基于性能感知预测的云服务推荐模型①

汪佳祯, 迟焕醒, 王木涵, 史双田

(中国石油大学(华东) 计算机与通信工程学院, 青岛 266000)

摘　要: 互联网上出现越来越多的云服务, 面对种类繁多的云服务, 如何准确地在众多云服务中把符合用户需求并且性能好价格低的服务推荐给用户成为云服务推荐的研究热点. 现有的服务推荐方法往往只是根据当前云服务的历史性能记录为用户进行推荐, 并没有充分考虑云服务的性能趋势. 针对上述问题, 本文提出了一种基于性能预测的服务推荐模型, 该模型利用共轭梯度改进人工神经网络对云服务的性能进行预测, 使用层次分析法对性能, 价格等因素进行综合比较计算, 最终为用户推荐最为合适的云服务. 实验结果表明, 使用改进神经网络对服务性能进行预测能够获得较高的准确度, 层次分析法可以综合考虑服务的性能与价格, 为用户推荐最为合适的云服务.

关键词: 云服务; 性能感知; 神经网络; 层次分析法

Cloud Service Recommendation Model Based on Performance Prediction

WANG Jia-Zhen, CHI Huan-Xing, WANG Mu-Han, SHI Shuang-Tian

(School of Computer & Communication Engineering, China University of Petroleum, Qingdao 266580, China)

Abstract: A growing number of cloud services have emerged on the Internet. In the face of a wide variety of cloud services, how to recommend high quality and low price service meeting the users' requirements to the user accurately has become a focus in cloud service recommendation field. Currently, many services recommendation methods are often just based on the current service status without taking into account the performance trend of cloud service. For this reason, this paper proposes a services recommendation model based on performance prediction. The model uses improved artificial neural network based on conjugate gradient to predict the performance of cloud services. Factors such as performance and prices can be compared and calculated by using AHP (Analytic Hierarchy Process), and then the most suitable cloud service would be recommended to the users. The experimental results show that the prediction accuracy would be higher by using improved neural network predicting service performance method, and AHP can recommend the most suitable service to the user according to comprehensively considering the performance and price of services.

Key words: cloud service; performance aware perception; neural network; analytic hierarchy process

1　引言

目前, 云服务以其"按需使用"的方式降低信息处理和存储投资为广大中小型企业所欢迎, 这意味着计算能力也可以作为一种商品通过互联网进行流通[1]. 随着云服务的发展, 越来越多的用户也开始使用云服务. 用户在使用云服务时, 不仅关注服务的功能, 同时也关注服务的性能及价格等问题. 例如在 IaaS 基础设施即服务(Infrastructure as a Service)中, 服务提供商会根据用户提出的不同服务性能需求进行收费. 在进行云服务推荐时, 主要根据云服务的性能进行推荐, 同时也适当考虑价格因素对推荐结果的影响.

已有的云服务推荐方法在处理服务性能时, 会根据服务注册中心的数据进行推荐[2], 并没有考虑到服务性能的动态变化. 例如云服务 A 性能比较高, 服务推荐系统会推荐用户使用该云服务, 那么随着用户的增多, 该服务的性能会显著下降; 如果继续向该服务推

① 收稿时间:2016-09-01;收到修改稿时间:2016-10-10　[doi:10.15888/j.cnki.csa.005755]

荐用户则会造成后续用户的使用体检降低. 现有的云服务推荐系统并没有考虑到云服务的性能由于使用该服务的用户数量变化而引起服务性能的变化. 针对该问题本文提出了基于性能预测的服务推荐模型, 该模型首先利用改进神经网络对服务性能进行预测, 其次通过层次分析法计算出性价比最高的服务推荐给用户.

本文安排如下: 第二节提出基于性能预测的云服务推荐模型, 对模型中的各个部分进行说明, 介绍了每一部分在服务推荐计算中的作用及步骤; 第三节针对模型中的关键部分及关键算法进行相关实验; 第四节分析实验结果, 验证本文提出的方法的有效性. 第五节对本文推荐方法进行总结; 第六节总结本文工作并对后续工作进行了展望.

2 基于性能预测的服务推荐模型

2.1 用户及服务数据收集

对用户及服务数据进行收集是性能预测的前提[3]. 对用户数据的收集包括用户端的监测数据: 响应时间, 花费等; 对服务的数据主要包括服务端 CPU、内存、硬盘、网络吞吐量的的运行情况及用户数量等信息. 通过获取这些数据, 可以分析出服务使用数量与服务性能之间的关系, 从而为服务推荐提供参考.

