基于类别特征选择与反馈学习随机森林算法的邮件过滤系统研究

孙 雪[1] 韩 蕾[1] 李昆仑[2]
[1](河北大学工商学院 河北 保定 071000)
[2](河北大学电信学院 河北 保定 071002)

摘 要 针对邮件过滤系统中普遍存在的维数灾难、类别主题差异和反馈信息缺失问题,提出一种基于类别特征选择与反馈学习随机森林算法的邮件过滤模型。该方法将隐含的 Dirichlet 模型引入到邮件的特征选择环节,在不同类型的邮件集中建立各自的生成模型,分别搜寻构成各个主题的特征信息,有效降低冗余信息和噪声数据对分类性能的影响。反馈学习随机森林算法发挥了决策树集成与反馈学习的优势,实现邮件过滤系统的自我调节,及时捕捉垃圾邮件的变化趋势。在公开的语料库 CCERT 和 Trec06 上进行测试,并与典型算法进行比较,实验结果表明所提算法的可行性和有效性。

关键词 LDA 模型 特征选择 反馈学习 随机森林算法 垃圾邮件过滤

中图分类号 TP3 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2015.04.016

ON EMAIL FILTERING SYSTEM BASED ON CATEGORY FEATURE SELECTION AND FEEDBACK LEARNING RANDOM FOREST ALGORITHM

Sun Xue[1] Han Lei[1] Li Kunlun[2]
[1](*College of Industral and Commercial, Hebei University, Baoding 071000, Hebei, China*)
[2](*College of Electronic and Information Engineering, Hebei University, Baoding 071002, Hebei, China*)

Abstract To solve the problems of "curse of dimensionality", "diversity in the categories topic" and "lack of feedback" commonly existed in email filtering system, we propose an email filtering method which is based on category feature selection and feedback learning random forest algorithm. It introduces the latent Dirichlet allocation (LDA) model to the feature selection link of email and builds the respective generation model in different type of email sets to search separately the feature information forming each subject, thus effectively reduces the impacts of redundant information and noise data on classification performance. The feedback learning random forest algorithm plays to the advantages of decision trees integration and feedback learning, realises the self-regulation of the email filtering system and can catch the change trend in spam promptly. The test is done on open corpus CCERT and Trec06, and the comparison is made with typical algorithm as well. Experimental results demonstrate the availability and effectiveness of the proposed algorithm.

Keywords LDA model Feature selection Feedback learning Random forest algorithm Spam filtering

0 引 言

电子邮件具有方便、快捷、性价比高等特点,已经成为人们生活中必不可缺少的一种重要通信手段。随着电子邮件的普及,垃圾邮件问题也日益凸显,它占用了大量的网络资源,干扰了人们的正常生活秩序,部分垃圾邮件已经成为病毒传播的载体,严重威胁着网络的安全稳定[1,2]。垃圾邮件具有破坏性强、传播速度快、危害范围广等特点,已经成为全球化公害。不断泛滥的垃圾邮件给邮件过滤系统带来了更加严峻的考验。如何有效地过滤和治理这些垃圾邮件成为当前的迫切要求。

1 相关工作

现有的垃圾邮件过滤技术[3]主要包括基于规则的过滤和基于内容的过滤两大类。基于规则的过滤方法如黑白名单过滤、信头分析、关键词过滤等,该方法加入的主观因素较多,抗干扰能力较弱,规则制定的好坏将直接影响邮件的过滤效果。基于内容的过滤方法属于文本分类范畴,近年来随着对模式识别和机器学习研究的深入,该领域的一些算法如决策树、贝叶斯、支持向量机 SVM 等被引入到垃圾邮件过滤的技术中,并取得重大进展,已经成为当前垃圾邮件处理所采用的一种主要技术手段。M. Sahami[4]最早将贝叶斯算法应用于垃圾邮件过滤系统中,随即引起该领域学者的广泛关注,越来越多的人致力于此方面的研究。Xu[5]提出了基于改进的 BP 神经网络邮件过滤算法,克服了传统算法训练时间较长、容易陷入局部最小等问题。文献[6,7]将基于 IP 的白名单过滤、基于规则的过滤和基于统计的贝叶斯过滤算法结合在一起,实现了垃圾邮件的综合过滤。

