第32卷 第20期 计 算 机 工 程 2006年10月
Vol.32 № 20 Computer Engineering October 2006

·博士论文· 文章编号：1000—3428(2006)20—0004—03 文献标识码：A 中图分类号：TP311

CFV-NB：基于概念特征向量的NB文档分类模型

何 丽[1,2]，刘 军[2]

(1. 天津大学管理学院，天津 300072；2. 天津财经大学理工学院，天津 300222)

摘 要：提出了一种基于概念特征向量的 NB 文档分类方法。该方法在未标注文档集上通过 SOM（Self-Organizing Maps）聚类产生若干初始文档类，并为每个文档类分配一个类标签，使用最大信息熵的方法建立每个文档类的概念特征向量。在概念特征向量空间上建立最终的文档分类器：CFB-NB。

关键词：文档分类；概念特征向量；NB 分类器

CFV-NB: Naïve-Bayes Documents Classification Model Based on Concept Feature Vectors

HE Li[1,2], LIU Jun[2]

(1. College of Management, Tianjin University, Tianjin 300072;

2. College of Technology, Tianjin University of Finance and Economics, Tianjin 300222)

【**Abstract**】This paper proposes a novel Naïve-Bayes document classification method based on the set of concept feature vectors. It produces some initial classes from the set of unlabeled Web documents by SOM clustering and distributes a label for each, and builds the corresponding concept feature vector for each initial class using the maximum entropy method. It builds the last CFV-NB document classifier based on the space of concept feature vectors.

【**Key words**】Document classification; Concept feature vectors; NB classifier

文档分类技术一直 IR 领域的一个重要研究课题，也是开发新一代智能搜索引擎的关键技术之一。近几年来，关于文档分类研究已获得了不少的研究成果。应用较为广泛的文档分类方法主要有：基于向量空间模型的 K-近邻分类方法，基于统计学习理论的支持向量机(Support Vector Machine, SVM)以及基于统计分析的贝叶斯分类方法和基于神经网络的文档分类方法等。但是，目前使用的文档分类方法大都建立在文档特征向量空间，存在文档特征向量的高维和人工参与过多等问题。本文根据同类文档在内容上的聚集特征，提出了一种基于概念特征向量的文档分类新方法。该方法通过概念特征向量的建立，不仅可以解决文档分类过程中的高维问题，还可以动态地建立用于文档分类的永久分类模式。基于概念特征向量的文档分类方法主要由以下 3 个步骤组成：

(1)获得初始文档类并建立对应的概念特征向量集；

(2)获得每个初始文档类的标注文档集；

(3)建立基于概念特征向量的 CFV-NB 文档分类模型。

1 概念特征向量及其表示

概念特征是特定文档集合中的所有文档在内容特征上的高层抽象，概念特征代表了文档在概念层的分类特征。

1.1 文档的结构化表示

无论是中文文档还是英文文档，一般是一种半结构化的数据，在进行分类之前，首先需要将文档集 U 中的每个文档转换为计算机可以直接处理的结构化数据。本文采用向量空间模型将每个文档 $d_j(d_j \in U, 1 \le j \le |U|)$ 表示为一个结构化的文档特征向量：

$$d_j = \{<t_1, w_{1j}(d_j)>, <t_2, w_{2j}(d_j)>, ..., <t_N, w_{Nj}(d_j)>\}$$

其中，t_i 表示对文档 d_j 中经过切词、分词和无用词去除后抽取的一个特征词；$w_{ij}(d)(1 \le i \le N)$ 表示特征词 t_i 在文档 d_j 中的权值，N 为根据全局字典设置的统一的文档特征向量长度。词 t_i 在文档 d_j 中的重要性权重 w_{ij} 用 TF*IDF 方法计算：$w_{ij} = tf_{ij} \times \log(n / idf_i)$，其中 tf_{ij} 表示词 t_i 在文档 d_j 中出现的频数，idf_i 表示出现词 t_i 的文档数，n 是文档集中的文档总数，即：$n = |U|$。为避免文档长度的影响，最后将 w_{ij} 规范化为

$$w_{ij} = \frac{w_{ij}}{\sqrt{\sum_{i=1}^{N}(w_{ij})^2}}$$

1.2 概念特征向量描述

概念特征是同一文档类文档内容特征的抽象，因此，在建立概念特征向量之前，首先需要建立初始文档类。在这里，建立初始文档类的目的是为了建立概念特征向量，因此，可以使用一般的文档聚类方法来实现。常用的文档聚类方法主

