云计算中的集群资源模糊聚类划分模型

刘伯成 陈庆奎
(上海理工大学光电信息与计算机工程学院 上海 200093)

摘 要 随着云计算应用的开展,计算机集群的作用越来越重要,但集群中计算机的性能良莠不齐,虽然能互联且能共享资源,但是很有可能因为某些集群内的计算机的性能不均衡或集群性能与并行任务资源需求不匹配而造成任务低效执行的后果。如何把物理集群(普通局域网互联的计算机构成)分为若干个性能均衡的逻辑集群是集群调度的关键。通过对计算机资源的模糊聚类来划分集群中的计算机,引进任务资源需求向量和最低误差容忍向量机制,把物理集群划分为若干个性能均衡或与并行任务资源匹配的逻辑计算机集群,使集群更易管理调度。对物联云运用此算法划分了网关集群、数据库集群和服务集群,验证了本算法,这种划分方法适合云计算应用。

关键词 云计算,计算机集群,资源,模糊聚类

Fuzzy Clustering Partition Model for Computer Cluster in Cloud Computing

LIU Bo-cheng CHEN Qing-kui
(Optical Electrical Information and Computer Engineering School, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract With the development of cloud computing applications, the computer cluster becomes more and more important, but the computer performance in cluster is so different. Even though they can connect and can share their resources each other, but when the computer performance is not balance and the cluster performance is not to match with the parallel tasks, it can cause the low efficiency for parallel task. So, it is an important problem how to partition and management computer cluster. The paper gave a way to solve it. Using the fuzzy clustering method according to the computer resources in computer cluster and through the vector of resource demand and the vector of lowest inaccuracy tolerance, we can divide the computer cluster into several classes(logical computer cluster) and make the every computer performance in one class to be more similar. It makes the computer cluster management to become easier. We partition the computers in lab into network gateway cluster, data base cluster, and service cluster for the internet of thing cloud using this model. This model is fit for the cloud computing applications.

Keywords Cloud computing, Computer cluster, Resource, Fuzzy clustering

1 问题的提出

随着互联网应用的飞速发展,信息资源呈爆炸趋势增长,海量信息的存储、查询、分析处理问题对现代计算机系统带来了极大的挑战。面对海量信息互联网应用和商业模式的快速普及,以云计算[1]为代表的新型网络计算模式应运而生,并且已经成为业界研究的热点。学术界对GOOGLE的云计算技术进行了应用探讨,如HADOOP[2]已经成为业界学习和探讨云计算技术的主要工具。云计算的物理基础就是计算机集群,通过成千上万个普通个人计算机系统,构建大规模的信息存储、信息查询和信息处理系统,已经被GOOGLE经典地实现。集群计算技术在过去的20年里得到了长足的发展[3,4],但这些技术大多是基于专有计算机集群平台的,如IBM深蓝计算机、曙光5000A并行计算机、天河1号计算机等。这些集群的特点是硬件资源统一、设备专有,并且软件系统也是专用的。另一方面,在现实生活的环境中,分布着大量各类性能的计算机系统,并且已经形成了成千上万个局域网络(由局域网互联的计算机集合构成物理集群),这些网络中的计算机资源绝大多数时间里是空闲的,造成了计算资源的很大浪费。如果能利用现有的计算资源,把空闲的CPU利用起来,这会是一件很开心的事情。然而,由于每个计算机的资源状况配置不同导致了每个计算机的性能也不一样,这些性能的不同会大大增加集群资源的调度和任务分配的难度,致使无法有效地利用集群资源,从而很难支持大规模信息处理。基于集群资源特征的应用模型研究已有很多[5,6]。文献[5]针对集群动态计算过程中的资源变换情况,运用机器学习机制来改善集群资源的应用。文献[6]运用动态迁移机制在集群资源不断变化的条件下,保证计算任务的执行。然而,随着云计算应用的开始,面对现实中的局域网络(物理集群)中的不同计算资源配置的计算机,如何把它们按能力聚类分组,形成一个个逻辑集群,且这些逻辑集群内的计算资源配置是相近的,可以支持某类并行计算,并且任务在逻辑集群内是均衡的,这些