收稿日期:2013-11-22。国家自然科学基金项目(60773062,61073121);河北省科技支撑计划项目(072135188);河北大学青年基金项目(2010Q17)。
孙雪,硕士,主研领域:数据挖掘。韩蕾,硕士。李昆仑,教授。

杨[8]基于网络协同过滤算法在 Enron 邮件集上对系统的过滤性能进行了测试。Song[9]把邮件收发者的关联度作为提取微博垃圾邮件特征词的标准。Chen[10]采用加权 SVM 算法对邮件进行分类。Gao[11]将动态邮件过滤技术应用到社交网络的垃圾评论在线识别系统。Bratko[12]设计了一种基于自适应统计数据压缩模型的邮件过滤算法,提高了系统的泛化性能。

目前提出的各种算法对垃圾邮件的过滤有一定的效果,但还存在一些不足:(1)垃圾邮件与正常邮件所阐述的主题内容是不同的,每种类型的邮件都有各自的表达方式和组织结构,而现有的垃圾邮件特征选择算法只考虑词特征对分类的影响而忽略了不同邮件类别中隐含的语义信息。(2)随着时间推移,垃圾邮件的种类和内容是不断变化的,且不同客户之间对邮件过滤系统的要求也存在差异性,如何实现邮件过滤系统的自我更新,满足不同客户间的差异化要求,也是一个完善的邮件过滤系统需要考虑的。

针对上述问题本文提出了一种基于类别特征选择与反馈学习随机森林算法,将隐含的 Dirichlet 模型引入到邮件的特征选择环节,在不同类型的邮件集中建立各自的生成模型,分别搜寻构成各个主题的特征信息,有效降低了冗余信息和噪声数据对分类性能的影响。同时将反馈学习与随机森林算法相结合,针对垃圾邮件种类实时变化的特点和客户群对邮件判别的差异性,实现邮件过滤系统的自适应调节。

2 基于类别信息的特征选择算法

2.1 LDA 生成模型

LDA[13](Latent Dirichlet Allocation)生成模型采用三层贝叶斯网络结构对文档信息进行建模,忽略文档中的句法结构和词语出现的先后顺序。该模型基于这样一种假设:文档是由若干个主题构成的,而主题又是由若干个词语构成的,例如一篇文档其内容可能涉及到经济、文化、历史等多个主题,每个主题都包含一些特定含义词语。LDA 模型生成文档流程如图 1 所示。

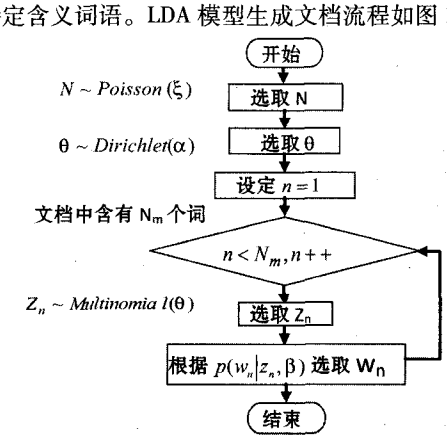

图 1 LDA 模型生成文档流程图

LDA 模型中特征项的权重可以表述为在隐含主题下生成特征的概率,V 表示文档集中包含不重复的词语总数;M 表示文档集中包含文档的个数;N 表示一篇文档中包含词的个数;N_m 表示第 m 篇文档中包含的词语个数;K 表示文档集中包含主题的个数;θ 服从 Dirichlet 分布 θ-$Dirichlet(\alpha)$,用 $1\times K$ 的列向量表示每个主题发生的概率,α 为 Dirichlet 分布的参数;Z 服从 Multinomial 分布 Z- Multinomial(θ),$p(z=i|\theta)=\theta_i$,$p(z|\theta)$ 表示给定 θ 时主题的条件分布;w 表示文档中所包含的词语分布;φ 也服从 Dirichlet 分布 φ-$Dirichlet(\beta)$ 用 $K\times V$ 的矩阵表示给定主题时词的条件分布。

