第34卷 第13期 计 算 机 工 程 2008年7月
Vol.34 No.13 Computer Engineering July 2008
·软件技术与数据库· 文章编号：1000—3428(2008)13—0061—03 文献标识码：A 中图分类号：TP391

基于 RSOM-Bayes 的网页分类方法

冯和龙[1]，夏胜平[2]
(1. 湖南铁路科技职业技术学院实训中心，株洲 412000；2. 国防科学技术大学电子科学与工程学院 ATR 重点实验室，长沙 410073)

摘 要：针对向量空间模型的网页分类计算复杂度高、不适用于大规模场景问题，该文采用 RSOM 和 BAYES 相结合的方法实现网页分类，利用 RSOM 神经网络树实现网页特征词的自动索引，利用 Bayes 实现网页的自动分类。结果证明其在特征空间维数、检索效率、样本容量及检索精度方面都具有良好的性能。

关键词：网页分类；RSOM 神经网络树；Bayes 方法；向量空间模型

Web Page Classification Method Based on RSOM-Bayes

FENG He-long[1], XIA Sheng-ping[2]
(1. Practical Centre, Hunan Railway College of Science and Technology, Zhuzhou 412000; 2. State Lab of Automatic Target Recognition, College of Electronic Science and Engineering, National University of Defense Technology, Changsha 410073)

【Abstract】 Most Web page classification methods are based on Vector Space Model(VSM), but it is not suitable for large scale application background with bad computation complexity. A new automated text classification method based on RSOM neural net tree and Bayes method is proposed, RSOM neural net tree is used in Web page index and Bayes method is used in automated Web page classification. The excellent performance of this method has been tested in feature dimension, performance, capacity and accuracy.

【Key words】 Web page classification; RSOM neural net tree; Bayes method; Vector Space Model(VSM)

1 概述

如何利用因特网海量信息资源找到有用的信息，是网页分类关注的重点。目前网页自动分类有许多方法，如Naïve Bayes, SVM, Boosting, KNN, 决策树[1-3]。这些方法大致可分为 2 类[4]，即基于向量空间模型方法和基于语义模型方式，目前的主要研究进展是在基于向量空间模型的网页分类方面。但基于向量空间模型的方法计算复杂度太高，不适合于大规模场景，如用NB和SVM方法进行文本分类时，输入的特征空间都是整个特征集，高达数万个。SOM网络作为一种无导师示范、具有自组织功能的神经网络，通过对输入模式的反复学习，可以使连接权矢量空间分布密度与输入模式的概率密度趋于一致，即连接权矢量空间分布能反映输入模式的统计特征。它在已知特征空间样本分布或大致分布的情况下可以取得很好的自组织效果。这时，直接使用SOM模型往往无法取得很好的效果。RSOM[5]是基于SOM网络进行的一种树状扩展，以SOM网络为基本节点用递归方法生成的层次化聚类树，在RSOM树中，空间上彼此靠近的数据点通常聚合在一起，RSOM树为海量数据索引提供了一种可行的办法。

本文通过对海量基本词汇 RSOM 库的构建，用特定样本的网页数据进行训练，构建特定主题的 RSOM 树，用 Bayes 方法实现网页的自动分类。

2 对基本词汇的 RSOM 树构造

基于词库的基本词汇树构造过程，实际就是一个 RSOM 的训练学习过程。将词库的每一个词描述成 30 维向量，除了存储每个词的内码，还可以包括每个词的属性。

设某一样本集U有p个样本$\boldsymbol{U}_k=[u_1{}^k,u_2{}^k,\cdots,u_n{}^k]^{\mathrm{T}}$，分属 c 个模式类，$\omega_i=\{U_k{}^i,k=1,2,\cdots,N_i\},i=1,2,\cdots,c$，将样本集$\boldsymbol{U}$分成$c$个子集，则有

$$\boldsymbol{U}=\bigcup_{i=1}^{c}\omega_i \tag{1}$$

每个子集ω_i表示同一模式类样本组成的集合。$\bar{m}_i$ 表示ω_i模式类的样本均值；$\bar{m}$ 为所有样本的均值；P_i为ω_i类在所有样本中所占的频度；$\boldsymbol{S}_{\omega i}$为$\omega_i$类的类内离差阵，$\omega_i$类与$\omega_j$的类间距离为$d^L(\omega_i,\omega_j)$，总的类内离差阵为$\boldsymbol{S}_w$，总的类间离差阵为$\boldsymbol{S}_B$，其定义见文献[6]，则类别可分性判据定义为

$$J=\frac{Tr[S_B]}{Tr[S_w]} \tag{2}$$

可知，J 越大，样本可分性越好；J 越小，样本可分性越差。不妨设定某个阈值 θ，当 J 小于阈值 θ 时，样本集不可分。RSOM 算法用该可分性判据来控制 RSOM 树的生长，θ 一般取值为 0.1 左右。

