基于神经网络改进的云环境下暴发式请求部署策略研究

陈 鹏[1] 马自堂[1] 孙 磊[1] 孙冬冬[2]
(解放军信息工程大学三院 郑州 450004)[1] (61579部队 北京 102400)[2]

摘 要 针对暴发式任务请求给云计算系统性能带来的影响,结合现有资源部署模型,提出了一种基于误差反向传播神经网络改进的资源部署模型来应对上述问题。模型判断出暴发式任务请求的始末时,自动启动网络模块,通过事先训练好的网络进行参数调整值的预测,以达到动态跟踪云计算系统底层资源与外界任务请求变化的目的。通过CloudSim对模型进行了仿真实验,结果证明,引入神经网络模块可有效提高现有系统的资源部署响应速度。

关键词 云计算,神经网络,资源部署,暴发式任务请求
中图法分类号 TP302.7 **文献标识码** A

Deployment Strategies Research on Cloud Computing under Bursty Workloads on Neural Network

CHEN Peng[1] MA Zi-tang[1] SUN Lei[1] SUN Dong-dong[2]
(The Third Institute, PLA Information Engineering University, Zhengzhou 450004, China)[1]
(61579 Army, Beijing 102400, China)[2]

Abstract Aiming at the degrading system performance that bursty workloads bring in cloud computing, a resource deployment model based on error back-propagation neural network was proposed to resolve the problems referred to above. A network module is started automatically when the beginning of bursty workloads is judeged. The prediction of parameter adjustment value is carried out by using pre-trained network to achieve the purpose of tracking dynamically the changing of underlying resource and outside world task in cloud computing system. The results of simulation in CloudSim prove that the response speed of resource deployment can be improved efficiently by bringing neural network module.

Keywords Cloud computing, Neural network, Resource allocation, Bursty workloads

1 引言

随着云计算在各个领域的不断拓展与深入,其发展也面临着各方面因素的考验。暴发式的任务请求(Bursty Workloads,BW)对云计算系统瞬时的响应性能提出了很高的要求[1],其中,暴发式任务请求可描述为在相对较短的时间内,存在着海量的并发式任务请求,瞬间造成队列堵塞,使云计算系统无法满足与用户签订的SLA协议。在应对暴发式任务请求给云计算系统带来的挑战时,美国惠普实验室提出了一种应对暴发式任务请求的资源部署算法Fastrack[2],该算法为云计算系统提供了从跟踪暴发式任务请求到相应的资源部署策略的应对模式,但其资源部署算法的自适应较差。而美国东北大学的Waleed Meleis等人提出的ARA模型[3]中的资源部署算法online ARA虽然可根据可用资源节点数,动态地调整资源部署策略,但其只是简单的线性变换,无法很好地拟合云计算系统资源部署问题非线性的特征。以上方法虽然对云计算环境下暴发式任务请求的跟踪、部署问题的解决取得了一定的效果,但要更加灵活有效地解决暴发式任务请求带来的对系统性能的冲击还需要采用一种更加高效、自适应性更强的方法。

由于人工神经网络在非线性映射、自学习和自适应方面拥有特殊优势[4],针对云计算环境下暴发式任务请求动态变化的特点,人工神经网络可有效解决此类问题。人工神经网络与人脑类似,是由海量的简单处理单元互联构成高度并行的非线性系统。它在结构与处理顺序上是并行和同时的,这可使神经网络拥有更快的计算速度,满足云计算快速响应请求的要求,而其自学习、自适应特性使系统能够改变自身的性能以适应环境变化的能力,通过接受外部输入的不断刺激来获取并积累知识,进而拥有很强的判断预测能力。误差反向传播(Error Back Propagation, BP)神经网络模型相比Hopfield网络、径向基函数网络等神经网络在训练时长方面虽有一定的劣势,但其在执行时间、判断信息的复杂度、信息容量和实用性方面具有优势[5],可很好地切合本文所面临的挑战。

本文提出了一种基于BP神经网络学习算法改进的资源部署模型来应对云计算环境下暴发式任务请求带来的挑战。针对现有模型无法根据云计算系统底层资源变化动态地做出调整等问题,通过引入反向传播学习算法,使系统在应对暴发