2.2 基于类别信息的 LDA 特征选择算法

基于类别信息的 LDA 特征选择算法其前提假设为:邮件内容是由若干个主题构成的,且垃圾邮件与正常邮件所构成的主题是不同的。该算法的主要思想是将训练集中的邮件按照类别信息分成正常邮件和垃圾邮件,在各自的邮件集上生成 LDA 模型,并利用 gibbs 抽样算法获取模型参数。图 2 给出了该算法的框图,图中标示的参数与前文定义相同。

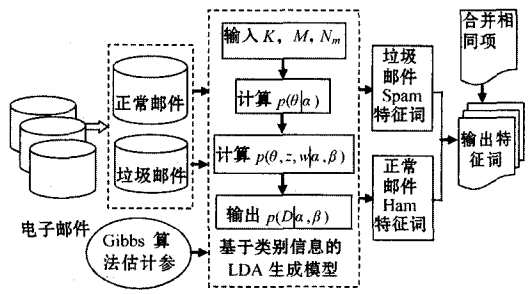

图 2 基于邮件类别信息的 LDA 特征选择算法框图

基于邮件类别信息的 LDA 特征选择算法的主要步骤如下:

Step1 输入参数 K_S,M_S,N_{mS},K_h,M_h,N_{mh};

Step2 利用 Gibbs 抽样算法估计 LDA 模型参数 α_s,β_s,α_h,β_h;

Step3 利用式(1)、式(2)计算 LDA 生成模型变量的条件概率:

$$p(\theta_s,z_s,w_s\mid\alpha_s,\beta_s)=p(\theta_s\mid\alpha_s)\prod_{n=1}^{N_m}p(z_{ns}\mid\theta_s)p(w_{ns}\mid z_{ns},\beta_s) \tag{1}$$

$$p(D_s\mid\alpha_s,\beta_s)=\prod_{m=1}^{M}\int p(\theta_{ms}\mid\alpha_s)\Big(\prod_{n=1}^{N_m}\sum_{z_{mns}}p(z_{mns}\mid\theta_{ms})p(w_{mns}\mid z_{mns},\beta)\Big)d\theta_{ms} \tag{2}$$

Step4 将垃圾邮件特征词 T_s 和正常邮件特征词 T_h 中包含的相同特征项合并;

Step5 输出基于邮件类别信息的 LDA 特征词。

注:角标 s 表示垃圾邮件,h 表示正常邮件。

3 基于反馈学习的随机森林算法介绍

3.1 基于反馈学习的邮件过滤系统

相关反馈[14]思想起源于 60 年代中期,最早被应用于信息检索领域,其作为一种有指导的学习过程,可以有效地提高系统的检索性能。反馈学习按照信息检索模型的不同可以分为基于向量空间模型的反馈算法、基于概率模型的反馈和基于布尔模型的反馈等。本文提出的邮件过滤系统采用向量空间模型,基于该模型的反馈算法如下式所示:

$$Q_{i'}=\alpha Q_i+\beta\sum_{rel}\frac{D_i}{N_{rel}}-\gamma\sum_{nonrel}\frac{D_i}{N-N_{rel}} \tag{3}$$

式中 Q_i 为未加入反馈的当前向量,$Q_{i'}$ 表示加入反馈信息后的调整向量,α、β 和 γ 为算法的调节参数,N 为用户反馈信息和邮件系统更新所包含的邮件数量,D_i 表示原始向量空间,N_{rel} 表示邮件与类别属性信息相关的文本数。

3.2　基于反馈学习的随机森林算法

随机森林[15]是一种基于决策树的集成学习算法,它融合了决策树与集成学习的优势。算法设计思想:在整个数据集上以一个固定概率分布随机生成多个子集,在每个子集上构造决策树,通过构建的多棵决策树来共同预测判别结果,有效降低了噪声数据对分类器性能的影响,且不会陷入过拟合。随机森林算法的执行过程如图 3 所示。