本文受上海信息技术领域重点科技攻关(09511501000),上海重点科技项目(09220502800)资助。

刘伯成 硕士生,主要研究方向网络计算;陈庆奎 教授,博士生导师,主要研究方向为网络计算、云计算、物联网技术,E-mail:chenqingkui@tom.com(通信作者)。

是很有现实意义的研究。可以根据不同计算能力的逻辑集群来分配其在云计算中的角色，如CPU能力强的逻辑集群可疑被分配以计算密集型任务，内存容量大、I/O能力强的逻辑集群可以用于海量信息并行处理。本文对物理集群内计算节点的计算资源，如CPU、内存、磁盘、网络等资源进行定量度量，然后运用模糊聚类[7,8]技术，按照任务对资源需求的情况，把普通局域网络中的计算机划分成若干个能力均衡的逻辑子集(逻辑集群)。分析与实践表明，这种划分是有效的。可以在配置Hadoop云模型时合理利用计算资源。该模型适合云计算应用。

2 划分模型

2.1 计算机参数

在讨论计算机参数时，本文只讨论集群内计算节点的最普遍的5个参数，它们是：①CPU：计算机运算速度，这个衡量标准很多，有的用MIPS来衡量，有的用MHZ，只要量纲统一即可；②MEMORY：内存，可以通过容量来度量；③Net Adapter：网卡，通过网卡的处理器、通信机制来度量；④I/O：硬盘读写速度，通过磁盘的I/O带宽或磁盘转速来度量；⑤NET：网络带宽。当然，还有其他很重要的参数，本模型对参数的个数不敏感，均可以同样处理。同时，本模型对参数的量纲也不敏感，只要同一参数的量纲一致即可，在聚类分析时，会通过数据的标准化消除量纲的影响。

2.2 参数权值

对于集群中的计算机的参数，在聚类时可以分派不同的权值，以便分类能根据计算任务特点划分集群，上述5类资源的权限分别表示如下：①CPU：K_1；②MEMORY：K_2；③Net Adapter：K_3；④I/O：K_4；⑤NET：K_5。其中，$K_1+K_2+K_3+K_4+K_5=1$。权值与属性参数个数相对应。计算任务的并行性和串行性本质上决定着集群的应用效率。而每个节点的运行效率也就是这个节点计算的本质，所以在模糊聚类时，对于对应于某个任务很重要的一个或几个至关重要的属性参数给于比较大的权值，这样，权值比较大的属性参数对于相似度的影像就比较大，集群的划分就能根据任务的特点聚类，而且适应任务。

2.3 模糊聚类分析

对集群计算机聚类一般分为3个步骤：(1)数据标准化；(2)建立模糊相似矩阵；(3)聚类。下面分别详细论述。

2.3.1 数据标准化

建立数据矩阵并去除数据矩阵中的量纲，而得到标准化的数据。

(1)建立数据矩阵

设被分类的物理集群为$C=\{C_1,C_2,\cdots,C_n\}$，n为集群中计算机的数目。每个计算机由m个属性参数表示其性能，例如本文实验中用5个属性来描述一台计算机，即x_1为CPU特性、x_2为内存MEMORY特性、x_3为Net Adapter网络适配器特性、x_4为I/O特性、x_5为网络NET特性。为了描述物理集群中的n台计算机，我们用向量$C_i=(x_{i1},x_{i2},\cdots,x_{im})$ $(i=1,2,\cdots n)$，表示第i个计算机的资源性能描述。于是，得到了集群的$n\times m$原始数据矩阵：