首先对所有原始训练样本用一个 SOM 网络进行训练，得到一组输出节点，之后按最近邻原则将所有原始训练样本分配到相应的节点，由此形成一个分类树的根节点。考察根节点所属输出节点，对分配到其中的样本进行可分性判决条件的检测，若不可分，则将该节点属性赋为叶节点，停止该节点的分解。若可分，则用与根节点 SOM 网络训练完全相同的算法对该节点进行训练，得到相应的 SOM 网络，并将该节点的样本分配到相应的输出节点，由此，通过采用递归的方法对所有节点进行类似的分析，直到没有节点需要进一步生长为止。这样就得到了一棵 RSOM 神经网络树，简称 RSOM 树。在 RSOM 树的生长过程中，有多种控制因子，包括节点样本的可分性判据、层数控制、样本数控制，以保

作者简介：冯和龙(1965－)，男，高级实验师，主研方向：信息处理，计算机应用；夏胜平，副教授、博士

收稿日期：2007-07-20 **E-mail：**fhl@hntky.com

证所训练得到的 RSOM 树具有优良的结构。具体训练方法见文献[5]。由上述 RSOM 树基本训练算法，经训练后最终构建了如图 1 所示的基本词汇 RSOM 树。

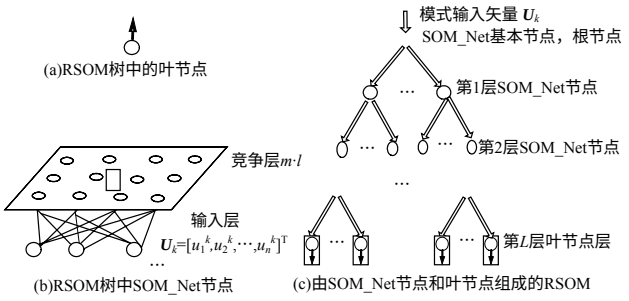

图 1　RSOM 树的基本结构

在一定训练样本数条件下，RSOM树中SOM网络的个数将有限增长，不妨设SOM_Net中网络基本结构为 3×3，若生成一个 5 层的网络，所需训练的SOM网络最大个数为 1＋9＋9^2＋9^3＋9^4＝7 381，此时叶节点数最大可达 66 429 个，设平均每个节点中包含 50 个样本，则整棵RSOM树样本容量达 3 321 450，为百万量级，增大网络基本结构，则网络容量将进一步快速增加，如网络结构变为 2×5，其他条件相同的情况下，网络容量可增加到 10^7，为千万量级。

3　基于 RSOM 树的词频统计学习

在得到基本词汇的 RSOM 树后，开始对该树进行第 2 阶段的学习。

(1)从网页搜索得到用户所关心的一定数量的网页，并加以标识。所得到的词汇根据其所属的文档类型赋予相应的类别属性，而且文档类型还可以具有多属性特征，如同一篇网页既可以是计算机方面的，也可以是自动控制方面的，这样，每个特征词可以具有多种类别属性。

(2)对文本切分处理后得到的特征词汇逐个用RSOM树进行匹配，找到RSOM树中相应的基本词，并进行词频统计。在此过程中，如果某个词匹配到RSOM树的某个叶节点，而其中不存在相应的基本词汇，那么就是出现了新的特征词汇，则将其添加到相应的叶节点中。设有特征词汇U_0，输入RSOM树，要求用经过第 1 阶段学习的RSOM树对所关心的文本进行词频统计学习。下面给出基于RSOM树对训练文本的词频进行统计学习的算法。