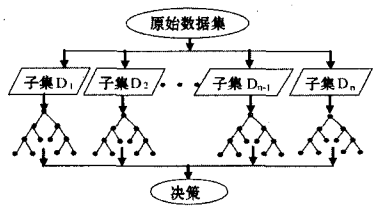

图 3　随机森林算法

基于反馈学习的随机森林算法描述:

输入:邮件训练集包括用户的反馈信息和新增加的邮件类型 $X = \{x_1, x_2, \cdots, x_n\}, x_i \in R^d$,测试集 $Y = \{y_1, y_2, \cdots, y_n\}$。

其中 n 表示训练集中包含的邮件个数,d 表示选取的特征项的维数。

输出:测试集中邮件类别 C。

算法步骤:

Step1　利用基于类别信息的 LDA 特征选择算法提取反馈信息的特征项,根据式(3)重新构建向量空间模型;

Step2　从邮件训练集抽取 K 个子集

利用 Bagging 算法从邮件集 $X = \{x_1, x_2, \cdots, x_n\}$ 中抽取 m 个样本,构造出同样大小的 K 个子集;

Step3　构造 K 棵决策树

①从原始 d 个特征中随机选取 F 个特征, $F = \log_2 d + 1$;

②利用式(4)和(5),分别计算 F 个特征的信息增益,选取信息增益最大的特征作为结点的分割属性。

$$\Delta_{info} = Ent(parent) - \sum_{j=1}^{2} \frac{N(\alpha)}{N} Ent(\alpha) \tag{4}$$

$$Ent(t) = -\sum_{i=1}^{2} p(i \mid t) \log_2 p(i \mid t) \tag{5}$$

式中的 $p(i \mid t)$ 表示节点 t 属于类别 j 的比例; N 表示父节点上的记录总数; $N(\alpha)$ 表示与子节点 α 相关联记录的个数; j 表示邮件类别, $j = 0$ 表示垃圾邮件, $j = 1$ 表示正常邮件;

③重复上面的方法,构造出决策树的各个分枝;

④重复执行①－③,直到 K 棵决策树构造完毕;

Step4　随机森林分类

①将测试数据输入到随机森林分类器模型;

②采用投票表决的方式对测试集中的邮件类别进行预测。

4　实　验

4.1　实验数据及评价指标

为验证算法的有效性,实验测试数据采用公开的垃圾邮件语料库 CCERT 和 Trec06,两个语料库均来源于真实邮件,保留了邮件的原有格式和内容。CCERT 是中国教育和科研计算机网紧急响应组提供的数据集分为 Jun 和 Jul 两个文件夹,Trec06 为国际文本检索会议提供的语料库,分为英文数据集(Trec06p)和中文数据集(Trec06c)两部分。中文邮件的预处理包括去除信头,分词,词性选择,过滤常用词,保留去掉词频和文档频度过高和过低的特征词,提取名词、动词、形容词和副词。英文邮件预处理包括去除停用词,词根还原(Stemming),去除信头三部分。文本表示选用向量空间模型(VSM)。

评价指标采用邮件过滤系统中经常采用的三个指标,即垃圾邮件的准确率(检对率)、正常邮件的召回率(检出率)和垃圾邮件的召回率,以及系统的精确率(判对率)。

4.2　基于类别信息的 LDA 特征选择算法

表 1、表 2 记录了基于类别信息的 LDA 特征选择算法(Categories-LDA)与其他特征选择算法卡方统计量(Chi)、信息增益(IG)、互信息(MI)、特征熵(TE)、文本证据权(WET)、期望交叉熵(ECE)在不同数据集上的实验结果。*sp* 表示垃圾邮件过滤的准确率,*sr* 表示垃圾邮件的召回率。

