM)[3],而目前应用向量空间模型进行的匹配和分类中,往往都是整个待分类文档的匹配和分类,而忽略了待分类文本中的段落特征。

目前针对段落的匹配机制也往往是针对传统的物理段落,即给不同的段落赋予不同的权值,从而使用这些段落进行匹配,这就带有一定的机械性。那是因为这些物理段落往往较短或者本身包含的信息过少,甚至有些段落包含对于分类有副作用的信息。特别是在过滤网络文本时,获得的网络数据文本文档往往都是一些附加信息,如果使用这些段落实施匹配,极其容易出现分类误差和匹配率较低的现象。

因此,本文从更加广泛的词义出发,建立了一种以特征词概念为中心的逻辑段落结构,在此基础上实现了基于段落的匹配机制,以体现段落个性化特点,提高分类效果。

2 基于语义特征的文本段落划分

段落化匹配过程中,最简单的就是依照自然段作为划分

到稿日期:2009-01-09 返修日期:2009-03-09 本文受国家自然基金(60873247),山东省自然基金(Y2006G20),山东省高新自主创新专项工程(2008ZZ28)资助。

朱振方(1981—),男,博士研究生,主要研究方向为网络信息安全、遗传算法等,E-mail:zhuzhfyt@163.com;刘培玉(1960—),男,教授,博士生导师,主要研究方向为计算机网络信息安全、网络系统规划、网络信息资源开发和软件开发技术;王金龙(1979—),男,博士生,讲师,主要研究方向为网络与网络资源管理。

依据[4]，但是如果单纯采用自然段作为划分段落依据，容易导致部分段落虽然特征明显但是特征词数较少，从而影响匹配率。而如果将特征项映射至概念级，无疑将有助于加强同一层次内段落间的聚合能力。

为此，需要通过特征词语对文本包含的语义特征进行分析[5]，更好地理解文本的主题思想，了解文本所表达的内容及采用的方式。

2.1 相关理论基础

2.1.1 概念

概念是事物本质特征的概括和抽象，不受词汇语种、多义性和歧义性的影响[6]。概念的产生和存在依附于词语，而词语能够表示其他事物，是因为人们头脑中有相应的概念。词语是概念的语言形式，概念是词语的思想内容。

2.1.2 概念词典

概念词典精确定义了词语及其所对应概念之间的映射关系，能够用来解决自然语言中存在的同义词与多义词问题。在概念词典中，词被划分为词形、词性和概念定义。词形指词的物理形态，词性说明该词的语法功能，概念定义由一个或多个基本属性以及它们与主干词之间的语义关系描述组成，这三者作为一条记录储存在词典中。

此处应用的概念词典主要是指北京大学计算语言学研究所开发的中文概念词典[7,8](Chinese Concept Dictionary，简称CCD)。

在该概念词典中，名词(或动词)概念之间的主要关系是上下位关系：概念 C' 称为概念 C 的下位概念(hyponymy concept)或概念 C 是概念 C' 的上位概念(hyponymy concept)当且仅当命题 C' is a kind of C 为真。定义概念 C 是概念 C' 的祖先概念或概念 C' 是概念 C 的子孙概念当且仅当存在概念 $C_1,C_2,\cdots,C_n$ 使得 C' 是 C_1 的下位概念，$\cdots$，C_n 是 C 的下位概念。

2.1.3 概念密度

概念密度表示相关概念在文本中的聚集程度，此处将其定义为：

$$g(c)=\sum_{t\in S}f(t)/K^{d-1}$$

其中，集合 S 是项集特征向量 P 中概念 c 的所有下位概念的项的集合，t 是属于集合 S 的特征项，$f(t)$ 是 t 的频率，d 是 t 的概念到概念 c 的最短路径长度，K 是常数，且 $K>1$。

2.1.4 概念映射

输入文本经过分词和停用词处理后，获取文本的物理结构信息，这里主要获得文本每个段落的项集特征量，经过概念映射后，得到概念集特征向量，具体如下：

设文本 T 具有 n 个自然段，P 表示自然段，则有如下组成关系

$$T=P_1\cup P_2\cup P_3\cup\cdots\cup P_n$$

而对每个段落 I 都可用项集特征向量 $P_i=(\langle t_{i1},d_{i1},f_{i1}\rangle,\langle t_{i2},d_{i2},f_{i2}\rangle,\cdots,\langle t_{im},d_{im},f_{im}\rangle)$ 表示，其中 t_{ij} 为特征项，d_{ij} 为分词时词典中获取的 t_{ij} 的概念码，f_{ij} 为特征项 t_{ij} 的频率。