基金项目：天津自然科学基金资助项目（033611011）；天津市高等学校科技发展基金资助项目（20041603）

作者简介：何 丽(1969－)，女，讲师、博士生，主研方向：Web 数据挖掘，决策支持系统；刘 军，副教授、博士生

收稿日期：2006-01-25 **E-mail**：renke21@vip.sina.com

要有：以 G-HAC（Group-average Hierarchical Agglomerative Clustering）等算法为代表的层次凝聚法(Hierarchical Agglomerative Clustering, AHC)，以 k-Means 算法为代表的平面划分法，以及以 WEBSOM 为代表的自组织映射方法。本文使用 SOM 网络的自组织特性对训练文档集进行聚类。使用该方法，一方面是为了在未标注训练文档集上自动产生建立概念特征向量需要的初始文档类；另一方面是为了获得不同概念特征向量之间的继承和层次关联。

概念特征向量是对每个初始文档类概念特征的一种向量表示。这里假设通过在未标注训练文档集U上进行 SOM 聚类，总共产生了 m 个有效的初始文档簇，用 $C=\{C_1,C_2,...,C_m\}$ 表示。其中，$C_i\,(1\le i\le m)$ 表示第 i 个文档簇的类标签。根据概念特征向量的定义，对产生的 m 个初始文档簇，应建立 m 个对应的概念特征向量。

本文仍然使用向量空间模型来表示概念特征向量，并将第 i 个初始文档簇的概念特征向量表示为

$$cf_i=(<t_{i1},w_{i1}(cf_i)>,<t_{i2},w_{i2}(cf_i)>,...,<t_{il},w_{il_i}(cf_i)>)$$

其中：w_{ij}（$j=1,2,...,l$）表示特征词 t_j 在概念特征向量 cf_i 中的权重，l_i 是概念特征向量的长度。不同文档簇对应的概念特征向量的长度可以不同。m 个文档类的概念特征向量集用 $CF=\{cf_1,cf_2,...,cf_m\}$ 表示。

1.3 概念特征词的抽取

文档特征向量通常是高维的。在文档特征向量中，不同权重的特征词对文档类别的贡献率也不同，通常只有 15~20 个特征词对文档类别的确定是至关重要的[1,3]。因此，建立基于概念特征向量的文档分类器的关键是如何从初始文簇所包括的文档特征词集合中抽取具有高类别贡献率的概念特征词。

目前，基于向量空间模型的有效特征词的抽取方法主要有：信息增益，互信息、期望交叉熵和x^2统计法等。本文采用最大信息熵方法来抽取每个初始文档簇的概念特征词。依据信息熵计算原理，某特征词的信息熵越大，则表示该特征词对特定文档类的类别贡献率也越高。特征词t_i在概念特征向量cf_k中信息熵计算公式定义为

$$E_{ik}=\log(f_{ik})\times(1+\frac{1}{\log(m)}\sum_{j=1}^{m}(\frac{f_{ij}}{n_i}\log\frac{f_{ij}}{n_i})) \tag{1}$$

其中，E_{ik} 表示特征词 t_i 在概念特征向量 cf_k 中的信息熵，f_{ik} 表示特征词 t_i 在文档簇C_k的文档中出现的频数，n_i 表示包含特征词 t_i 的文档簇的个数，m代表概念特征向量的个数，等于SOM聚类产生的初始文档簇的个数。将 E_{ik} 进一步标准化为

$$E_{ik}=\frac{E_{ik}}{\sqrt{\sum_{i=1}^{n_k}(E_{ik})^2}} \tag{2}$$

其中，n_k表示文档类C_k中出现的特征词的个数。E_{ik}表示一个特征词 t_i 的对某概念特征向量的信息熵的值越大，表明t_i对该概念特征向量的贡献越大，对文档分类的影响也越大。因此，通过设置一个概念特征词选择阈值 λ，就可以获得每个概念特征向量中的特征项。当 $E_{ik}\ge\lambda$ 时，将特征词 t_i 选出作为概念特征向量cf_k的一个特征词，否则，将该特征词过滤掉。不同的初始文档类，由于其涵盖内容的广泛性程度不同，产生的概念特征向量长度也可能不一样。信息熵E_{ij}表明了概念特征词t_j在概念特征向量cf_i中的权重w_{ij}。

2 CFV-NB 分类模型

由于贝叶斯分类模型具有分类速度快和分类准确率高等优势，因此，本文采用朴素贝叶斯分类模型(Naïve-Bayes, NB)来验证基于概念特征向量的文档分类方法的有效性，并建立基于概念特征向量的 NB 分类模型：CFV-NB 文档分类器。