表 1　基于邮件类别信息的 LDA 特征选择算法与其他特征选择算法在 Trec06c 数据集上的实验对比结果

特征选择函数	Mul-Naive Bayes 分类器		C-SVC 分类器		J48 分类器	
	sp	sr	sp	sr	sp	sr
MI	0.734	0.998	0.78	0.994	0.713	0.990
Chi	0.973	0.867	0.995	0.841	0.958	0.981
TE	0.820	0.927	0.883	0.870	0.980	0.848
Wet	0.878	0.924	0.823	0.987	0.868	0.974
ECE	0.973	0.865	0.995	0.840	0.962	0.979
Categories - LDA	0.964	0.897	0.922	0.989	0.960	0.980

表 1 记录了在 Trec06c 中文数据集上的不同特征选择算法的实验对比结果,从数据集中随机抽取 19 364 篇垃圾邮件和 9 881 篇正常邮件作为训练数据,其余的 23 490 篇垃圾邮件和 11 885 篇正常邮件作为测试数据,以 Chi 特征选择算法得到的特征词权重作为向量空间模型的权重系数,并选取多项式分布模型贝叶斯 Multinomial-Naive Bayes、标准支持向量分类 C-SVC 和 J48 决策树三种分类器对不同特征选择算法选取的特征值进行分类。

表 2　基于类别信息的 LDA 特征选择算法与其他特征选择算法在 Trec06p 数据集上的实验对比结果

特征选择函数	Mul-Naive Bayes 分类器		C-SVC 分类器		KNN 分类器	
	sp	sr	sp	sr	sp	sr
MI	0.664	0.999	0.656	0.999	0.663	0.998
Chi	0.989	0.893	0.950	0.991	0.910	0.990
TE	0.966	0.883	0.946	0.989	0.925	0.988
Wet	0.988	0.795	0.989	0.932	0.852	0.997
ECE	0.990	0.885	0.949	0.992	0.915	0.989
Categories-LDA	0.985	0.928	0.970	0.973	0.938	0.985

表 2 记录了不同特征选择算法在 Trec06p 英文数据集上的实验对比结果,从数据集中随机抽取 11 505 篇垃圾邮件和 6651 篇正常邮件作为训练数据,其余的 12 440 篇垃圾邮件和

6143 篇正常邮件作为测试数据，选用向量空间模型（VSM）作为文本表示方法，以词频—文档频度（tf-idf）函数作为权重系数，采用 Multinomial-Naive Bayes、C-SVC 和 KNN 三种分类器对不同特征选择算法选取的特征值进行分类。

从表 1、表 2 可以看出采取不同的特征选择算法对邮件过滤系统影响的程度有所不同，同一种特征选择算法在不同分类器上表现的性能存在差异，而基于类别信息的 LDA 特征选择算法的性能几乎不受数据集和分类器的影响，表现比较平稳，没有较大波动，说明该算法有着较好的鲁棒性和适应性。

分析图 4 的折线图可以发现相同一种特征选择算法在不同的语料库上表现出的性能有所不同，这是由于邮件集自身差异导致的。一般情况下垃圾邮件与正常邮件在内容表述上区别很大，特征词较明显，但垃圾邮件的内容呈现多样性的特点，使得对分类有价值的特征词与低频词相混淆，不易提取。现有的特征选择算法存在一定的局限性，如信息增益算法的初衷是找到在一类中出现频率高在其他类中出现频率低且在其他类出现频率高在该类出现频率低的特征词，但其对垃圾邮件专有词与噪声词的区别度不高，导致分类性能不佳。MI 算法虽然给在某一类出现频度较高且在其他类别出现的频度较低的词赋予很高的互信息值，但垃圾邮件的种类繁多很难找出这样的特征词，使得该算法在邮件过滤中性能不好。总之特征选择算法受语料库规模，类别数量均衡，数据偏斜等各种因素影响。

