基于二阶隐马尔科夫模型的云服务QoS满意度预测

贾志淳[1,2] 李 想[1] 于湛麟[1] 卢 元[1] 邢 星[1,2]
(渤海大学信息科学与技术学院 辽宁 锦州 121013)[1] (渤海大学自动化研究院 辽宁 锦州 121013)[2]

摘 要 随着云计算相关技术的迅速发展,云服务组件的QoS预测成为云计算中一个重要的研究课题。实现对QoS值的准确预测是该领域的研究难点。QoS常用来衡量不同云服务组件的性能,基于不同候选组件的QoS值,可以容易地选出最优的组件。对于同一个云服务组件,不同的用户提供的QoS值并不一定相同。针对不同的用户,有个性化的组件QoS值才能进行准确的选择。如果用户的QoS不能由单一的云服务组件满足,则应该考虑组件组合,在这种情况下,需要预测其QoS能力,以保证用户需求得到满足。文中设计了云服务组件的QoS满意预测模型,该模型使用二阶隐马尔科夫模型构建QoS满意度预测模型,通过考虑前两个状态对当前状态的影响,能够有效提高预测精度。最后,通过所构建的原型系统和具有2507个真实Web服务的QWS数据集,并应用Matlab仿真环境验证了所提方法的有效性。

关键词 云服务,二阶隐马尔科夫模型,服务选择
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.09.049

QoS Satisfaction Prediction of Cloud Service Based on Second Order Hidden Markov Model

JIA Zhi-chun[1,2] LI Xiang[1] YU Zhan-lin[1] LU Yuan[1] XING Xing[1,2]
(College of Information Science and Technology, Bohai University, Jinzhou, Liaoning 121013, China)[1]
(Institute of Automation, Bohai University, Jinzhou, Liaoning 121013, China)[2]

Abstract With the rapid development of cloud computing technology, QoS prediction of cloud service components has become an important research issue in cloud computing. Accurate prediction of the QoS value is of great difficulty in this research field. QoS is often used to measure the performance of different cloud service components. Based on the QoS values of different candidate components, it is easy to choose the best one. For the same cloud service component, the QoS values provided by different users are not necessarily the same. For different users, personalized component QoS values are needed so that accurate selection can be made. If the user's QoS cannot be satisfied by a single cloud service component, the component composition should be considered. In this case, its QoS capability should to be predicted to meet the user's needs. This paper presented a QoS satisfied prediction model of cloud service component. The model uses a second order hidden markov model to construct the QoS satisfaction predictive model. By considering the influence of the previous two states on the current state, the proposed method can effectively improve the prediction accuracy. Finally, in the Matlab simulation experiment environment, the effectiveness of the proposed method is prove by the prototype system and QWS data set with 2507 real web services.

Keywords Cloud service, Second order hidden markov model, Service selection

1 引言

云计算[1-5](Cloud Computing)是互联网的使用过程中相关服务种类的增加、使用和交付的一种模式,通常被应用于互联网中以获取相应的服务和资源,以满足自身的发展需求。用户可以通过手机、电脑等方式接入数据中心,按自己的需求进行运算。现阶段广为接受的是美国国家标准与技术研究院(NIST)给出的定义:云计算是一种按使用量付费的模式,这种模式提供可用的、便捷的、按需的网络访问,利用可配置的计算资源共享池(资源包括网络、服务器、存储、应用软件、服务等),这些资源能够以最省力和无人干预的方式获取和释放[6]。

云计算平台下的应用服务有五大基本特征:按需自助服务、无处不在的网络访问、资源池化、弹性负载、服务可计量,云计算的技术优势可以满足用户各种各样的个性化需求。在云计算环境下,云服务一般部署在分布式异构环境中,用户只需要通过浏览器在云端访问和使用云服务。文献[7]给出了

到稿日期:2018-07-25 返修日期:2018-10-26 本文受国家自然科学基金(61503036,61603054),辽宁省自然科学基金(20170540016)资助。
贾志淳(1982-),女,博士,副教授,主要研究方向为Web服务组合、故障诊断、云计算,E-mail:zhichun.jia@bhu.edu.cn(通信作者);李 想(1994-),女,硕士生,主要研究方向为云计算、Web服务组合;于湛麟(1963-),男,副教授,主要研究方向为数据挖掘;卢 元(1993-),女,硕士生,主要研究方向为云计算、Web服务组合;邢 星(1982-),男,博士,副教授,主要研究方向为社交网络挖掘、社会计算、推荐系统等。