为此定义概念映射 $\Phi(P,\lambda):P\xrightarrow{\Phi}Q$，其中，$P$ 为项集特征向量，Q 为概念集特征向量

$$Q=(\langle c_{i1},g_{i1}\rangle,\langle c_{i2},g_{i2}\rangle,\cdots,\langle c_{is},g_{is}\rangle)$$

其中，λ 层是概念结点的代码，g_{ij} 是 c_{ij} 的概念密度。

2.2 基于概念的文本表示模型的构建

本文构建基于语义概念的文本表示模型是为了弥补向量空间模型在语言知识和领域知识中的不足，同时为实现基于概念的段落化匹配提供段落划分依据。模型构建过程如图1所示。

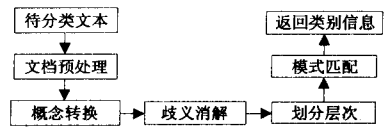

图1 基于概念的段落化匹配机制应用流程

2.2.1 文档预处理

按照传统的文本分类方法，对文本进行分词，把文本表示成一段词语序列，并计算其权值信息。本文应用项目组自行改进的TF-IDF统计方法计算特征项的权重。设总的文档数为 N，包含词条 t 的文档数为 n，其中某一类 C 中包含词条 t 的文档数为 m，则 t 在 C 类中计算公式为

$$IDF=\log(\frac{m}{n}*N)$$

如果在某一类 C 中包含词条 t 的文档数量大，而在其它类中包含词条 t 的文档数量小，则 t 能够代表 C 类的文本的特征，具有很好的类别区分能力。如果除 C 类外，包含词条 t 的文档数为 k，则公式的变形形式为：

$$IDF=\log(\frac{m}{m+k}*N)$$

2.2.2 概念变换

经过预处理的待分类文本可以表示成以概念词语及其权重为个体的向量，而概念变换则是通过查询概念词典得到每一个词语对应的一个或多个概念。以概念来表征文本特征，不但可以正确地表示文本的本质内容，同时，利用概念的抽象性还可将数个同义词语归结为一个概念。用概念来衡量特征词对类别的影响，获取关键概念及其他概念与关键概念的关系，就能模拟人类的分类过程并达到较高的准确率。

2.2.3 词义消歧

通常我们认为，多义词在某个特定的文本中表示的意义往往只有一个，也就是特征的语义局域性。词义消歧[9]的目的就是从一个词的所有可能的意思中剪除不相关的语义而保留正确的语义。

在文本分类过程中，分词和概念转换时会出现未登录词、没有概念标注的词和一词具有多个概念标注等情况。

1. 未登录词、无概念标注词语

在含有词 ω 的段落中，统计共现词频数 $f(\omega t)=l$，l 为 ω 与 t 共同出现的句子数。获取频数最大者 t 的概念结点 c，将 ω 的概念标注定义为 c 的子结点，c 为其父结点。

2. 一个词语具有多个概念标注

假设词典中 ω 有 m 个概念标注 $c_1,c_2,\cdots,c_m$，在含有词 ω 的段落中，统计共现概率函数为 $h_\omega(c_i)=\frac{1}{D}\sum f(c_i,t)$，$D$ 为 c_i 的子结点数，T 是一棵以 c_i 为顶点的子树，$f(c_i,t)$ 是 t 在段落中的频率，取共现概率函数最大者为 c 的概念标注。

2.3 应用特征词聚类的文本段落划分方法

同一层次的若干自然段，由于共同支持该层次所表达主题思想在概念上具有很强的集聚性，在使用的频率上也往往具有很大的相同之处。因此，通过特征词的聚类算法能够实

现文本逻辑段落的有效划分，从而实现文本的段落化匹配。

设文本 T 具有 n 个自然段，K 个层次，用 H 表示文本层次，P 表示自然段，则有如下组成关系

$$H_1H_2\cdots H_k=(P_{i_1}\cdots P_{i_2-1})(P_{i_2}\cdots P_{i_3-1})\cdots(P_{i_k}\cdots P_{i_{k+1}-1})$$