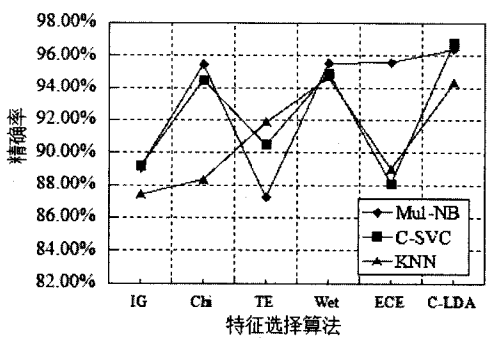

图 4 基于类别信息的 LDA 特征选择算法与其他特征选择算法在 CCERT-Jun 语料库上的实验精确率比较图

4.3 基于反馈学习的随机森林算法

基于反馈学习的随机森林算法主要包括两方面的反馈：一是针对新出现的垃圾邮件的形式和内容的反馈；二是针对客户的个性化需求的反馈，如有的客户喜欢网上购物，则与网上购物相关的邮件对该客户而言就是正常的邮件，这与常规的垃圾邮件的判断准则是冲突的，一个好的邮件过滤系统必须充分考虑到客户的需求。

表 3 中给出了在不同数据集上，Naive Bayes Mul、C-SVC、Bagging、KNN、J48、AdaBoostM1、RandomForest 和 Feedback-Rf（基于反馈学习的随机森林算法）七个分类器的邮件过滤性能比较。*sp* 表示垃圾邮件过滤的准确率，*hr* 表示正常邮件的召回率。基于反馈学习的随机森林算法在三个数据集上选取的反馈邮件的数量分别为：从 CCERT 语料库的 Jun 文件夹中随机抽取 3000 篇垃圾邮件和 2000 篇正常邮件作为反馈信息，从 Trec06p 英文语料库中随机抽取 2000 篇垃圾邮件和 1500 篇正常邮件作为反馈信息，从 Trec06c 中文语料库上随机抽取 4000 篇垃圾邮件和 3000 篇正常邮件作为反馈信息。

表 3 基于反馈信息的随机森林算法与其他分类器在不同数据集上的实验对比结果

特征选择函数	Trec06 数据集		Trec06p 数据集		CCERT-Jun 数据集	
	sp	hr	sp	hr	sp	hr
Naive	0.962	0.925	0.985	0.974	0.976	0.935
C-SVC	0.961	0.922	0.970	0.944	0.979	0.943
Bagging	0.959	0.916	0.965	0.932	0.980	0.948
KNN	0.968	0.937	0.938	0.878	0.978	0.940
J48	0.962	0.924	0.959	0.922	0.976	0.936
AdaBoostM1	0.950	0.924	0.833	0.631	0.898	0.709
R-Forest	0.983	0.967	0.984	0.970	0.988	0.968
Feedback-Rf	0.989	0.978	0.988	0.978	0.993	0.980

图 5 给出了基于反馈学习的随机森林算法与 Naive Bayes Mul、C-SVC、Bagging、KNN、J48、AdaBoostM1、RandomForest 算法的精确率比较结果。特征选择算法采用基于类别信息的 LDA 特征选择算法，从图中可以看出带反馈学习的随机森林算法性能要优于随机森林算法，这说明通过采取增大训练集的方式确实可以提高分类器的性能。对比分类器在邮件过滤系统中的精确度发现，不同分类器在同一种语料库上的表现各不相同，而同一种分类器在不同语料库上的精确度也存在差异，有的数值比较接近有的则相差很多，但这并不说明哪种分类器的性能不好，影响分类器的因素很多如数据集的规模、内容、分类器的参数、预处理的效果等，要根据实际情况选择适当的分类器，这样才能从根本上改善分类效果，提高分类精度。反馈学习效果受反馈数据集的规模和邮件的内容影响较大。一般而言，反馈学习都能在一定程度上提高邮件过滤系统的性能。