到稿日期:2013-04-21 返修日期:2013-08-10 本文受武器装备预研重点基金资助项目(9140A15060311JB5201)资助。
陈 鹏(1988-),男,硕士生,CCF学生会员,主要研究方向为云计算、系统优化,E-mail:skyskyasd@163.com;马自堂(1962-),男,教授,主要研究方向为信息安全、密码系统工程;孙 磊(1973-),男,博士,副研究员,主要研究方向为云计算基础设施可信增强、可信虚拟化技术;孙冬冬(1975-),女,讲师,主要研究方向为计算机软件应用。

式任务请求的同时可根据云计算系统实时的可用资源节点情况做出调整，进一步提升云计算系统对任务请求的响应速度，进而提高用户体验。

2 现有 BW 环境下资源部署模型

提供高质量的服务是云计算服务的核心目标，暴发式任务请求环境对云计算系统的冲击及对系统带来的影响严重降低了用户体验。现有的应对暴发式任务请求资源部署模型在解决该问题上的思路主要可以归纳为以下两点。

2.1 判断暴发式任务请求始末

模型的首要任务为判断暴发式任务请求的到来，即通过引入负载监听指数 I 对每单位时长的任务请求量及变化率进行定量分析。

其表达式为：

$$I=SCV(1+2\sum_{k=1}^{\infty}\rho_k) \tag{1}$$

式(1)中变量的平方系数(Squared Coefficient of Variation，SCV)为一个固定长度任务请求量的平方系数，ρ_k 代表自相关系数，是一种用来寻找随机变量与系统自身关系的统计学方法。假设一个时序的系统随机变量值为$\{X_n\}$，其中 $n=(0,1\cdots\infty)$，则：

$$\rho_k=\frac{E[(X_t-\mu^{-1})(X_{t+k}-\mu^{-1})]}{\sigma^2} \tag{2}$$

式中，μ^{-1}代表均值，σ^2 为变量$\{X_n\}$的方差。

现有算法均是从任务请求量的绝对值量与变化率两个角度来度量每一单位时间长度到达的任务请求量是否满足其所定义的暴发式任务请求，具体内容参见文献[6]。此类算法可以很好地判断暴发式任务请求的始末，与此同时也可有效避免当任务请求量在适当范围内波动时，将其误判为暴发式任务请求状态的发生。

2.2 资源部署策略设计

当前云计算系统中主流的资源部署策略有 HP Laboratory 提出的一种改进的模拟退火算法、OpenNebula 等开源软件使用的 First-Come-First-Server 机制[7]、刘进军设计的基于组播技术实现的部署模型[8]等，这些部署策略(模型)的共同特征是均以单方面性能、目的为目标，追求其最优值。此类部署策略在暴发式任务请求环境下会面临云计算系统在通常情况下没遇到的问题，即在某一瞬时时刻系统的性能会有较大幅度的下降。

为应对暴发式任务请求对系统性能带来的冲击，目前可参考的解决方案主要是针对主流的策略基于某一(些)特定的需求所提出的改进，其可以归纳为表 1 所列。

表 1 The high level of the Bursty Workloads Allocation

```
input
  N, the number of available site.
  K, the candidate sites (1<K<N).
  I, the state of system (burst or non burst).
  The algorithm of Bursty Workloads Allocation
1. if (detect the start of burst)
2. {set K to Nub;}// Nub is close to N; such as Nub=1/2N.
3. set k to Nus;// Nus to be a small value; such as 1.
4. end if
5. analysis all sites Si;//1<i<N
6. select S={S1,S2,…,Sk}//select out the best K sites.
7. select S'=uniform(1,K)//under the random measure.
8. submit the job to S'.
```

从表 1 可以看出，算法是通过调整部署节点的数量来完成系统优化的，即算法避免了暴发式任务请求的到来使云计算系统短时产生局部热点而造成系统性能下降。但此类算法也明显存在一个不足点：对于 Nu_b、Nu_s 值的设定模式过于单一，无法与云计算系统底层资源实时高度变化的情况相结合，其性能有待进一步提高。

3 基于 BP 神经网络改进的部署模型

针对以上问题，通过引入误差反向传播神经网络使系统具有自适应、自学习能力，实现对部署参数的动态调整，提高云计算系统应对暴发式任务请求的响应性能表现。本文提出的自适应模型的主要工作流程如图 1 所示。