在云计算环境下的服务系统架构。

云计算环境下的服务系统架构主要分为 3 层:软件层、平台层和基础设施层。用户通过浏览器向软件层中的代理发送云服务组合请求,使平台层的云服务组合引擎与代理根据用户请求进行通信,生成满足用户 QoS 需求的云服务组合执行计划,接着通过服务选择器从所有的组合云服务中选择满足用户 QoS 的最优组合云服务。基础设施层根据平台中云服务的组合执行计划来控制实际的资源分配。

由于用户体验对促进云计算发展起到关键作用,因此将云服务组件与用户的服务质量(QoS)相匹配是非常重要的。对于每个用户请求,如果单个云服务组件没有完全满足用户的 QoS,则提供者应该选择适当的云服务组合组件,以满足用户的需求。这就需要对云服务组合的组件的 QoS 满意度进行预测。

在云计算中的预测方法方面,文献[8]提出了一种基于组件偏好的贪婪方法来完成组件排序,该方法利用其他用户的使用经验,识别并聚合了两个组件之间的首选项,以生成组件的排名。文献[9]在协同过滤技术和客户满意度的估计方法的基础上,提出的一种创新的个性化云服务选择排名预测方法。文献[10]提出了适合于 QoS 感知的 Web 服务选择的改进粒子群优化算法(iPSOA)。文献[11]利用云服务的连续监测数据,提出了一种多值协同方法,通过对潜在用户的时间序列进行分析来预测未知的 QoS 值。

在隐马尔科夫模型的方法研究方面,文献[12]提出了一种隐马尔科夫概率模型,可预测 Web 服务的响应时间,并对服务进行定量排序,从而在运行时从功能等效的 Web 服务列表中选择一个最优的 Web 服务。文献[13]提出了一种基于响应时间 QoS 参数和隐马尔科夫模型选择 Web 服务的方法。文献[14]提出了利用隐马尔科夫模型(HMM)对 Web 服务状态预测机制进行研究,并指出 HMM 能够通过分析和识别长时间运行的 Web 服务生成的错误模式来预测 Web 服务的未来异常行为。文献[15]提出了融合攻击图模型和隐马尔科夫模型的方法,该方法能有效计算状态转移序列的最大概率,然后通过该对偶模型精确地推断攻击意图,为网络安全管理员提供了良好的配置。文献[16]针对云计算中多租户环境的竞争特性,提出了云计算环境下基于隐马尔科夫模型的云资源分配模型。

2 基本模型

隐马尔科夫模型(Hidden Markov Model)是一种应用很广泛的双重随机的统计概率模型,是由一个不可观测的马尔科夫链(称为状态过程)和与其每个状态相关联的表示观测结果的随机过程(称为观测过程)组成的整体。隐马尔科夫模型先前的理论和应用大多数是基于传统的一阶隐马尔科夫模型[17-20],即假定一个状态仅与前一个状态有关,且观测序列相互独立。基于这个假设可以推导出简单而有效的学习算法和识别算法。然而传统的隐马尔科夫模型是存在缺陷的,如一阶隐马尔科夫模型无法表示更远状态距离间的依赖关系,这样就忽略了很多有用的统计特征。

一个二阶隐马尔科夫模型可以由下列参数描述:

(1)N 为模型中 Markov 的状态数目,记 N 个状态为 s_1,s_2,…,s_N,t 时刻 Markov 链所处状态为 x_t,显然 $x_t \in (s_1, s_2, \cdots, s_N)$。

(2)M 为每个状态对应的可能观测值数目,记 M 个观测值为 v_1,v_2,…,v_M,t 时刻观测到的观测值为 O_t,其中 $O_t \in (v_1, v_2, \cdots, v_M)$。

(3)π 为初始状态概率矢量,$\pi=(\pi_1, \pi_2, \cdots, \pi_N)$,其中,$\pi_i = Pr(x_1 = s_i)$,$1 \leqslant i \leqslant N$。

(4)A_1,A_2 为状态转移概率矩阵,$A_1 = (a_{ij})_{N \times N}$,$A_2 = (a_{ijk})_{N \times N \times N}$,其中,$a_{ij} = Pr(x_{t+1} = s_j \mid x_t = s_i)$,$1 \leqslant i, j \leqslant N$,$a_{ijk} = Pr(x_{t+1} = s_k \mid x_t = s_j, x_{t-1} = s_i)$,$1 \leqslant i, j, k \leqslant N$。