2.2 基于性能感知的服务推荐模型

云服务性能指的是一个云服务响应处理用户请求的能力[4]. 在基于性能感知的服务推荐模型中, 用户根据自身需求提出服务申请, 服务资源计算模块计算所需服务资源; 服务资源列表中存储各个服务的实时资源状态用于性能预测; 性能预测模块预测服务性能并将结果交予层次分析系统; 层次分析系统给出服务推荐序列, 同时更新服务资源列表. 图 1 为基于性能感知的云服务推荐模型的结构图.

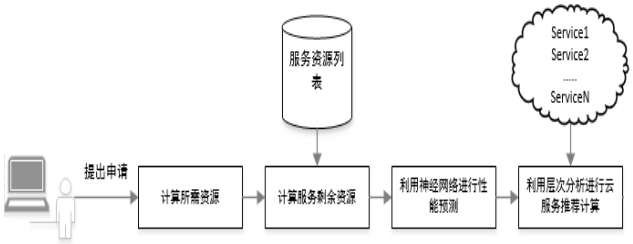

图 1 基于性能感知的云服务推荐模型

该模型主要包括三个部分:

1) 服务资源列表. 以 4 元组的形式存储存储服务资源使用剩余情况, 格式为: <CPU_Count, BandWidth , Mem_size, HardDisk>分别表示 CPU 使用率(%)、网络带宽(Mb/s)、内存大小(GB)及硬盘(GB). 需要进行性能预测计算时, 从服务资源列表中获取服务资源信息, 计算推荐完毕后根据服务推荐情况更新列表.

2) 性能预测: 根据历史记录训练改进 BP 神经网络, 根据服务资源剩余进行性能预测并将预测结果交予服务计算推荐模块.

3) 服务推荐计算: 利用层次分析法, 通过性能构造比较矩阵并计算权重, 得到服务推荐优先级列表并为用户推荐.

2.3 改进神经网络算法

服务的性能主要与服务资源有关(CPU, 内存, 带宽等), 由于性能与资源之间的关系是非线性的[5],难以建立一个确定的映射关系, 而神经网络的优点在于能够利用已有的历史数据不断训练调整, 可以满足预测要求, 预测结果够最大程度接近实际情况. 除上述优点外, 神经网络也有自身缺点: 可能陷入局部最小、训练时间长、网络冗余大. 为了减少训练时间, 提高训练精度, 本文采用共轭梯度法来优化神经网络. 模型拓扑结构包括输入层(input), 隐层(hide layer)和输出层(output layer), 如图 2 所示.

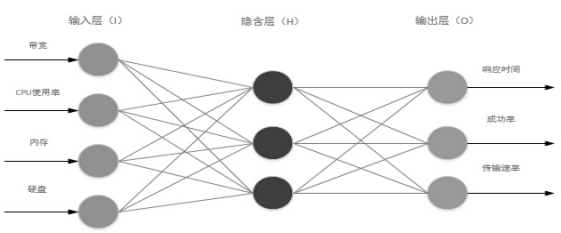

图 2 神经网络结构图

在输入层, 输入参数分别为 CPU 使用率(%), 网络带宽(Mb), 内存大小(GB)以及硬盘大小(GB), 输出为响应时间(S), 服务调用成功率(%),传输速 β_k 率(Kb/s), 隐含层数为 3 层.

使用共轭梯度法改进后的神经网络搜索方向为上一次搜索方向与负梯度方向共轭方向.

设 g_0 为梯度方向, p_0 为它的负方向, 有:

$$x_{i+1} = x_i + \alpha_i p_i$$

选取共轭方向作为新的搜索方向, 在当前梯度的负方向上增加一次搜索方向:

$$p_k = -g_k + \beta_k p_{k-1}$$

对于修正系数的选取步骤如下:

1) 仿真次数设为 k, 随机产生 n 维权向量 ω_k, 尺度因子 λ_b 取 0. 影响参数 σ 及调整参数 λ 为 5.0×10-5 和 5.0×10-7.

$$W_{ij}(t+1) = W_{ij}(t) - \eta * d_i^k * X_i^{k-1} + \alpha \Delta W_{ij}(t)$$

$$\Delta W_{ij}(t) = W_{ij} - W_{ij}(t-1)$$

2) 令 k 循环自增, 有

$$\sigma_k = \frac{\sigma}{|p_k|},\quad s_k = \frac{G'(\omega_k + \sigma_k p_k) - G'(\omega_k)}{\sigma_k},\quad \delta_k = p'_k s_k$$

3) 对尺度做调整

$\delta_k = \delta_k + (\lambda_k - \lambda_b) \bullet |p_k|^2$, 若不大于零, 有

$$\delta_k = 2(\lambda - \frac{2\delta_k}{|p_k|^2}),\quad \delta_k = -\delta_k + \lambda_k |p_k|^2,\quad \lambda_k = \lambda_b .$$