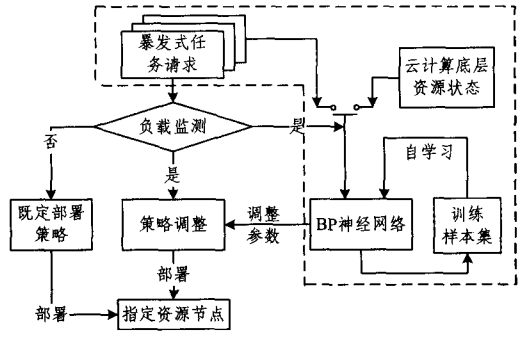

图 1 基于 BP 网络改进的云环境应对 BW 资源部署模型

Step 0 确定神经元。首先分析云计算系统的输入与期望的输出项的值、类型，按照 BP 神经网络的要求，明确与数据相对应的神经元的类型、输入输出神经元的种类。

Step 0′ 训练样本集。通过对样本集(样本集的选取规则由下节说明)的学习，根据差错的大小和方向调整连接权值，使输出模式和网络预期的输出模式相同，这样才能使该模型拥有预测系统参数的能力。

Step 1 触发开关。由负载监测模块实时监测任务请求的变化情况，当模块监测到暴发式任务请求到来的时候，触发开关，使 BP 神经网络调取任务请求与云计算底层资源的相关信息。

Step 2 量化 BW 值。通过量化暴发式任务请求值，可使之更适应 BP 神经网络输入方式，加快网络处理数据的速度。

Step 3 量化资源池指标。对收集的云计算底层资源池信息进行标准化处理，可使神经网络定量地分析底层资源的变化，为接下来的参数预测提供保障。

Step 4 参数预测。参数预测为本模型的核心任务，通过前期的模型训练与对输入信息的收集、处理，使部署模型具有实时预测资源部署参数的能力，优化云计算系统应对暴发式任务请求的能力，切合云计算系统动态可伸缩的特征。

通过前两步的预测模型学习过程，使模型拥有预测能力。随后对采集的输入值进行量化后输入到 BP 神经网络模型中处理，产生的输出值即为云计算系统应对暴发式任务请求的参数调整量。

4 分析与实验

本文以 1998 年世界杯官方网站任务请求量的数据为模拟环境[9]，对提出的基于 BP 神经网络的云环境下暴发式任务请求资源部署模型进行应用实验，先确定神经网络的结构与初始参数；通过对样本的提取与学习，对本模型完成训练后，在仿真环境下进行模拟测试，分析、总结其性能表现[10]。

4.1 确定网络结构

构建一个满足云中暴发式任务请求环境的 BP 神经网络

模型时，对输入神经元的选取十分关键，因为其能够直接影响数据预测的结果。通过综合分析暴发式任务请求资源部署模型的特征，选取 3 个影响因子作为本神经网络的输入，它们分别为任务请求量的负载监听指数 I、云计算资源池拥有的资源节点数 N_t、资源池某时刻可用的节点数 N。其中 I 可以定量地分析单位时间段内任务请求的强度，N_t 可判断该云计算系统应对暴发式任务请求的能力，N 可以作为系统当前性能的度量指标。

本模型的主要目的为通过引入 BP 神经网络模块，对云计算系统在应对暴发式任务请求时的资源部署策略做出优化，使其为用户提供更好的体验效果。BP 神经网络模块输出的调整参数为 Nub' 值与 Nus' 值，利用这两个输出值对云计算系统部署方案进行动态调整，以满足实时跟踪底层资源的变化情况。其结构如图 2 所示。