(5)观测值概率矩阵 $B=(b_{jk})_{N \times M}$,其中,$b_{jk} = Pr(o_t = v_k \mid x_t = s_j)$,$1 \leqslant j \leqslant N$,$1 \leqslant k \leqslant M$,那么一个二阶隐马尔科夫模型可简记为 $\lambda=(N, M, \pi, A_1, A_2, B)$或简记为 $\lambda=(\pi, A_1, A_2, B)$。

3 云服务组件的 QoS 满意度预测模型

本文提出一个基于二阶隐马尔科夫的云服务组合预测模型,在该模型中加入某一时刻状态转移与历史状态的关联性,以提高预测的准确性。

3.1 预测模型训练

我们提出了一个基于二阶隐马尔科夫的 QoS 满意预测模型,用于预测云服务组件的组合是否满足用户的 QoS。目前研究的 QoS 属性因参数取值范围和表示意义不统一,很难在不作处理的基础上进行合理的比较,因此为了排除属性本身的限制条件,对服务的 QoS 属性进行归一化处理,其范围控制在[0,1]之间,具体归一化公式为:

$$\hat{x}_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

其中,$x_{\min}$ 和 $x_{\max}$ 分别为云服务组件的 QoS 属性 x 的最小值和最大值,x_i 为归一化之前的数值,$\hat{x}_i$ 为归一化之后的数值。

使用基本的二阶隐马尔科夫模型,假设云服务组件的 QoS 归一化后,$\hat{x}_i$ 为观测服务质量的状态序列 v_1,v_2,…,v_M。定义 $b_{kj} = Pr(O_t = V_l \mid x_t = s_j)$,其中,$1 \leqslant j \leqslant N$,$1 \leqslant l \leqslant m$。$b_{kj}$ 是观测服务质量值的概率,在 t 时刻所运行的服务组件是 s_j 的条件下,观测服务质量序列 O 的值是 V_l 的概率,其中 $V_l \in (v_1, v_2, \cdots, v_M)$。

为减少计算 $Pr(O|\lambda)$的复杂度,定义前向变量为:

$$\alpha_t(i,j) = Pr(O_1, O_2, \cdots, O_t, x_{t-i} = s_i, x_t = s_j \mid \lambda) \tag{2}$$

其中,$2 \leqslant t \leqslant T$,$1 \leqslant i, j \leqslant N$。根据动态规划的原理,有如下前向算法:

步骤 1 初始化

$\alpha_t(i,j) = \pi_i b_i(O_1) a_{ij} b_j(O_2)$,$1 \leqslant i, j \leqslant N$

步骤 2 递归迭代

$$\alpha_{t+1}(j,k) = \sum_{i=1}^{N} \alpha_t(i,j) a_{ijk} b_k(O_{t+1})$$

其中,$2 \leqslant t \leqslant T-1$,$1 \leqslant j, k \leqslant N$。

步骤 3 终结

$$Pr(O|\lambda) = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_T(i,j)$$

同理,定义后向变量为:

$$\beta_t(i,j) = Pr(O_{t+1}, \cdots, O_T \mid x_{t-i} = s_i, x_t = s_j, \lambda) \tag{3}$$

其中,$2 \leqslant t \leqslant T$。

因此,有如下后向算法:

步骤 1　初始化

$\beta_T(i,j)=1,1\leqslant i,j\leqslant N$

步骤 2　递归迭代

$$\beta_t(j,k)=\sum_{i=1}^{N}\beta_{t+1}(i,j)a_{ijk}b_i(O_{t+1})$$

其中 $2\leqslant t\leqslant T-1,1\leqslant j,k\leqslant N$。

步骤 3　终结

$$Pr(O|\lambda)=\sum_{i=1}^{N}\sum_{j=1}^{N}\beta_2(i,j)\pi_i b_i(O_1)a_{ij}b_j(O_2)$$

由式(2)和式(3)定义的前向和后向变量,有:

$$Pr(O|\lambda)=\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\alpha_t(i,j)a_{ijk}b_k(O_{t+1})\beta_{t+1}(j,k)\qquad(4)$$