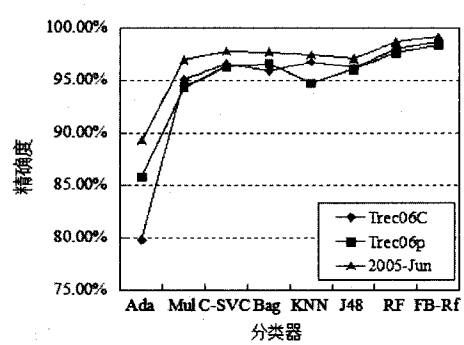

图 5 基于反馈学习的随机森林算法与其他分类器的精确率比较

5 结 语

针对邮件过滤系统中普遍存在的维数灾难，类别主题差异和反馈信息缺失问题，提出了一种基于类别特征选择与反馈学习随机森林算法的邮件过滤模型，在不同类型的邮件集中建立各自的生成模型，分别搜寻构成各个主题的特征信息，有效地克服了 LDA 模型中由于忽略类别主题差异性而造成的系统过滤性能下降的影响。由于垃圾邮件种类具有实时变化的特点且不

同客户群间对邮件过滤系统存在着差异性,本文将反馈学习理论与随机森林算法相结合,继承了决策树与集成学习的优势,有效地提高了邮件过滤系统的性能。垃圾邮件过滤属于文本分类范畴,因此用本文的方法来解决文本分类和信息过滤有潜在的应用价值。

由于 LDA 模型基于的前提假设为"Bag of words",在构建模型过程中其忽略了语法结构和词出现的先后顺序已经词语间的相互关系,下一步考虑在特征选择环节加入文档的其他信息,缩短特征空间与样本空间的描述差异。

参 考 文 献

[1] Caruana G, Li M. A survey of emerging approaches to spam filtering [J]. ACM Computing Surveys, 2012, 44(2): 1-27.

[2] Guzella T S, Caminhas W M. A review of machine learning approaches to Spam filtering[J]. Expert Systems with Application. 2009, 36(7): 10206-10222.

[3] Delany S J, Buckley M, Greene D. SMS spam filtering: Methods and data[J]. Expert Systems with Applications, 2012, 39(10): 9899-9908.

[4] Sahami M, Dumais S, Heckerman D E A, et al. Bayesian approach to filtering junk E-mail [C]//Proceeding of AAAI-98 Workshop on Learning for Text Categorization Technical Report WS-98-05. 1998: 55-62.

[5] Xu H, Yu B. Automatic thesaurus construction for spam filtering using revised back propagation neural network[J]. Expert Systems with Applications, 2010, 37(1): 18-23.

[6] 李玉峰,郜晓晶. 中文垃圾邮件过滤综合方法[J]. 计算机应用与软件,2011,28(8):219-221,226.

[7] 奚建荣. 基于综合过滤技术的邮件过滤终端研究[J]. 计算机应用与软件,2011,28(6):186-188,235.

[8] 杨震,赖英旭,段立娟,等. 邮件网络协同过滤机制研究[J]. 自动化学报,2012,38(3):399-411.

[9] Song J, Lee S, Kim J. Spam filtering in Twitter using sender-receiver relationship[C]//Int. Symp. Recent Advances in Intrusion Detection (RAID), 2011.

[10] Chen X, Liu P, Zhu Z, et al. A method of spam filtering based on weighted support vector machines[C]. IT in Medicine & Education, 2009.

[11] Gao H Y, Chen Y, Lee K, et al. Towards online spam filtering in social networks [EB/OL]. 2012. http://www.cs.northwestern.edu/~ychen/Papers/NDSS12_spam.pdf.

[12] Bratko A, Filipic B, Cormack G, et al. Spam Filtering Using Statistical Data Compression Models[J]. The Journal of Machine Learning Research, 2006(7): 2673-2698.

[13] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003(3): 993-1022.

[14] Croft W B, Harpe D J. Using probabilistic models of document retrieval without relevance [J]. Journal of Documenration, 1979 (35): 285-295.

[15] Breiman L. Random forests [J]. Machine Learning, 2001(45): 5-32.