算法 1

(1)初始化，载入训练样本集U_0，设RSOM树中学习的特征词个数为$Count_T$，总计有M类词汇，每类词汇学习的计数器为$Count_T=\{Count_T^{(j)}(i),j=1,2,\cdots,M\}$，每个词对$M$类词汇学习的属性计数器$Count_t(i), i=1,2,\cdots, Count_T$，其中，$Count_t(i)$为该词对每一类的计数器$Count_t(i)=\{Count_T^{(j)}(i), j=1,2,\cdots,M\}$，表示$M$类词汇的学习次数计数向量。

(2)按序从U_0中取一个样本U_t，将根节点SOM网络当作获胜节点网络SOM_Net。

(3)用获胜节点中SOM_Net保存的归一化参数对U_t归一化为$\overline{U}_t$。

(4)将$\overline{U}_t$输入当前 SOM_Net，采用 SOM 网络获胜节点求解方法求得相应的获胜节点。

(5)若获胜节点为叶节点，转(7)。

(6)转(3)。

(7)检查当前叶节点中包含的样本数p_{Lf}，若$p_{Lf}>p_{lf}^{\max}$，则对当前叶节点用文献[5]的方法进行进一步的RSOM生长。

(8)设当前叶节点中包含p_{Lf}个样本，将当前样本与叶节点中的p_{Lf}个样本进行匹配处理。

1)若存在与该样本精确相等的特征词，若该词尚无类别属性计数器数组，则设定其计数器数组为$Count_t(i), Count_T=Count_T+1, i=Count_T$。

2)若不存在与该样本精确相等的特征词，则将该样本添加到当前叶节点中，对该词设立类别属性计数器数组$Count_t(i), i=Count_T+1$。

将$Count_t(i), Count_T$中对应的项加 1；若该词有多个属性项，则将$Count_t(i), Count_T$中每一个对应的项加 1。

(9)将当前样本从U_0中删除，结束对一个样本的处理。

(10)检查U_0是否非空。非空，返回(2)。

(11)根据 RSOM 树中确定特征词及类的先验分布。

1) 对每个特征词U_i的学习次数计数器$Count_t(i), i=1,2,\cdots,Count_T$进行频度统计，得到

$$\pi(U_i|H_j)=Count_t^i(j)/Count_T^{(j)}, j=1,2,\cdots,M, i=1,2,\cdots,Count_T \tag{3}$$

2)对每个类的特征词的学习次数$P_T{}^j$进行频度统计：

$$\pi(H_j)=Count_T^{(j)}/\sum_{k=1}^{M}Count_T^{(k)}, j=1,2,\cdots,M,\ j=1,2,\cdots,M \tag{4}$$

(12)结束学习。

由此，可得到一定主题范围内的特征词 RSOM 树。由算法可知，这种词频统计处理的速度很快。以前述最大 5 层的 RSOM 树为例，其规模可达 3 百多万，在学习过程中对一个特征向量进行词频统计处理只需进行 5 次 9 个节点的 SOM 网络求获胜节点的处理，以及最多含 50 个样本的叶节点样本匹配处理，而且对每个词的处理速度是均衡的，如此完成一次预分成 30 类的、3 000 个经切分的短篇文档，平均每篇约 200 个词计 60 万个特征词汇的学习不到 1 小时。

4　基于 RSOM 树的 BAYES 网页分类

在完成 RSOM 树第 2 阶段的学习之后，即可以根据从网页上搜索得到的文档对网页进行自动分类。

设某个网页的特征文档τ经自动切分后，得到文档词汇集$U_T=\{U_i,i=1,2,\cdots\}$，由算法 2 已经得到了每一个词对文本分类的统计信息，可用Bayes方法求得$U_i,i=1,2,\cdots,l$条件下各类的后验概率：

$$\pi(H_j|U_i)=\pi(U_i|H_j)\times\pi(H_j)/(\sum_{k=1}^{M}(\pi(U_i|H_k)\times\pi(H_k))) \tag{5}$$

其中，U_i为文本文档词汇集U_T中的元素，$i=1,2,\cdots,l$；$j=1,2,\cdots,M$表示总计有M类词汇。

不难发现，有些词汇对分类的后验概率$\pi(H_j|U_i)$，$j=1,2,\cdots,M$各项比较均衡，其含义是这些词不具有强的“专业性”，因而在分类过程中，也不需要将U_T中的全部词汇当作文档的特征词。事实上，设U_i条件下的后验概率$\pi(H|U_i)=[\pi(H_1|U_i),\pi(H_2|U_i),\cdots,\pi(H_M|U_i)]'_{M\times1}$，并设列向量$\boldsymbol{V}=[1/M,1/M,\cdots,1/M]'_{M\times1}$，用下式对其均衡性进行考察：

$$d(\boldsymbol{V},\pi(H|U_i))=(\pi(H|U_i)-\boldsymbol{V})'\times(\pi(H|U_i)-\boldsymbol{V}) \tag{6}$$