其中，$i_1=1\leqslant i_2\leqslant\cdots i_k\leqslant i_{k+1}-1=n$。

设文本 T 的特征向量为$(c_1,c_2,\cdots,c_m)$，则设 $P_i=(\omega_{i1},\omega_{i2},\cdots,\omega_{im})$为第 i 段的特征向量。其中 ω_{ij} 是概念 c 在第 i 段中概念密度。

将 n 个段落划分为 K 个层次，则所有可能的分法共有 C_{n-1}^{K-1} 种，设 $S(n,K)$ 是任一种分法，其中，$S(n,K)=\{i_1=1,i_1+1,\cdots,i_2-1\},\cdots,\{i_j,i_j+1,\cdots,i_{j+1}-1\},\cdots,\{i_K,i_K+1,\cdots,n\}$。

有序聚类就是寻找一种分法使 K 个层次内差异尽可能小，而层次间的差异尽可能大。设 $D(i_j,i_{j+1}-1)$表示第 j 层内的差异量，则误差函数为

$$E(S(n,K))=\sum_{j=1}^{K}(i_j,i_{j+1}-1)$$

为了使上述总体误差函数达到最小，寻求最优的 K 分法，相当于把 n 个段落分成两个部分，将前一部分进行最优 $K-1$ 分法，然后再考虑后一部分的误差，由此寻找到最优 K 分法。

设 $S(n,K,c_K)$是使总体误差函数达到最小的分法，其中 c_K 是上述最佳分法的分割点 i_k，则有如下递推公式

$$E(S_0(n,K,c_K))=\min\{E(S_0(i_{K-1},K-1,c_{K-1})+D(i_K,n))\}$$

文本层数的确定关系到文本逻辑结构的建立，它可以通过给定阈值 ℓ，当$|E(S(n,K+1)-E(S(n,K))|\leqslant\ell$ 时，则最优层数为 K。

3 文本分类的段落化匹配实现

在采用 VSM 进行分类的过程中，VSM 依赖于两个文本所共同包含的特征项的多少，因此往往是那些冗长的文本易于取得较高的相似度，因为包含的特征项较多，增加了共现的几率，实际上仅仅是偶然的提及或者出现的语境不同，这就给分类造成一定的困难。而一个文本真正属于一个类别必须存在与该类别相关的段落，如首段或者末段，这样可以防止“假相关”现象[10]。

而基于概念和关联扩充的文本分类机制不仅利用概念和关联扩充降低文本特征项之间的相关性和歧义性，同时在文本与类别特征向量相关的基础上，考察段落特征向量与文本类别特征向量之间的关系，最后确定分类的类别，从而实施分类。

3.1 逻辑段落概念词语的单一性

上述提出的基于概念的文本结构分析产生的逻辑段落既可以实现段落化匹配，还能够减少无效段落对于匹配的干扰。但是，由于实施以概念为中心的匹配需要选取能够代表中心概念的词语，也就不可避免地会出现同义词匹配的误差。同时，如“教师”一词可以有效地区别教育类文献，但是如果选取“教师”一词作为概念中心词，则如果出现“教授”、“讲师”之类的词语则不能有效辨识，为此，需要以“教师”一词为中心进行相关词语的词义扩充以及相关知识的相近搭配扩充。

为解决该问题，本文参考文献[11]引入了关联词语扩充解决词义搭配扩充问题，引入同义词概念扩充进行概念词的同义词扩充，提高匹配率。

3.2 基于概念的概念扩充和关联词语扩充

3.2.1 关联词语扩充

关联扩充的目的在于主题词在搭配方面的扩充。关联扩充的依据是关联矩阵，关联矩阵来自于相应语料库的统计结果，表明某一词与其它词之间的搭配频率，在一定程度上可以提供该词出现的上下文环境信息。

选取相应的文本集作为统计语料，主要计算词汇之间的共现频率，计算的单位为句子，在统计时选取实词参加运算，滤去虚词和停用词，以减少运算量和提高词汇特征的表现能力。虚词指数词、量词、介词等。停用词为高频词和一些不常用的低频词，如工作、研究、进行、认为等。

另一个值得考虑的因素是词汇之间的距离，计算的单位是句子，所以两个词汇必须是句子内相邻词，其关联强度随着距离的增大而减小，超过一定距离时，可以认为无关联。因此，将距离因素加入关联系数表达式中。