其中,$2\leqslant t\leqslant T-1,1\leqslant j,k\leqslant N$。求取 λ 使 $Pr(O|\lambda)$ 最大,是一个泛函极值问题。由于给定的训练序列有限,因此不存在一个最佳方法来估计 λ。在这种情况下,Baum-Welch 算法[21]利用递归的思想,使 $Pr(O|\lambda)$ 局部最大,最终得到模型参数 $\lambda=(\pi,A_1,A_2,B)$。

定义 $a_t(i,j,k)$ 为给定训练序列 O 和模型 λ 时,Markov 链在时刻 $t-1$ 处于状态 s_i,在时刻 t 处于状态 s_j,在时刻 $t+1$ 处于状态 s_k 的概率,即:

$$a_t(i,j,k)=Pr(x_{t-1}=s_i,x_t=s_j,x_{t+1}=s_k|O,\lambda)\qquad(5)$$

可推导出:

$$a_t(i,j,k)=\frac{\alpha_t(i,j)a_{ijk}b_k(O_{t+1})\beta_{t+1}(j,k)}{Pr(O|\lambda)}\qquad(6)$$

那么,时刻 $t-1$ 以及时刻 t 时 Markov 链处于状态对 s_i,s_j 的概率为:

$$\begin{aligned}a_t(i,j)&=Pr(x_{t-i}=s_i,x_t=s_j|O,\lambda)\\&=\sum_{k=1}^{N}a_t(i,j,k)=\frac{\alpha_t(i,j)\beta_t(i,j)}{Pr(O|\lambda)}\end{aligned}$$

因此,$\sum_{t=2}^{T-1}a_t(i,j)$ 表示从状态对 s_i,s_j 转移出去的次数的期望值,而 $\sum_{t=2}^{T-1}a_t(i,j,k)$ 表示从状态对 s_i,s_j 转移到状态 s_k 的次数的期望值。由此推导出 Baum-Welch 算法中的重估公式:

$$\pi_i^*=\sum_{j=1}^{N}\sum_{k=1}^{N}a_2(i,j,k)\qquad(7)$$

$$a_{ij}^*=\frac{\sum_{k=1}^{N}a_2(i,j,k)}{\sum_{j=1}^{N}\sum_{k=1}^{N}a_2(i,j,k)}\qquad(8)$$

$$a_{ijk}^*=\frac{\sum_{t=2}^{T-1}a_t(i,j,k)}{\sum_{t=2}^{T-1}a_t(i,j)}\qquad(9)$$

$$b_{kj}^*=\frac{\sum_{t=2,O_t=V_t}^{T-1}\sum_{i=1}^{N}\sum_{j=1}^{N}a_t(i,j,k)}{\sum_{i=1}^{N}\sum_{j=1}^{N}a_t(i,j,k)}\qquad(10)$$

我们获得一个新的模型 $\lambda^*=(\pi^*,A_1^*,A_2^*,B^*)$,这里 $\pi^*=(\pi_i^*),A_1^*=(a_{ij}^*),A_2^*=(a_{ijk}^*),B^*=(b_{kj}^*)$。为了证明 $Pr(O|\lambda^*)\geqslant Pr(O|\lambda)$,重复训练模型,逐步调整模型参数,直到 $Pr(O|\lambda^*)$ 收敛,最后,得到预测模型 λ^*,用于评估云服务组件组合的 QoS 满意能力。

3.2　组合服务选择

在云环境中,用户需要可随时获取、服务质量满足需求、低成本、灵活且易于使用的云服务。在云服务组合组件的集合中,有大量功能相似的组合云服务组件,因此需要对组合的云服务的 QoS 满意度进行预测,然后将满足 QoS 需求的组合服务提供给用户。如图 1 所示。

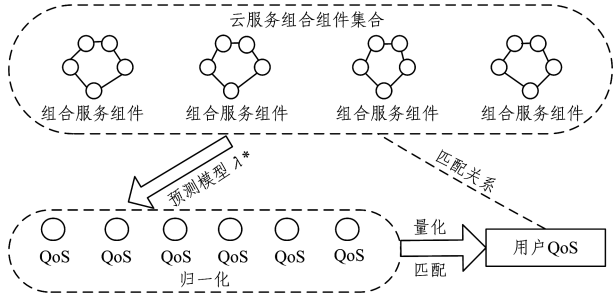

图 1　云服务组合选择模型

Fig. 1　Cloud service combination selection model

4　实验与分析

为了验证所提出的 QoS 满意预测模型的可行性,本文开发了一个支持该工作的原型系统,部署在服务器集群上。服务器集群由 3 台 IBM x3850 M2 服务器组成,4 个 Xeon E7320 2.13 GHz 内核和 8GB RAM。在云环境中设置 6 个虚拟机,其运行 Windows XP 操作系统。所有虚拟机都配置了 1 GB 内存和 1.2 GHz CPU。