在朴素贝叶斯分类模型中，对一个待分类文档 X，当且仅当 $\forall j\ P(C_i\mid X)>P(C_j\mid X),1\le j\le m,\ j\ne i$ 时，分类器才将 X 分配到文档类 C_i 中。根据贝叶斯分类原理，文档 d_i 属于文档类 C_j（$j=1,2,...,m$）的后验概率为

$$P(C_j\mid d_i)=\frac{P(d_i\mid C_j)P(C_j)}{P(d_i)} \tag{3}$$

因为 $P(d_i)$ 对所有文档类 C_j 来说是个常量，可以忽略对它的计算。所以建立 CFV-NB 文档分类器的关键是计算 $P(C_j)$ 和 $P(d_i\mid C_j)$，$P(C_j)$ 和 $P(d_i\mid C_j)$ 的计算过程就是分类器的建立过程。

2.1 构造标注文档集 L

分类是一种有监督学习方式。在建立 CFV-NB 文档分类模型之前，首先必须建立与每个概念特征向量代表的类标签对应的标注文档集 L_j（$j=1,2,...,m$），以获得 NB 分类模型中的 $P(C_j)$ 的值。

本文使用基于距离的相似性测量方法来产生标注文每个标注文档集 L_j（$j=1,2,...,m$）。对训练文档集U中的每个文档 d_i $(1\le i\le|U|)$，计算其与每个概念特征向量 $cf_j\in CF$（$j=1,2,...,m$）之间的距离相似度 $s_j=sim(d_i,cf_j)$（$j=1,2,...,m$），最后将 d_i 分配到最相似的概念特征向量对应的标注文档集 L_r 中，即

$$s_r=\max_{1\le j\le m}s_j$$

文档 d_i 与概念特征向量 cf_j 之间的相似度 $sim(d_i,cf_j)$ 测量使用应用广泛的余弦距离测量函数 $sim(d_i,cf_j)$，即：

$$sim(d_i,cf_j)=\cos(d_i,cf_j)=\frac{w_{i1}*w_{j1}+...+w_{i|V|}*w_{j|V|}}{|d_i|*|cf_j|} \tag{4}$$

其中，

$$|d_i|=\sqrt{\sum_{k=1}^{|V|}w_{ik}^2}\text{，}|cf_j|=\sqrt{\sum_{k=1}^{|V|}w_{jk}^2}$$

V 表示包含在概念特征向量集CF 中的所有词的集合，$|V|$ 表示V 中包含的词总数。$sim(d_i,cf_j)$ 的值越大，表示二者的相似度越高。

考虑不同文档类的文档在内容特征上的关联和交叉情况，本文在标注文档集的建立过程中引入了模糊理论的思想。根据标注文档集 L_j（$j=1,2,...,m$）中每个文档 d_i（$i=1,2,...,|L_j|$）与对应概念特征向量 cf_j 相似度的大小，为 d_i 分配一个属于文档类 C_j 的隶属概率 $P(C_j\mid d_i)$。当 $sim(d_i,cf_j)$ 值大于等于规定阈值 Λ_j 时，令 $P(C_j\mid d_i)=1$，否则令 $P(C_j\mid d_i)=sim(d_i,cf_j)/\Lambda_j$，其余的 1-$P(C_j\mid d_i)$ 为其他各个文档类平均共享。不失一般性，

本文将 Λ_j 定义为 L_j 中的所有文档与概念特征向量 cf_j 相似度值的平均，实现算法如下：

输入：概念特征向量集 $CF = \{cf_1, cf_2, ..., cf_m\}$ ；

训练文档集 $U = \{d_1, d_2, ..., d_n\}$ ；

输出：每个文档类的标注文档集 $L = \{L_1, L_2, ..., L_m\}$ ；

过程：

(1)初始化 $L_i = \phi, (i = 1, 2, ..., m)$ ；

(2)对 U 中的每个文档 d_i ，根据式(2)计算它与所有 $cf_j \in CF$ $(j = 1,2,..., m)$ 之间的相似度：$s_j = sim(d_i, cf_j)$ ；

(3)令 $s_r = \max\{s_1, s_2, ..., s_m\}$ ；

(4)若（$s_r > 0$)则 $L_r = L_r \cup \{d_i\}$ ；

(5)计算每个 $L_j \in L$ 的相似度阈值：

$$\Lambda_j = \sum_{i=1}^{|L_j|} sim(d_i, cf_j) \Big/ |L_j|$$

(6)计算文档对对应标注文档集的隶属度：

FOR j =1 to $|L|$ DO

FOR i =1 to $|L_j|$ DO

若 $sim(d_i, cf_j) \geq \Lambda_j$ ，则

$\{P(C_j | d_i) = 1; P(C_k | d_i) = 0$

$(k \neq j, k = 1,2,.., m);\}$

否则：{令 $P(C_j | d_i) = sim(d_i, cf_j) \big/ \Lambda_j$ ；

$P(C_k | d_i) = 1 - P(C_j | d_i) \big/ m - 1$

$(k \neq j, k = 1, 2, .., m);\}$

2.2 建立 CFV-NB 分类模型

在 CFV-NB 分类模型中，$P(C_j)$ 和 $P(d_i | C_j)$ 的计算过程就是分类器 CFV-NB 的建立过程。

设 D 为根据上述算法获得的所有标注文档的集合，$|D|$ 表示 D 中包含的文档总数。则 CFV-NB 分类模型中每个初始文档类的先验概率 $P(C_j)$ 可估计为