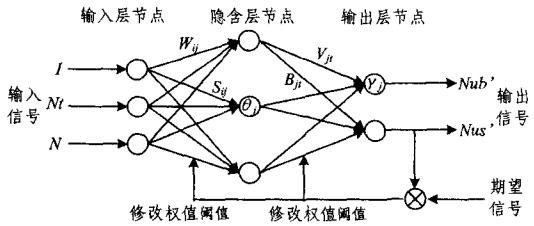

图 2　误差反向传播神经网络模块结构图

4.2　学习算法

设 BP 神经网络输入模式向量为 $X^k=(I^k,N_t^k,N^k)^T(k=1,2,\cdots,m)$，其中 m 为学习模式对个数；对应的输入模式期望输出为 $Y^k=(N_{us}'^k,N_{us}^k)^T$；中间隐含层的净输入向量为 $S^k=(s_1^k,s_2^k,s_3^k)^T$，输出向量为 $B^k=(b_1^k,b_2^k,b_3^k)^T$；输出层的净输入向量为 $L^k=(l_1^k,l_2^k)^T$，实际输出向量为 $C^k=(c_1^k,c_2^k)^T$；输入层至隐含层的连接权值为 $W=\{w_{ij}\}(i=j=1,2,3)$，隐含层至输出层的连接权值为 $V=\{v_{jt}\}(j=1,2,3,t=1,2)$；隐含层各单元的阈值为 $\theta=\{\theta_j\}(j=1,2,3)$，输出层各单元的阈值为 $\gamma=\{\gamma_t\}(t=1,2)$。本文提出的 BP 神经网络模型采用 Sigmoid 函数作为转移函数，因为本模型需要网络的输出值为一个实际的实数，而不是类似单层感知机模型中的二值离散值。Sigmoid 函数的数学表达式为：$f(x)=\dfrac{1}{1+e^{-x}}$。

BP 算法描述如下：

1）初始化。为连接权值 W、V 及阈值 θ、γ 赋予[−1，+1]区间的随机值。

2）随机选取一个学习模式对 (X^k,Y^k) 提供给网络。

3）计算隐含层的输入输出值。

$$S^k=\sum_{i=1}^{3}w_{ij}X^k-\theta_j,j=1,2,3;B^k=f(S^k),j=1,2,3$$

4）计算输出层各个神经元的净输入和实际输出。

$$L^k=\sum_{j=1}^{3}v_{jt}B^k-\gamma_t,t=1,2;C^k=f(L^k),t=1,2$$

5）根据给定的期望输出，计算输出层各神经元的矫正误差。

$$d_t^k=-\frac{\partial E^k}{\partial l_t^k}=\frac{\partial E^k\partial c_t^k}{\partial c_t^k\partial l_t^k}=(y_t^k-c_t^k)f'(l_t^k),t=1,2$$

6）计算隐含层各个神经元的校正误差。

$$e_j^k=[\sum_{t=1}^{2}v_{jt}d_t^k]f'(s_j^k),j=1,2,3$$

7）修正隐含层至输出层的连接权值和输出层神经元的阈值，其中 α 为学习速率，在这里取 0.5。

$$\Delta v_{jt}=-\alpha\frac{\partial E^k}{\partial v_{jt}}=-\alpha\frac{\partial E^k\partial c_t^k}{\partial c_t^k\partial v_{jt}}$$

$$=\alpha\partial_t^k f'(l_t^k)b_j^k=\alpha d_t^k b_j^k,j=1,2,3,t=1,2$$

$$\Delta\gamma_t=\alpha d_t^k,t=1,2$$

8）修正输入层至隐含层的连接权值和隐含层神经元的阈值，其中 β 为学习速率，在这里取 0.5。

$$\Delta w_{ij}=-\beta\frac{\partial E^k}{\partial w_{ij}}=-\beta\frac{\partial E^k\partial b_j^k\partial s_j^k}{\partial b_j^k\partial s_j^k\partial w_{ij}}$$

$$=-\beta[\sum_{t=1}^{2}(-d_t^k)v_{jt}]f'(s_j^k)x_i^k$$

$$=\beta e_j^k x_i^k(i=j=1,2,3)$$

$$\Delta\theta_j=\beta e_j^k,j=1,2,3$$

9）随机选取下一个学习模式对提供给网络，重复 3）直至全部学完。

10）判断网络全局误差 E 是否满足精度要求，即 $E<\varepsilon$。若满足，则结束，否则更新网络学习次数。

11）网络输出预测修改的参数值，部署策略随即做出相应调整。

4.3　网络训练

完成了网络的结构与算法设计，接下来就是对 BP 神经网络进行样本集的训练，明确网络走向，达到预测参数的目的。以 1998 年世界杯官方网站某一时间段的任务请求量数据为基础，在选取时间段样本时，因网站的任务请求量是一个动态变化的过程，故选取的样本要照顾到不同强度的暴发式任务请求量，这样选取样本可使其不具备“特殊性”，避免了网络只能学习到某种特定的规律，从而降低网络的推广能力。首先计算每分钟任务请求量的 I 值，云计算资源池规模限定在 1000～10000 个计算节点里，某时刻可用节点数限定在 0.5～0.9 倍资源池节点数的范围内。用于反馈的网络期望输出信号由人工实现实验完成，部分样本如表 2 所列，根据 min-max 规范化的方法对原始数据进行归一化处理。转换公式为：