4.1　原型系统

图 2 展示了扩展 Java 代理开发框架(JADE)原型系统的主要架构。JADE 是由一组称为代理的组件组成,它们通过交换消息来执行任务并相互作用。我们的原型系统由组合和预测服务组成。该组合包括目录文件(DF)、协调器代理(CA)、BPEL 和 AEL 文件。DF 为发布代理提供了一个黄页服务,所有代理必须在 DF 中注册。CA 用于执行诸如启动和处理代理等管理操作。它负责从 BPEL 文件中读取组合信息,组织一个分布式工作流,并在 AEl 文件中记录执行信息。CA 读取来自 DF 的代理的寄存器信息,并调用这些代理进行组合。这些代理节点分布在不同的虚拟机上。最后,使用推出的预测模型对组合服务组件的 QoS 满足能力进行评估预测。

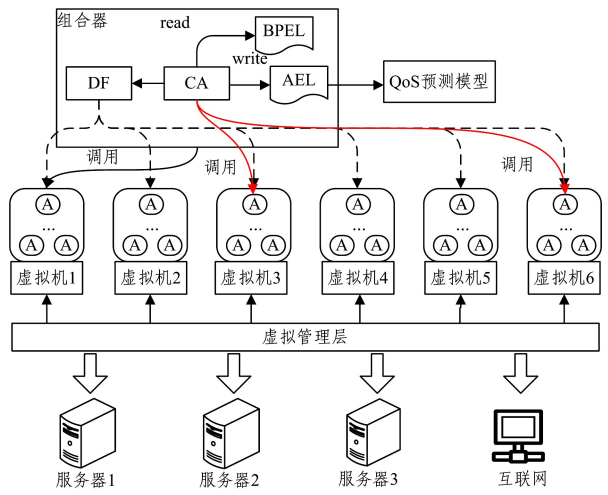

图 2　扩展 Java 代理开发框架(JADE)原型系统的主要架构

Fig. 2　Main architecture of prototype system for extending Java agent development framework

4.2 数据分析

为了验证 QoS 满意预测模型具有可行性，使用 QWS 数据集[22]在 Matlab 环境中进行仿真实验。QWS 收集了网络中 2507 个真实的 Web 服务，每个服务包含 11 个属性。

在仿真实验中，随机从 QWS 中选取 100 个服务，并对服务的 QoS 属性进行归一化处理。然后，将处理后的数据作为输入值在 Matlab 中运行 Baum-Welch 算法，对预测模型进行训练，从而得到二阶隐马尔科夫模型预测模型。利用该预测模型，针对天气查询系统和飞机购票系统两个实例进行预测，并获得仿真结果。将二阶预测模型、一阶预测模型的预测值和两个实例的实际 QoS 满意度数据进行对比，结果如表 1 所列。不难看出，在天气查询系统中，一阶隐马尔科夫模型和二阶隐马尔科夫模型与实际值的差值分别为 0.035 和 0.012。在飞机购票系统中，一阶隐马尔科夫模型和二阶隐马尔科夫模型与实际值的差值分别为 0.022 和 0.003。综上得出，在原来的基础上加入某一时刻状态转移与历史状态的关联性，可使二阶隐马尔科夫预测模型的预测值更接近实际值。该仿真实验验证了本文给出的 QoS 满意预测模型的合理性以及有效性。

表 1 模型对比数据

Table 1 Model contrast data

	实际值	一阶隐马尔科夫模型	二阶隐马尔科夫模型
天气查询系统	0.367	0.332	0.355
飞机购票系统	0.598	0.576	0.601

结束语 本文提出了一个基于二阶隐马尔科夫的云服务 QoS 满意度预测模型。为了提高 QoS 满意预测模型的准确度，本文在原来的基础上加入某一时刻状态转移与历史状态的关联性，把二阶隐马尔科夫算法应用到 QoS 满意预测模型中，来预测组合服务的 QoS 满意度。最后，本文通过仿真实验验证了所提出的 QoS 满意预测模型的合理性以及有效性。仿真结果表明，该模型具有较高的预测精度。

参 考 文 献

[1] BOKHARI M U, MAKKI Q, TAMANDANI Y K. A Survey on Cloud Computing [J]. International Journal on Computer Science & Engineering, 2018, 5(6): 302-311.