可以统计保留$d(\boldsymbol{V},\pi(H|U_i)), i=1,2,\cdots,l$中最大的$K$个不重复的$U_i$作为表征该文档$\tau$的特征词向量集$U_\tau$，值得注意的是每一个$U_i$可出现多次，则：

$$U_\tau=\{U_i',num\_U_i'|i=1,2,\cdots,K\} \tag{7}$$

对U_τ中重复出现的特征词U_i'采用乘法方式增强其后验并进行归一化处理如下：

$$\pi(H_j|U_i')=\pi(H_j|U_i')^{num\_U_i'}/(\sum_{j=1}^{M}\pi(H_j|U_i')^{num\_U_i'}) \quad (8)$$

$\pi(H_j|U_i'), j=1,2,\cdots,M$ 中可能存在等于 0 的项，其含义是在某类型的文档中没有出现过该词，但可能出现的情况是，对其他的特征词，相应的项可能很大，甚至等于 1，为了保证验后信息不被某个和几个等于 0 的项掩盖，在进行文本分类的判决时，采用加性统计量，仍然用后验概率的符号表示：

$$\pi(H_j|U_\tau)=\frac{1}{K}\times\sum_{i=1}^{K}\pi(H_j|U_i'), j=1,2,\cdots,M \quad (9)$$

对 $\pi(H_j|U_\tau)$ 中的项按降序排列，取前 2 项构造如下规则：

若 $\pi(H_m|U_\tau)/\pi(H_q|U_\tau)\geq\lambda_{\min}$，则该文档判决为第 m 类；若 $\pi(H_m|U_\tau)/\pi(H_q|U_\tau)<\lambda_{\min}$，则该文档判决为第 m 和第 q 类 (10)

其中，$\lambda_{\min}$可取为 2，可知，本文算法对一个文档可判定为具有 2 类属性，当然，也可采用Winner-Take-All的准则进行决策，即该文档判决为第m类。这样就完成了网页的自动分类。

其算法流程如下：

算法 2

(1)初始化，载入训练后的特征词RSOM树；搜索某个网页的一个文档，对文档内容自动进行切分，得到文档词汇集U_T={U_i,i=1,2,⋯,l}。

(2)从U_T中取一个需处理的样本U_i，将根节点SOM网络当作获胜节点网络SOM_Net，用算法 1 中的步骤(3)～(6)进行处理求得当前样本在RSOM树中所属的叶节点。

(3)设当前叶节点中包含p_{Lf}个样本，将当前样本与叶节点中的p_{Lf}个样本进行匹配处理。

1)若存在与该样本精确相等的特征词，且其计数器数组为P_t^i不等于 0，用式(5)求得 $\pi(H|U_i)$，以用式(6)求得的$d(\boldsymbol{V},\pi(H|U_i)$作为$\pi(H|U_i)$的度量指标，将$\pi(H|U_i)$按降序存入长度为l的向量列表；将该词设为已处理状态，返回(2)。

2)若不存在与该样本精确相等的特征词，将该词设为已处理状态，返回(2)。

(4)从$\pi(H|U_i)$列表中按降序取K个不重复的U_i，并对每个U_i的重复次数进行统计，得到文本特征词向量集U_τ＝{U_i ' ,$num\_\ U_i$ ' , i=1,2,⋯,K} 并用式(8)进行处理得到相应的 $\pi(H|U_i')$。

(5)用式(9)构造决策统计量，由式(10)进行判决。

(6)若自动分类的文档数目小于Doc_$num_{\max}$，返回(1)。

(7)将自动分类的Doc_$num_{\max}$篇文档及其自动分类结果提交用户进行验证，若自动分类结果错误，则修正给出正确的分类属性。

(8)统计自动分类的平均正确率P_C，若P_C低于$P_C^{\min}$，提请用户确认可在后台对新近分类的文档按算法 1 进行积累学习，此时载入的是近阶段用于自动分类的特征词RSOM树，训练完毕形成最新的RSOM树循环用于自动分类。