4) 计算评价参数

$$D_k = \frac{2\delta_k [G(\omega_k) - G(\omega_K + \partial_K P_K)]}{\mu_k^2}$$

其中 $\mu_k = p_k g_k$ 是步长, $\alpha_k = \mu_k \div \delta_k$. 当评价参数大于等于零时, $\omega_{k+1} = \omega_k + \alpha_k p_k$. 假如 mod$(k,n)$ =0, 按原来梯度方向重新计算, 否则 $\beta_{k+1} = \frac{(|g_{k+1}|^2 - g'_{g+1} g_k)}{\mu_k}$. 若评价参数大于 θ_1, 减小尺度因子从而减少误差; 若评价参数小于 θ_2, 增加尺度因子; 如此计算最终得到学习结果.

2.4 服务推荐计算

层次分析法把研究对象作为一个系统, 按照分解, 比较判断, 综合的思维方式进行决策, 是系统分析的重要工具, 非常适合解决多目标决策问题[6], 因此本文选定该方法解决服务推荐问题.

2.4.1 层次分析法在服务推荐计算中的步骤

1) 建立层次模型结构: 目标层为最优服务; 准则层为服务性能指标: 响应时间, 传输速率, 服务调用成功率, 价格; 方案层为备选服务, 构造层次模型如图 3.

2) 构造比较矩阵: 若服务性能指标数为 i, 那么备选服务对应服务性能指标的比较矩阵有 i 个, 分别 $A_1, A_2, ..., A_j$ 每个比较矩阵大小为 j(j 是备选服务数). A_r 对应的是服务性能指标 r, 那么 A_r 的最大特征值对应的归一化特征向量就是全体备选服务相对于服务性能指标 r 的权重向量, 那么权重矩阵 $W_{All} = [W_1, W_2 .. W_i]$ 就是全体备选服务在全体服务指标下的权重向量集合.

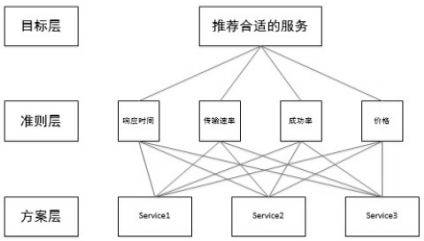

图 3　层次模型图

3) 计算单排序权向量并做一致性检验;

4) 计算总排序权向量并做一致性检验;

5) 根据服务性能指标的权重向量 W_j 和备选服务的权重矩阵 W_{All}, 得到备选服务的综合值 $W_j = W_{All} * W_p^T$ 最终推荐值最大的服务 Max_{wj}

2.4.2 服务推荐实例

服务 1: 调用成功率 94%; 平均响应时间 4.141 秒; 传输速率 0.985Mbps, 收费 0.33 元/时

服务 2: 调用成功率 99.2%; 平均响应时间 11.043 秒; 传输速率 0.209Mbps,收费 0.51 元/时

服务 3: 调用成功率 99.2%; 平均响应时间 2.9 秒; 传输速率 0.02Mbps, 收费 0.4 元/ 时

构造比较矩阵: $\begin{pmatrix} 1 & 2 & 1/2 & 1/4 \\ 1/2 & 1 & 1/5 & 1/2 \\ 2 & 5 & 1 & 2 \\ 4 & 2 & 1/2 & 1 \end{pmatrix}$ 通过计算我们求得这个矩阵的最大特征值 λ =4.256. CI= $\frac{4.256-4}{4-1}$ =0.085. 查阅数据可知该四阶矩阵对应 RI=0.90. 判断矩阵一致性 CR= $\frac{C_I}{R_I}$ =0.094<0.1 通过检验, 对应归一化权重向量为 W^T ={-0.1515, 0.0983,0.4366, -0.3136}(考虑到响应时间, 费用越小越好)

比较矩阵为 $\begin{pmatrix} 1 & 0.35 & 1.4279 \\ 2.667 & 1 & 3.8079 \\ 0.7003 & 0.2626 & 1 \end{pmatrix}$ 权重向量 W_T^T ={0.225,0.6138,0.1612} 同理可得 W_K^T ={0.7895,0.1052,0.1052} W_F^T ={0.791,0.1678,0.0412} W_{pri}^T ={0.2661,0.4113,0.3226}

则权重向量构成的矩阵为 $W_T^T W_F^T W_K^T W_{pri}^T$, $W_T^T W_F^T W_K^T W_{pri}^T \times W^T = (0.1502, 0.1591, -0.1180)^T$, 该计算结果即为服务推荐序列, 值越大则推荐优先度最高, 因此向用户推荐服务 2.