[2] SHARMA P, SOOD S K, KAUR S. Security Issues in Cloud Computing [J]. International Journal of Advanced Computer Research, 2017, 2(3): 1-11.

[3] ÅŽICLOVAN A, PATRUT B. The Future of Cloud Computing [J]. Brand Broad Research in Accounting Negotiation & Distribution, 2016, 3(3): 47-68.

[4] COMUZZI M, JACOBS G, GREFEN P. Understanding SLA Elements in Cloud Computing [J]. Ifip Advances in Information & Communication Technology, 2017, 408: 385-392.

[5] SINGH S, CHANA I, SINGH M. The Journey of QoS-Aware Autonomic Cloud Computing [J]. IT Professional, 2017, 19(2): 42-49.

[6] MELL P M, GRANCE T. The NIST Definitionof Cloud Computing[C]// National Institute of Standards & Technology. Gaithersburg, 2011.

[7] WANG D, YANG Y, MI Z. A genetic-based approach to web service composition in geo-distributed cloud environment [J]. Computers & Electrical Engineering, 2015, 43(C): 129-141.

[8] ZHENG Z, ZHANG Y, LYU M R. CloudRank: A QoS-Driven Component Ranking Framework for Cloud Computing[J]. Proceedings of the IEEE Symposium on Reliable Distributed Systems, 2010, 23(3): 184-193.

[9] DING S, WANG Z, WU D, et al. Utilizing customer satisfaction in ranking prediction for personalized cloud service selection [M]. Elsevier Science Publishers B. V., 2017.

[10] SUN Q. An improved Particle Swarm Optimization Algorithm for QoS-aware Web Service Selection in Service Oriented Communication[J]. International Journal of Computational Intelligence Systems, 2010, 3(sup1): 18-30.

[11] MA H, ZHU H, HU Z, et al. Multi-valued collaborative QoS prediction for cloud service via time series analysis[J]. Future Generation Computer Systems, 2016, 68: 275-288.

[12] AHMED W, WU Y, ZHENG W. Response Time based Optimal Web Service Selection[J]. IEEE Transactions on Parallel & Distributed Systems, 2015, 26(2): 551-561.

[13] CANTÓNPUERTO D G, MOOMENA F, UCCETINA V. QoS-Based Web Services Selection Using a Hidden Mar-kov Model [J]. Journal of Computers, 2017, 12(1): 48-56.

[14] PRASANGA R, WIJESIRIWARDANA C, WEERASURIYA G, et al. States Prediction of Web Services Using Hidden Markov Model [C]// ITRU Research Symposium. Sri Lanka, 2015: 16-20.

[15] LIU S, LIU Y. Network security risk assessment method based on HMM and attack graph model [C]// IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/distributed Computing. IEEE, 2016: 517-522.

[16] WEI W, FAN X, SONG H, et al. Imperfect Information Dynamic Stackelberg Game Based Resource Allocation Using Hidden Markov for Cloud Computing [J]. IEEE Transactions on Services Computing, 2018, PP(99): 1-1.

[17] BAUM L E, PETRIE T. Statistical Inference for Probabilistic Functions of Finite State Markov Chains[J]. Annals of Mathematical Statistics, 1966, 37(6): 1554-1563.

[18] RABINER L R. A tutorial on hidden Markov models and selected applications in speech recognition [J]. Proceedings of the IEEE, 1989, 7(2): 257-286.

[19] XU L, FAN W, SUN J, et al. An HMM-based Over-Segmentation Method for Touching Chinese Handwriting Recognition [C]// International Conference on Frontiers in Handwriting Recognition. IEEE, 2017: 343-348.

[20] GHARACHEH M, DERHAMI V, HASHEMI S, et al. Proposing an HMM-based approach to detect metamorphic malware [C]// Fuzzy and Intelligent Systems. IEEE, 2016: 1-5.

[21] MATUZ B, BLASCO F L, LIVA G. On the Application of the Baum-Welch Algorithm for Modeling the Land Mobile Satellite Channel[C]// 2011 IEEE Global Telecommunications Conference. Kathmandu, Nepal, 2011, 44: 1-5.

[22] AL-MASRI E, MAHMOUD Q H. Investigating Web Services on the World Wide Web [C]// International Conference on World Wide Web. Beijing, China, 2008: 795-804.