设 $K>1$ 为允许的最大关联距离常数，l 为词 t_i 与 t_j 之间的距离。若 t_i 与 t_j 是同一个句子中距离小于 K 的相邻词，则其局部关联系数为：$r_{ij}=\dfrac{\log_2(\frac{K}{l})}{\log_2 K}$，否则令其关联系数 $r_{ij}=0$。

设 tf_i 表示 t_i 的出现频数，tf_j 表示 t_j 的出现频数，S_{ij} 表示 t_i 与 t_j 的共现句的集合，即同时包含 t_i 与 t_j 的句子集合，关联矩阵 $A=(a_{ij})$，称 a_{ij} 为词 t_j 与 t_j 之间的关联系数，则有：

$$a_{ij}=\frac{\sum_{t\in S}r_{ij}}{tf_i+tf_j}$$

通过如上运算形成关联矩阵，对于类别的原始特征向量 P_0 中每一项 t_i 在关联矩阵中选取与之关联度最大的前 L 个词作为关联扩充，经过整理获得类别的关联特征向量 $P_0=(\langle a_1,\omega_1\rangle,\langle a_2,\omega_2\rangle,\cdots,\langle a_n,\omega_n\rangle)$。

3.2.2 同义概念扩充

概念扩充就是将若干个低级概念节点归结为较高级概念节点，这是一个迭代过程，直至所有的概念节点彼此独立。为了防止概念扩充后的概念层次过高而造成含义过于笼统，失去具体含义，可以在概念扩充过程中将其限制在指定的概念层次之下。选择合适的概念层次，将其作为扩充的临界层 λ。

定义概念扩充 $\Phi(P,\lambda):P_0\rightarrow P_c$。其中 P_0 为文本特征向量，P_c 为概念特征向量。$P_0=(\langle t_1,f_1\rangle,\langle t_2,f_2\rangle,\cdots,\langle t_n,f_n\rangle)$和 $P_c=(\langle c_1,g_1\rangle,\langle c_2,g_2\rangle,\cdots,\langle c_n,g_n\rangle)$是概念结点的代码，其层数小于或者等于临界层 λ，g_1 是 c_1 的概念密度。其中概念密度为$\dfrac{\sum_{t\in S}f(t)}{K^{d-1}}$。

它表示概念在特征向量中的集聚程度。其中集合 S 是概念 c 的所有下位概念的项的集合。t 是属于集合 S 的特征项，$f(t)$是 t 的权重，d 是 t 的概念结点到概念 c 的最短路径长度，K 是常数($K>1$，如 $K=\sqrt{2}$，$K=2$ 等)。

对于非登录词和没有概念标注的词汇，将其作为一个 λ 层的概念结点，其子结点集合为空，权重设为频率值。

3.3 段落化文本分类实现

基于概念和关联扩充的文本分类机制不仅可以利用概念和关联扩充降低文本特征项之间的相关性和歧义性，同时也

可以在文本与类别特征向量相关的基础上，考察段落特征向量与文本类别特征向量之间的关系，最后确定文档的类别，从而决定是否加以分类。

假设文本的分类为 $C=\{C_1,C_2,\cdots,C_m\}$，C 为文本集，C_i $(i=1,2,\cdots,m)$为划分的类别，$C_i=(\langle t_{i1},f_{i1}\rangle,\langle t_{i2},f_{i2}\rangle,\cdots,\langle t_{is},f_{is}\rangle)$，$t_{ij},f_{ij}$ $(j=1,2,\cdots,s)$是类别 i 的主题词表及其权重值，称 C_i 为类别的原始特征向量。

设待分类文本为 T，$T=\{P_0,P_1,\cdots,P_n\}$，P_i $(i=1,2,\cdots,n)$为文本段落，$P_i=(\langle t_{i1},\omega_{i1}\rangle,\langle t_{i2},\omega_{i2}\rangle,\cdots,\langle t_{is},\omega_{is}\rangle)$，$t_{ij},\omega_{ij}$ $(j=1,2,\cdots,s)$是段落 i 的主题词表及其权重值，称 P_i 为段落的原始特征向量。值得指出的是 P_0 为文本标题。

对于文本类别原始特征向量 C_i $(i=1,2,\cdots,m)$和段落原始特征向量 P_i $(i=1,2,\cdots,n)$实行概念扩充操作，最终获得类别特征向量 LC_i 和段落特征向量 LP_i。即：