$$A=\frac{X-V_{\min}}{V_{\max}-V_{\min}}$$

表 2　网络调整参数预测归一化处理前的样本数据

序	系统输入(PerMin)			系统输出(PerMin)	
(单位长)	I	N_t	N	Nub′	Nus′
1	3650	9000	7000	131	0
2	3650	9000	4800	362	0
3	3650	5000	4000	247	0
4	3650	5000	3200	413	0
5	2780	8000	6000	118	0
…	…	…	…	…	…

按照上一小节学习算法流程迭代样本数据，反复训练 BP 神经网络，直到网络满足收敛条件，停止训练。

4.4　性能测试与分析

本文利用云计算仿真软件 CloudSim 进行实验仿真[11]，通过对 CloudSim 扩展编译，添加判断暴发式任务请求的方法，增加并行的 BP 神经网络学习模块以实现本文所提出的方法。对训练完毕的网络输入剩余样本进行性能测试，其性能表现如图 3 所示。实验首先对不同时刻的任务请求量进行相应的负载监听指数变换，随后判断暴发式任务请求的到达与结束时刻(图中的 A、B 区间)，通过观察响应时间的对比图，发现增加了 BP 神经网络模块后的系统响应时间明显缩短，但在[A-c]、[d-B]段，增加了网络模块的系统性能略逊于现有系统，造成这种现象的主要原因是由于引入 BP 神经网络模块后，其占用一定的系统资源，造成系统性能的小幅度下降，在暴发式任务请求强度并不是十分剧烈的时候，引入网络

模块给系统性能带来提升的效能要小于模块占用系统资源的性能，所以造成了小幅度的系统响应时间增加，但随着暴发式任务请求强度的增加，其所占系统资源的比重不断减小，其给系统性能提升带来的表现逐渐显现。因此得出结论，通过引入 BP 神经网络模块可使云计算系统在应对暴发式任务请求时的性能得到有效提高，本文提出的模型是一种可行的方法。

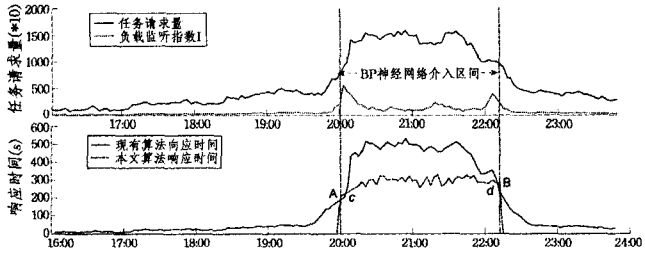

图 3　添加 BP 神经网络模块前后性能对比

结束语　本文针对暴发式任务请求对云计算系统性能带来的冲击，设计了基于 BP 神经网络算法改进的资源部署模型，它能够在判断暴发式任务到来的同时，改变资源部署策略，并通过 BP 神经网络模块，实时优化调整参数，达到优化系统全局的目的。本文首先分析了现有的应对暴发式任务请求资源部署模型，并归纳总结了现有的优势与不足，随后引入网络模块来解决动态跟随底层资源与请求量变化的问题。从部署模型的应用原理与流程出发，对其网络结构与学习算法进行设计，最后通过仿真实验对模型进行验证，结果表明本文方法可行、有效，对提高云计算系统应对暴发式任务请求能力起到积极作用。

参 考 文 献

[1] 文雨，孟丹，詹剑锋. 面向应用服务级目标的虚拟化资源管理[J]. 软件学报，2013，24(2)：358-377

[2] Caniff A, Lu Lei, Mi Ning-fang, et al. Fastrack for Taming Burstiness and Saving Power in Multi-Tiered Systems[C]//22nd International Teletraffic Congress (ITC 22). Amsterdam, the Netherlands, September 2010

[3] Tai Jiang-zhe, Meleis W, Zhang Jue-min, et al. ARA: Adaptive Resource Allocation for Cloud Computing Environments under Bursty Workloads, 978-1-4673 [R]. Northeastern University, Boston, USA, 2011