通过这样的一个过程，自动分类系统学习到的文档数量不断增加，其性能也不断提升，自动分类系统处于这样一个循序渐进的进化过程之中。

5 试验分析

在实验中，选择 SOM 网络结构为：输入层神经元个数为 20 个，输出层神经元个数为 3×3 个，RSOM 树共 5 层。首先选用 10 万条基本词汇、10 万条常用专业词汇及 2 万条人名、车名、地名等共 22 万词汇以及常用英语词汇对 RSOM 树进行训练，在 P4 2.4 GHz，内存 1 GB 的双 CPU 计算机上，花费 6 s 完成一棵 RSOM 树的训练，得到了基本词汇的 RSOM 树。

试验 1：下载 5 000 个中文网页， 通过人工方式将其分为 20 类，即文化生活、宗教种族、天文地理、计算机、音乐、电信、环境、数学、物理、生物、信息、机械、石油、航空、能源、电力、农牧林、机电仪器、电机工程、地理等。将 5 000 篇文档分成 2 个集合：一是训练集，包含 3 000 篇文档，另一个是测试集，包含 2 000 篇文档。

用 3 000 篇训练文档来进行 RSOM 树的词频统计学习。首先对文挡进行切分，平均每篇得到约 200 个特征词，3 000 篇文档共计 60 万个特征词汇，然后，经过 45 min 的训练完成了 RSOM 树的词频统计学习。

对 2 000 篇文档进行了测试，分类准确率为 97.8%，每篇文档识别时间 0.9 s。为了说明文中方法的有效性，笔者把它与支持向量机(SVM)、K-近邻算法(KNN)等方法进行了对比实验。从表 1 可知，本文方法具有很高的分类准确率和很快的处理速度，完全适用于在线实时网络分类。

表 1　3 种方法的分类比较

方法	分类准确率/(%)	每篇文档处理时间/s
SVM	92.1	50.0
KNN	89.5	380.0
文中方法	97.8	0.9

试验 2：选用 Reuter 发布的 Reuter Corpus(Volume 1:English Language, 1996-08-20~1997-08-19)，该全集以网页的方式发布。数据分布和测验结果如表 2、表 3 所示。

表 2　数据分布

标号	代码	类别领域	训练数	测试数
1	C13	REGULATION/POLICY	1 542	518
2	C18	REGULATION/POLICY	2 233	722
3	C21	PRODUCTION/SERVICES	1 379	459
4	E12	MONETARY/ECONOMIC	1 097	288
5	E21	GOVERNMENT FINANCE	2 023	638

表 3　测试结果　(%)

类号	Naïve Bayes 方法		RSOM+Bayes 方法	
	Precision	Recall	Precision	Recall
类 1	66.93	81.27	82.58	81.79
类 2	85.53	82.69	93.45	85.34
类 3	92.86	76.47	94.37	87.23
类 4	71.76	84.72	87.56	88.45
类 5	95.23	81.50	97.74	94.81

6 结束语

随着 Internet 数据急剧增长，如何从海量的网页信息中高效、快速地检索出所需信息是当前许多应用领域的重要问题。鉴于网页数据海量、高维的特性，所构造的索引树能很好地适应这些特点。本文采用基于 RSOM 和 Bayes 相结合的方法实现网页的自动分类。实验证明，无论从特征空间维数、检索性能、样本容量及检索精度方面都具有良好的性能。

参考文献

[1] Mitchell T. Machine Learning[M]. [S. l.]: McGraw, Hill, 1996.

[2] Slattery S. Hypertext Classification[D]. Pittsburgh: Carnegie Mellon University, 2001.

[3] Yang Yiming, Slattery S, Ghani R. A Study of Approaches to Hypertext Categorization[J]. J. Intelligent Info. Syst., 2002, 18(2/3): 219-241.

[4] 孙建涛, 沈　抖, 陆玉昌, 等. 网页分类技术[J]. 清华大学学报: 自然科学版, 2004, 44(1): 65-68.

[5] 夏胜平, 张乐锋, 胡卫东, 等. 基于 RSOM 树模型的机器学习原理与算法研究[J]. 电子学报, 2005, 33(5): 937-944.

[6] 孙即祥. 现代模式识别[M]. 长沙: 国防科技大学出版社, 2002.