$LC_i=\Omega(\Phi(C_i,\lambda),l),(i=1,2,\cdots,m)$

$LP_i=\Omega(\Phi(P_i,\lambda),l),(i=1,2,\cdots,m)$

定义段落特征向量和类别特征向量的相似度为：

$$\mathrm{sim}(C_i,P_j)=\frac{LC_i^T LC_j}{||LC_i^T||\,||LC_j||}$$

文本与类别 CP_i 的相似度定义为：

$$\mathrm{sim}_{Class}(T,C_i)=W_T\mathrm{sim}(T,C_i)+W_H\mathrm{sim}(P_0,C_i)+W_F\mathrm{sim}(P_1,C_i)+W_L\mathrm{sim}(P_n,C_i)$$

其中，W_T,W_H,W_F,W_L 为可调参数，表示文本各个段落在分类过程中的重要性。文本 T 属于类别 C_K，则 $K=\underset{1\leqslant i\leqslant m}{\mathrm{argmax}}\ \mathrm{sim}_{Class}(T,C_i)$。

为了防止因简单的提及、偶然的出现或者分类语境不同而造成的分类错误，要考察重要段落的分类趋势。同时文本中的各个段落对于文本主题的表现能力也有差异，这就需要通过简单的综合评价获取该文本对于分类的影响程度。在依靠全局相似度进行预选的基础上，进行更为确切的局部相似度运算，来判断文本归属的类别。

4 模拟实验数据

4.1 实验环境

实验过程中，假定测试文本集合 $S=\{x_1,\cdots,x_n\}$，同时使用二进制序列 $y_1,\cdots,y_n$ 表示相应的测试文本是否归属于一个类 C，若 $y_i=1$ 时，表示 $x_i\in C$；若 $y_i=-1$ 时，表示 $x_i\notin C$。再假定分类器 $f(x_i)\in\{-1,1\}$ 表示对测试文本集合 S 中测试文本的类别归属判断。

将测试样本集合 S 表示为 $S=S_{正}\cup S_{反}$，称 $S_{正}$ 为测试文本的正例集合，且 $S_{正}=TP\cup FN$；$S_{反}$ 为测试文本的反例集合，且 $S_{反}=FP\cup TN$。其中：$TP=\{x_i\in S|f(x_i)=1\wedge y_i=1)\}$，$FN=\{x_i\in S|f(x_i)=-1\wedge y_i=1)\}$，$FP=\{x_i\in S|f(x_i)=1\wedge y_i=-1)\}$，$TN=\{x_i\in S|f(x_i)=-1\wedge y_i=-1)\}$[12]。

基于以上这些定义，重新定义准确率、召回率如下：

$$prec(f,s)=\frac{|TP|}{|TP|+|FP|}$$

$$rec(f,s)=\frac{|TP|}{|TP|+|FN|}$$

4.2 实验结果分析

4.2.1 文本分类实验

在下面的分类实验中，我们将采用 4.1 节中所表述的准确率和召回率两个参数来评估分类的性能。同时，在应用上述方法进行文本层次划分过程中，参照文献[13]给出的划分文本层次的结果，不同长度的文本其层数大都在 2～6 之间，此处取 6。

训练文档采用了复旦大学计算机信息与技术系国际数据库中心自然语言处理小组李荣陆提供的测试语料，共 9804 篇文档，分为 20 个类别。在实验中使用了文档数超过 1000 篇的类别中的计算机、经济、政治、经济以及体育 5 个类别，随机选取 1000 篇文本，其中 200 篇作为测试文本集合，其余 800 篇作为训练文本集合，4 个类共 4000 篇文本。在这 4000 篇训练文本中分别取 200，800，1600，3200，4000 共 5 组训练文本集合进行封闭测试分类实验。相关数据集的有用参数如表 1 所列。

表 1 训练集统计数据

编号	文本数量	词条数目	概念数
1	200	3983	8060
2	800	9038	11364
3	1600	12795	13030
4	3200	18107	14487
5	4000	19696	14894

实验中采用基于原始空间向量模型(VSM)以及基于文中所涉及的方法进行分类比较实验，并应用作者项目组[14]自行设计实现的文本分类器，相关分类准确率如图 2 所示。