3 实验

3.1 实验准备

为了检验本文中改进神经网络算法对服务性能预测的准确性, 本实验通过在亚马逊弹性计算云(Amazon elastic computer cloud,Amazon EC2)上运行SEPCWeb2009 软件, 监测 EC2 运行的 CUP 使用率等服务性能变化情况. SEPCWeb2009 是由标准性能评估公司(Standard Performance Evaluation Corpiration, SPEC)开发的软件基准测试软件, 用于测试 Web 服务器的静态和动态页面响应能力. 本文选择激活函数为 purelin 函数, 训练函数 trainlm 函数, 训练目标误差 0.000001, 最大迭代次数 5000, 学习率 0.1.

3.2 BP 神经网络训练与验证

一共进行 1000 组测试, 选取前 900 组对神经网络进行训练, 取最后 100 组作对照试验. 试验结果如图 4-图 6 所示.

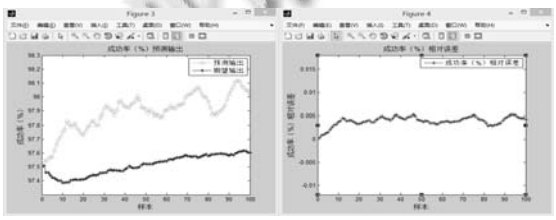

图 4 服务调用成功率预测对比图

图 4 是服务调用成功率的预测结果, 左图是实际预测调用成功率和期望预测调用成功率的对比图, 右图为预测误差.

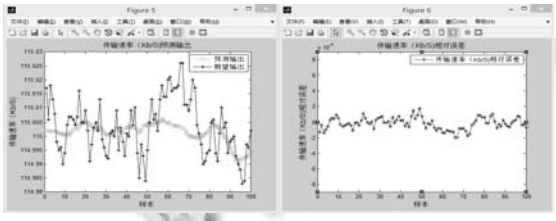

图 5 传输速率预测对比图

图 5 是传输速率的预测结果, 左图是实际预测传输速率和期望预测传输速率的对比图, 右图为预测误差.

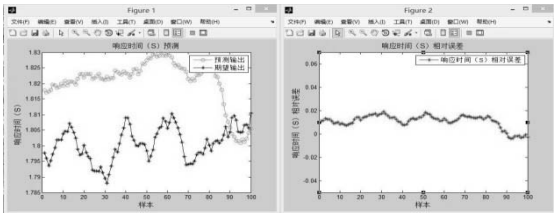

图 6 响应时间预测对比图

图 6 是响应时间的预测结果, 左图是实际预测响应时间和期望预测响应时间的对比图, 右图为预测误差.

3.3 层次分析法效果验证

本文在 4 台电脑上各自部署一个服务, 同一个用户申请 8 次相同服务, 对服务性能进行预测后利用层次分析法计算推荐优先级队列, 情况如下:

	第一次	第二次	第三次	第四次	第五次	第六次	第七次	第八次
服务 1	0.7147	0.5693	0.201	0.201	0.201	-1.264	-1.264	-1.264
服务 2	0.3114	0.3114	0.3114	-0.087	-0.087	-0.087	-3.1751	-3.1757
服务 3	0.2205	0.2205	0.2205	0.2205	-1.2452	-1.2452	-1.2452	-1.2452
服务 4	-0.1864	-0.1864	-0.1864	-0.1864	-0.1864	-0.1864	-0.1864	-4.932

服务 1 共推荐 3 次, 服务 2 推荐 2 次, 服务 3 推荐 2 次, 服务 4 推荐 1 次, 服务推荐队列变化情况的发生是由于服务性能随着用户的使用而不断变化, 实验证明该方法可以根据服务资源变化而导致服务性能波动的情况为用户推荐出当前状态下最优的服务, 从而避免因为推荐用户过多而导致服务性能下降, 用户使用体验降低情况的发生, 有效提高使用体验; 同时, 层次分析法还将性能与价格结合在一起计算, 符合云服务推荐的要求.

4 实验结果及分析

通过上述实验我们可以发现, 使用改进神经网络对服务性能进行预测, 误差波动范围比较小, 在实际应用中增加训练样本的数量还可以进一步提高预测精度, 该实验结果证明该方法可以用来对服务性能进行预测.

层次分析法在模拟服务推荐时, 当有新的用户被推荐至某服务, 服务推荐队列随之发生变化, 证明该方法可以在服务推荐计算时有效考虑服务性能的波动, 动态向用户推荐服务, 从而避免服务性能下降而导致用户使用体验降低情况的发生.