$C=[\ C_i\mid C_i=(x_{i1},x_{i2},\cdots x_{im}),1\leqslant i\leqslant n]$

(2)标准化数据

在物理集群中的计算机的m个属性参数有不同的量纲，为了使不同的量纲也能进行比较，就要做适当的变换，把不同纲量的数据根据模糊矩阵的要求规范到区间[0,1]上。本文给出平移极差变换，更多变换请参见文献[7,8]。平移差极变换方法如下：

$$x''_{ik}=\frac{x'_{ik}-\min\limits_{1\leqslant i\leqslant n}\{x'_{ik}\}}{\max\limits_{1\leqslant i\leqslant n}\{x'_{ik}\}-\min\limits_{1\leqslant i\leqslant n}\{x'_{ik}\}},(k=1,2,\cdots,m)$$

显然，有$0\leqslant x''_{ik}\leqslant 1$，每个属性参数的均值为0，标准差为1，且消除了量纲的影像。

2.3.2 建立模糊相似矩阵

要被分类的物理集群为$C=\{C_1,C_2,\cdots,C_n\}$，每个计算机描述为$C_i=(x_{i1},x_{i2},\cdots,x_{im})(i=1,2,\cdots,n)$，$K=(K_{i1},K_{i2},\cdots,K_{im})$为资源权值向量。把性质相近的归为一类，要根据某种方法求出任意两个计算机C_i和C_j之间的相似度$r(C_i,C_j)$，记为r_{ij}，显然有$r_{ii}=1$和$0\leqslant r_{ij}=r_{ji}\leqslant 1$。由节点相似度$r_{ij}$可以构建物理集群$C$的计算节点间的一个模糊相似矩阵$R$。

$$R=\begin{pmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nn} \end{pmatrix}$$

$r_{ii}=1,0\leqslant r_{ij}=r_{ji}\leqslant 1,1\leqslant i,j\leqslant n$

求模糊相似度r_{ij}的方法非常多，本文给出绝对值指数法，关于更多方法请参见文献[7,8]。绝对值指数法如下：

$$r_{ij}=e^{-\sum_{k=1}^{m}K_k|x_{ik}-x_{jk}|}$$

可用r_{ij}表示X_i与X_j的相似系数，这种方法适用于$x_{ij}>0$的情况，这里$1\leqslant i\leqslant n,1\leqslant k\leqslant m$。

2.3.3 聚类

有了模糊相似矩阵R还是不够的，因为R还没有传递性，比如把计算机彼此之间的相似度大于或等于阈值α的计算机算作是一类，那么当$r_{ij}\geqslant\alpha$且$r_{jk}\geqslant\alpha$时未必有$r_{ik}\geqslant\alpha$，即当X_i与X_j同类且X_j与X_k同类时，X_i与X_k未必同类，所以还要在R的基础上求出R的等价闭包(记为R^*)。求传递闭包只是根据模糊相似矩阵来聚类的一种方法，其他聚类方法也很多，更多的方法可以参见文献[7,8]。

$$R^*=\begin{pmatrix} r^*_{11} & \cdots & r^*_{1n} \\ \vdots & \ddots & \vdots \\ r^*_{n1} & \cdots & r^*_{nn} \end{pmatrix}$$

$r^*_{ii}=1,0\leqslant r^*_{ij}=r^*_{ji}\leqslant 1,1\leqslant i,j\leqslant n$

(1)闭包聚类法

由于根据计算机间m个相关性属性参数构造出的模糊关系矩阵R往往是模糊相似矩阵，但不一定是模糊等价矩阵，因此必须计算等价关系阵R^*。计算过程如下：

循环对R做如下操作：

$$r'_{ij}=t_{ij}=\bigvee_{k=1}^{n}(r_{ik}\wedge r_{jk})=\max_{1\leqslant k\leqslant n}\{r_{ik}\wedge r_{jk}\}$$

直到r_{ij}保持不变，此时$r_{ij}=r^*_{ij}$，即$R=R^*$。

(2)划分阈值α

在模糊聚类分析中，对于各个不同的阈值$\alpha(0\leqslant\alpha\leqslant 1)$，可以得到不同的分类，从而形成了一种动态的聚类图，这对于全面了解物理集群中的计算机是比较直观形象的。但是当要根据任务的特性聚类时，要根据具体任务特点来确定一个阈值α。相似度在这个阈值内的计算机被认为是性能相同的节