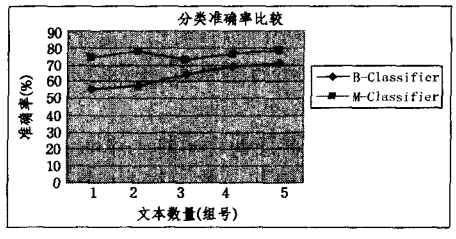

图 2 分类精度比较图

图 2 所示的分类精度比较图中，B-Classifier 代表采用基本 VSM 模型进行文本表示的分类器，即图中 B-Classifier 对应的曲线所示；M-Classifier 代表采用文中基于概念的段落化匹配机制的向量空间模型进行文本表示的分类器，即图中 M-Classifier 对应的曲线所示。

从图 2 可以看出，当训练文本集合的规模从 200 篇到 4000 篇逐渐增大时，前者的分类准确率从 85.11%上升到 88.36%；而后者的只从 45.36%上升到 60.86%。非常明显，采用文中改进的文本表示模型进行文本表示的分类器，其分类准确率总是比采用基本向量空间模型进行文本表示的分类器的高。

同样，在图 3 中，B-Classifier 代表采用简单 VSM 模型进行文本表示的分类器，即图中 B-Classifier 对应的曲线所示；M-Classifier 代表采用文中基于概念的段落化匹配机制的向量空间模型进行文本表示的分类器，即图中 M-Classifier 对应的曲线所示。

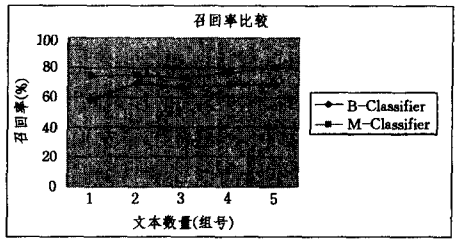

图 3 召回率比较图

从该曲线图可以看出，当训练文本集合的规模从 200 篇到 4000 篇逐渐增大时，前者的分类召回率从 84.4%上升到

(下转第 256 页)

参 考 文 献

[1] Santa A, Guido A, Daniela P, et al. An Image Adaptive, Wavelet-based Watermarking of Digital Images[J]. Journal of Computational and Applied Mathematics, 2007, 210(1/2): 13-21

[2] 陈晨，成礼智. 基于奇异值的 DWT 域公开零水印技术[J]. 通信学报，2006，27(11A)：81-84

[3] 刘彤，裘正定. 小波域自适应图像水印算法研究[J]. 计算机学报，2002，25(11)：1195-1199

[4] 陈青苏，祥芳，王延平. 采用小波变换的鲁棒隐形水印算法[J]. 通信学报，2001，22(7)：42-45

[5] Zhang Li, Qian Gong-bin, Xiao Wei-wei. Geometric Distortions Invariant Blind Second Generation Watermarking Technique Based on Tchebichef Moment of Original Image[J]. Journal of Software, 2007, 18(9): 2283-2294

[6] 康显桂，黄继武，等. 抗仿射变换的扩频图像水印算法[J]. 电子学报，2004，32(1)：7-12

[7] Zhang Li, Sam Kwong, Gang Wei. Geometric Moment in Image Watermarking[C]// Proceedings of the 2003 International Symposium on Circuits and Systems. 2003, 2: 25-28

[8] 胡玉平，韩德志，羊四清. 抗几何变换的小波域自适应图像水印算法[J]. 系统仿真学报，2005，17(10)：2470-2475

[9] 张仁昌，耿国华. 基于奇异值分解和小波变换的抗几何失真数字水印新方法[J]. 计算机应用与软件，2007，24(7)：33-35

[10] Wang Xiang-yang, Wu Jun, Niu Pan-pan. A New Digital Image Watermarking Algorithm Resilient to Desynchronization Attacks[J]. IEEE Transactions on Information Forensics and Security, 2007, 2(4): 655-663

[11] Bas P, Chassery J M, Macq B. Geometrically Invariant Watermarking Using Feature Points[J]. IEEE Transactions on Signal Processing, 2002, 11(9): 1014-1028

[12] Lee Hae-Yeoun, et al. Evaluation of Feature Extraction Techniques for Robust Watermarking[J]. Lecture Notes in Computer Science, Germany: Springer, 2005, 3710: 418-431