$$P(C_j) = \frac{\sum_{i=1}^{|D|} P(C_j | d_i)}{|D|} \tag{5}$$

假定每个文档特征词的抽取均符合朴素贝叶斯假设，则文档 d_i 属于文档类 C_j 的后验概率可近似为

$$P(d_i | C_j) = \prod_{k=1}^{|V|} P(w_k | C_j) \tag{6}$$

考虑可能存在 $P(d_i | C_j) = 0$ 的情况，这里使用拉普拉斯平滑，使：

$$P(w_k | C_j) = \frac{1 + \sum_{i=1}^{|D|} n(w_k, d_i) P(C_j | d_i)}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D|} n(w_s, d_i) P(C_j | d_i)} \tag{7}$$

其中，$n(w_k, d_i)$ 为特征词 w_k 在文档 d_i 中出现的次数。对一个新文档 d_i，根据 $P(C_j)$ 和 $P(d_i | C_j)$ 的计算结果，就可以得到 d_i 属于文档类 C_j（$j = 1,2,..., m$）的后验概率 $P(C_j | d_i)$，最后将文档 d_i 分配到后验概率最大的文档类中。

3 实验分析与总结

为了验证概念特征向量对文档的分类效果的影响，本文使用了 20_ newsgroups(http://www-2.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes/20_newsgroups.tar.gz) 数据集中 SCI、REC 和 TALK 这 3 个主题下的 900 篇文档，每个主题下抽取的文档数均为 300 篇。随机抽取其中的 600 篇文档组成训练文档集，其余的 300 篇组成测试文档集。该实验的目标是验证基于概念特征向量的文档分类方法的有效性，并以分类准确率作为评价标准。准确率定义为正确分类的文档数与文档集中同类文档总数的比值，实验结果如图 1。

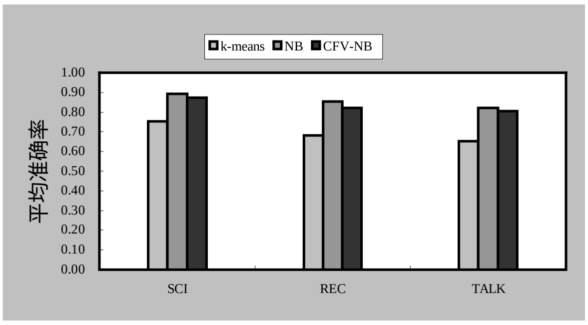

图 1 分类准确率对照图

从图 1 可以看出，CFV-NB 文档分类器的平均分类准确率非常接近先验概率正确情况下的传统 NB 分类器，而且比基于文档特征向量的 k-means 聚类方法的分类准确率平均高出 13.7%。实验结果证明，基于概念特征向量的 CFV-NB 文档分类方法是有效的。它不仅获得了很好的文档分类效果，还解决了传统文档分类过程中人工参与过多和基于文档特征向量文档分类方法中的高维计算问题，基本实现了文档分类过程的自动化。除此之外，在 CFV-NB 分类模型建立过程中产生的概念特征向量可以作为永久的分类模式来维护和使用，这些分类模式在基于概念层的交互式搜索引擎和增量式文档分类方法的实现中具有重要的实际应用价值。

参考文献

1 Weng Sung-Shen, Liu Chih-Kai. Using Text Classification and Multiple Concepts to Answer E-mails[J]. Expert Systems with Applicatons, 2004, 26(4) : 529-543.

2 Smith K A, Ng A. Web Page Clustering Using a Self-organizing Maps of User Navigation Patterns[J]. Decision Support Systems, 2003, 35(2): 245-256.

3 张义忠，赵明生，梁久祯. 基于自组织特征映射的网页分类研究[J]. 信息与控制, 2003, 32 (2): 108-112.

4 王爱华，张 铭，杨冬青等. PCCS 部分聚类分类：一种快速的 Web 文档聚类方法[J]. 计算机研究与发展, 2001, 38(4): 416-421.