综上, 本文提出的基于性能预测的云服务推荐模型可以对云服务性能进行准确预测, 并根据预测结果为用户进行云服务推荐.

5 相关工作

目前, 针对云服务推荐问题国内外学者展开了深入研究, 并取得了一定成果. 文献[7]针对服务推荐过程中性能预测不准确的问题, 通过使用排队论的方法建立服务性能预测模型以此来预测性能, 提出一种优

化的性能预测模型. 但是该模型主要考虑 CPU 与性能的关系而忽略网络、带宽以及硬盘等对于服务性能的影响.

文献[8]使用相似度计算来计算各服务的性能, 根据结果为用户进行服务推荐. 该方法使用相似度系数用于描述数据性能之间的线性相关关系, 在实际并不能很好地解释客观性能数据之间的相似性.

文献[9]提出云服务选择框架, 该框架根据候选服务的服务质量及虚拟机的模拟参数做为参考为用户进行云服务推荐. 该方法只是根据服务注册信息进行模拟计算推荐, 并没有考虑到性能的动态变化情况. 文献[10]提出一种基于信任的云服务推荐系统, 该系统分别计算云服务的直接和推荐信任度, 根据信任度的计算结果为用户推荐最可信的服务.

上述研究在进行云服务推荐时都根据性能预测结果为用户进行服务推荐[1], 但是并没有考虑到服务性能变化的动态性, 使用上述推荐系统会出现由于用户增加而导致的服务性能严重下降情况的发生(服务调用丢失率高, 响应时间长, 传输速率慢等). 针对现有推荐模型的不足, 本文使用通过共轭梯度法改进后的神经网络来动态预测服务性能; 在进行服务推荐时, 综合考虑多种因素, 使用权重向量来代表每一个具体影响因素, 向量的不同值代表不同因素对于推荐的影响程度[12].

6 总结

本文针对服务性能动态变化的问题, 提出一种基于性能感知预测的云服务推荐模型, 该模型使用改进人工神经网络对服务性能进行预测, 采用层次分析法计算推荐优先级. 通过试验证明人工神经网络预测法对于不同资源环境下的服务性能有较高的预测准确率, 层次分析法可以根据服务性能变化而改变服务推荐优先级. 本文主要不足是在层次分析法中构建权重矩阵时, 各因素的权重向量值的确定方法相对固定, 没有根据每个用户的实际偏好量身定制, 这是后面工作应解决的问题.

参考文献

1 Gmbh ICE. Cloud-Managed Wi-Fi Market to Reach $2.5 Billion by 2018, IDC Says. Wifi Wlan, 2014.

2 Garcia DF, Garcia J, Entrialgo J, et al. A QoS control mechanism to provide service differentiation and overload protection to Internet scalable servers. IEEE Trans. on Services Computing, 2009, 2(1):3–16.

3 Tsesmetzis D, Roussaki I, Sykas E. QoS-aware service evaluation and selection. European Journal of Operational Research, 2008, 191(3):1101–1112.

4 Jiang D, Pierre G, Chi CH. Autonomous resource provisioning for multi-service web applications. International Conference on World Wide Web, WWW 2010. Raleigh, North Carolina, USA. April, 2010. 471–480.

5 Lorenzi L, Mercier G, Melgani F. Support vector regression with kernel combination for missing data reconstruction. IEEE Geoscience & Remote Sensing Letters, 2013, 10(2): 367–371.

6 Shao L, Zhang J, Wei Y, et al. Personalized QoS prediction for Web services via collaborative filtering. IEEE International Conference on Web Services. 2007. 439–446.

7 Han SM, Hassan M, Yoon CW, et al. Efficient service recommendation system for cloud computing market. International Conference on Interaction Sciences: Information Technology, Culture and Human. ACM. 2009. 839–845.

8 Zeng L, Benatallah B, Dumas M, et al. Quality driven web services composition. Proc. of the 12th International Conference on World Wide Web. ACM. 2003. 411–421.

9 Köksalan M, Zionts S. Multiple criteria decision making in the new millennium. Lecture Notes in Economics & Mathematical Systems, 2001, 31(5): 358.

10 Yao X, Liu Y. A new evolutionary system for evolving artificial neural networks. IEEE Trans. on Neural Networks, 1997, 8(3): 694–713.

11 Kong D, Zhai Y. Trust based recommendation system in service-oriented cloud computing. International Conference on Cloud and Service Computing. 2012. 176–179.

12 Kant K, Won Y. Server capacity planning for web traffic workload. IEEE Trans. on Knowledge & Data Engineering, 1999, 11(5): 731–747.