[13] Mikolajczyk K, Schmid C. Scale & Affine Invariant Interest Point Detectors[J]. International Journal of Computer Vision, 2004, 60(1): 63-86

(上接第 230 页)

89.55%；而后者的只从 48.11%上升到 63.25%。很明显，与分类准确率变化情况相似，采用本文设计的文本表示模型进行文本表示的分类器，其分类召回率总是比采用基本向量空间模型进行文本表示的分类器的高。

从以上两图可以看到，基于文中所涉及的方法在处理训练文本集合小的情况下，与基于词根的文本表示模型相比，能挖掘出更多的表现训练文本集合内容的语义特征，从而提高了文本分类的准确率和召回率。

4.2.2 信息过滤效果测试实验

由于文中设计的基于上述概念分析的段落化分类策略最终要应用到基于内容的信息过滤中，因此，试验中还将上述分类器应用于网络信息过滤的测试实验。试验中，对色情、暴力和合法 3 个类别进行训练和测试，训练和测试文档集均选自搜狗大规模语料库，每个类别各选取 1200 篇，其中 1000 篇用于训练，200 篇用于测试。其中合法类别是从非暴力和色情的文档集合中随即选取的 1200 篇文档。训练后测试结果如表 2 所列。

表 2 过滤效果测试统计数据

类别	文本数量	有效过滤数		准确率(%)	
		整体匹配	段落化匹配	整体匹配	段落化匹配
色情	200	187	192	93.50	96.00
暴力	200	176	184	83.00	92.00
合法	200	143	161	71.50	80.05

从表 2 可以看出，改进算法表现出了较好的过滤效果，同时，对色情暴力等具有鲜明特色的类别具有更好的分类效果，而最终要过滤的就是该类不良信息。

综合上述数据可以看出，上述匹配和分类策略应用于信息过滤具有较好的过滤不良信息的能力，因此上述方法的应用是有效的。

结束语 本文针对当前信息过滤中分类系统由于某些文档匹配度过低导致分类错误的现象，采用段落作为匹配元素，以提高过滤效果。同时，针对自然段落匹配过程中存在的一些问题，引入基于概念的文本段落划分方法。实验证明，基于该文本段落划分的段落匹配机制能够获得较好的召回率和准确率，能够有效地实现基于内容的文本信息过滤。

参 考 文 献

[1] 程妮，崔建海，王军. 国外信息分类系统的研究综述[J]. 现代图书情报技术，2005(6)：30-38

[2] 田范江，等. 进化式信息分类方法研究[J]. 软件学报，2000，11(3)：328-333

[3] 庞剑锋，卜东波，白硕. 基于向量空间模型的文本自动分类系统的研究与实现[J]. 计算机应用研究，2001(9)：23-29

[4] Hind J. Organizational Patterns in Discourse, Syntax and semantics: Discourse and Syntax[M]. New York: Academic Press, 1979

[5] 赵丰年，刘林，商建云. 基于概念的文本过滤模型[J]. 计算机工程与应用，2006，42(4)：186-188

[6] 张俐，王宝库，姚天顺. 从英文 WordNet 到中文 WordNet[C]// 中文信息处理国际会议论文集. 北京：清华大学出版社，1998：355-360

[7] 于江生，俞士汶. 中文概念词典的结构[J]. 中文信息学报，16(4)

[8] 阎蓉，张蕾. 一种新的汉语词义消歧方法[J]. 计算机技术与发展，2006，16(3)

[9] 郑文贞. 段落的组织[M]. 福州：福建人民出版社，1984

[10] Yang Y. An evaluation of statistical approaches to Text Category[J]. Journal of Information Retrieval, 1999, 1(1/2): 67-88

[11] 郑海，林鸿飞. 基于段落匹配的文本分类机制[J]. 计算机工程与应用，2004(28)：174-176

[12] Stephan B, Andreas H. Boosting for text classification with semantic features[C]// Proceedings of the MSW 2004 Workshop at the 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Seattle, WA, USA, 2004

[13] 林鸿飞，战学刚，姚天顺. 基于概念的文本结构分析方法[J]. 计算机研究与发展，2000，37(3)：324-328

[14] Zhu Zhenfang, Liu Peiyu, Lu Ran. Research of text classification technology based on genetic annealing algorithm[C]// ISCID 2008. 2008: 265-269