(上接第 37 页)

响应时间的预测误差率。由于 CPU 等的服务时间 S 很短,响应时间主要受访问次数 V 的影响。由于负载 1 的命中率相较于负载 2 偏低,产生较大的预测误差,所以导致对输入参数 $V_{(物理读,硬盘)}$ 的估算不准确,而 $V_{(查询进程,逻辑读)}$ 只是访问次数平均值,使得最终对负载 1 的响应时间估算的准确率不高。

将 8 组实验数据按照缓冲池从小到大的顺序按 2、2、3 的组成形式进行分组,对于负载 1,MOL 算法的误差分别为 16.6%、14.81%、12.95%,而 SRVN 算法的误差分别为 20.18%、18.27%、18.21%;对于负载 2,MOL 算法的误差分别为17.16%、13.52%、4.1%,SRVN 算法的误差分别为10.05%、6.12%、3.22%。随着命中率的增大和稳定,访问次数逐渐趋于稳定,整个模型的响应时间预测也越来越准确。

MOL 和 SRVN 两种算法与实际值之间存在误差的另外原因是:由于两种算法在对资源争用时的等待队列长做近似估算时采用不同的方法:MOL 算法使用 Linearizer MVA 算法,而 SRVN 算法使用 Bard-Schweitzer MVA 算法。模型只描述同步服务,没有利用 SRVN 算法能够描述两阶段服务的能力,这可能是影响 SRVN 算法预测精度的另一原因。

4 结 语

本文提出了基于缓冲池的 DBMS 分层排队网络模型,对其求解算法——MOL 算法和 SRVN 算法进行了对比分析了,总结了采用两种算法所需的模型参数及其获取方法。示范了建立 DBMS 分层排队网络模型的方法和步骤。

参 考 文 献

[1] 赵建光,施剑,牛保宁. 数据库系统交易型负载自适应管理[J]. 计算机工程与应用,2013,49(6):131-134.

[2] 黄翔,王伟,张文博,等. 面向性能剖析的 Web 应用自动性能建模方法[J]. 软件学报,2012,23(4):786-801

[3] Zhang M, Niu B. Utility Functions in Autonomic Workload Management for DBMSs[J]. International Journal on Advances in Intelligent Systems, 2012, 5(1&2): 66-75.

[4] 徐增敏,张昆,丁勇,等. 基于动态视图的数据库性能调优[J]. 计算机应用与软件,2012,29(12):58-60.

[5] 姜林枫. 基于主动对象/行为图的主动面向对象数据库建模机制的研究与应用[J]. 计算机应用与软件,2013,30(4):177-179.

[6] 修树娟,张永茂. 合理设计缓冲池,提高数据库效率[J]. 科技信息,2010(30):I0247-I0247.

[7] 郭忠南,孟凡荣. 关系数据库性能优化研究[J]. 计算机工程与设计,2006,27(23):4484-4490.

[8] 李斌臣,苏桂平. 基于排队网络模型的数据库系统瓶颈问题研究[J]. 微电子学与计算机,2012,29(6):100-103.

[9] 贺娜,杨扬. 资源同时占有平均值分析方法的 C/S 系统应用[J]. 计算机与现代化,2004(4):50-52.

[10] Mirco Tribastone. A Fluid Model for Layered Queueing Networks. transaction on software engineering, 2013, 39(6): 744-756.

[11] 边学工,胡瑞敏,陈军,等. 基于分层排队网络模型的 MCU 性能预测及优化研究[J]. 计算机学报,2004,27(2):209-215.

[12] Rolia J A, Sevcik K A. The Method of Layers[J]. IEEE Trans. Software Engineering, 1995, 21(8): 689-700.

[13] Woodside C M, Neilson J E. The Stochastic Rendezvous Network Model for Performance of Synchronous Client-Server-Like Distributed Software [J]. IEEE Trans. Computers, 1995, 44(8): 20-34.

[14] 赵杰,牛保宁. 基于缓冲池描述的 DBMS 分层排队网络模型[J]. 计算机工程与设计,2013,34(11):3971-3976.

[15] Suri R, Sahu S. Approximate Mean Value Analysis for Closed Queuing Networks with Multiple-Server Stations[C]//Proceedings of the 2007 Industrial Engineering Research Conference, 2007.