[4] 高刃，唐龙，伍爵博. 基于神经网络的无线传感器网络数据预测应用研究[J]. 计算机科学，2012，30(5)：44-47

[5] 马锐. 人工神经网络原理[M]. 北京：机械工业出版社，2010

[6] Tirado J M, Higuero D, Isaila F, et al. Predictive Data Grouping and Placement for Cloud-based Elastic Server Infrastructures [C]//2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing. IEEE DOI/CCGrid, 2011：285-294

[7] C12G Labs S. L. Private cloud computing with OpenNebula 1. 4 [EB/OL]. http://opennebula. org/_ media/software: ecosystem:private_cloud_computing_with_opennebula_1. 4. pdf, 2010

[8] 刘进军，赵生慧. 面向云计算的多虚拟机管理模型的设计[J]. 计算机应用，2011，31(5)：1417-1419

[9] Arlitt M, Jin T. Workload characterization of the 1998 World Cup Web site[R]. HPL-1999- 35R1. HP Laboratories, 1999

[10] 李强，郝沁汾，肖利民，等. 云计算中虚拟机放置的自适应管理与多目标优化[J]. 计算机学报，2011，34(12)：2253-2264

[11] Rodrigo N C, Ranjan R, Beloglazov A, et al. CloudSim: A Toolkit for Modeling and Simulation of Cloud Computing Environments and Evaluation of Resource Provisioning Algorithms[R]. Cloud Computing and Distributed Systems Laboratory, Australia, 2010

(上接第 238 页)

子节点的局部信息传播轨迹获取算法(K-DFS, ELPS)，以从微博社会网络中获取信息传播轨迹；(4)提出了信息传播倡导者发现算法(KAD)；(5)提供了充分的实验在真实的微博社会网络(新浪微博、Twitter、Flickr、Douban)中讨论分析本文所提算法的效果及效率。在实验部分，我们验证了信息传播轨迹是一个可用于研究社会网络中信息传播规律不错的数据结构。除此外，本文所提方法可不局限于使用在微博领域的信息传播规律分析领域中，还可用于分析其他 SNS 应用、在线广告投放或电子商务领域的其他应用中。我们未来的工作方向为：提高所提算法性能，根据观察 SNS 应用中用户对不同新闻事件话题信息的传播方式，尝试建立新闻事件与微博信息传播轨迹相关的联系。

参 考 文 献

[1] Granovetter M. The strength of weak ties[J]. American Journal of Sociology, 1973, 78(6)：1360-1380

[2] Huberman B A, Adamic L A. Information Dynamics in the Networked World [J]. Lect. Notes Phys. , 2004, 650：371-398

[3] Kossinets G, Kleinberg J M, Watts D J. The structure of information pathways in a social communication network[C]// Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2008：435-443

[4] Kossinets G. Effects of missing data in social networks[J]. Social Networks, 2006, 28：247-268

[5] Laumann E, Marsden P, Prensky D. The boundary specification problem in network analysis[J]. Applied Network Analysis, 1983(10)：18-34

[6] Leskovec J, Adamic L A, Huberman B A. The dynamics of viral marketing[J]. ACM Transactions on the Web (TWEB), 2007, 1 (1)

[7] Liben-Nowell D, Kleinberg J. Tracing information flow on a global scale using Internet chain-letter data[J]. Proc. Natl. Acad. Sci. USA, 2008, 105(12)：4633-4638

[8] Bakshy E, Rosenn I, Marlow C, et al. The role of social networks in information diffusion[C]//WWW. 2012：519-528

[9] 樊鹏翼，王晖，姜志宏，等. 微博网络测量研究[J]. 计算机研究与发展，2012(4)：691-699

[10] Qin L, Yu J X, Chang L. Keyword search in databases: the power of RDBMS[C]//SIGMOD Conference. 2009：681-694

[11] Illenberger J, Kowald M, Axhausen K W, et al. Insights into a spatially embedded social network from a large-scale snowball sample[C]//The European Physical Journal B-Condensed Matter and Complex Systems. 2011：1-13

[12] Song Xiao-dan, Chi Yun, Hino K, et al. Identifying opinion leaders in the blogosphere[C]//CIKM 200. 72007：971-974

[13] Zafarani R, Liu H. Social Computing Data Repository at ASU [OL]. http://socialcomputing. asu. edu. Tempe, AZ: Arizona State University, School of Computing, Informatics and Decision Systems Engineering, 2009

[14] http://jung. sourceforge. net/