点。用统计方法确定 α 值能适应一般性的计算任务，也是一种很好地确定 α 的方法。α 的计算方法参见 2.3.4 节。

(3)聚类划分集群

对第(1)步所得的闭包矩阵 R^* 和根据用户需求而定义的阈值 α 进行逻辑集群划分，划分过程可以如下计算：

$M_L=R^*-M_\alpha$

式中，$M_\alpha=[t_{ij}\mid t_{ij}=\alpha,1\leqslant i,j\leqslant n]$，$M_L$、$M_\alpha$ 均为 $n\times m$ 矩阵。

从矩阵 M_L 中清除值不大于 0 的元素，剩下的元素的下标所对应的计算节点，既为同一个逻辑集群。因为 M_L 具有对称性，故要去掉重复的元素。

2.3.4 阈值的确定

设一个并行计算任务为 T，其需要一组性能相近的计算机协同来执行。任务 T 需要对计算机各类资源有个需求意向，所以用 $D(d_1,d_2,d_3,\cdots,d_m)$来表示 T 的资源需求向量，d_i 表示 T 对第 i 类资源的需求度，m 是所需的资源数目。

由前面叙述所知，D 的分量需要规约为[0,1]之间的数值，可以用如下计算公式规约：

$$\overline{d_i}=d_i/\sum_{j=1}^{m}d_j$$

式中，$1\leqslant i\leqslant m$。这样，$\overline{d_i}$即位任务 T 对第 i 类资源需求的权值。需求向量 D 变换为需求权值向量$\overline{D}=(\overline{d_1},\overline{d_2},\overline{d_3},\cdots,\overline{d_m})$。由于计算机之间的能力是绝对存在差异的，故需要用户或系统给出一个最低误差容忍度向量 $E=(e_1,e_2,e_3,\cdots,e_m)$，$e_i$ 为第 i 类资源的最低误差容忍度，且 $0\leqslant e_i<1$，e_i 值越大说明该因素越重要，要求其误差就越小。故可以计算任务 T 的集群划分阈值 α，计算公式如下：

$$\alpha=\sum_{i=1}^{m}d_i\times e_i$$

由于$\sum_{i=1}^{m}d_i=1$，且 $0\leqslant e_i<1$，因此 $\alpha\leqslant 1$。

故可以利用 α 来构造 2.3.3 节所描述的 M_c 矩阵。

3 模型有效性讨论

有关 R 的闭包 R^* 是一个模糊等价关系的问题已经在文献[7,8]中详细描述，这里不再赘述，仅讨论阈值、需求权向量、最低误差容忍度之间关系和实际有效性问题。

由 2.3.4 节可知，$\sum_{i=1}^{m}d_i=1$，故 $\alpha=\sum_{i=1}^{m}d_i\times e_i$ 的大小主要取决于最低误差容忍度向量 $E=(e_1,e_2,e_3,\cdots,e_m)$。

如果所要划分的物理计算机集群是能力同构的，且如果当 E 的每个分量都趋于 1 时，说明任务对每类资源的最低误差容忍都是很小的，故 α 趋于 1，那么由此 α 构成的 M_c 矩阵的元素值都趋于 1，故 $M_L=R^*-M_\alpha$ 的正数元素就很少，故构建的逻辑集群规模很小。

如果所要划分的物理计算机集群是能力同构的，即均是同样配置的计算机，那么 R^* 元素的值均为 1。这时不管 E 的每个分量取值如何，M_L 中元素值均大于 0，故能力同构的物理计算机集群被划为一个逻辑集群，这说明该模型在边界值条件下也成立。

物理计算机集群的计算节点间计算性能差异越大，R^* 元素的值就相对越小，这时如果阈值 α 取值过大，那么 M_L 的正值元素就很少，能够划分成一个逻辑集群的几率就小。故在性能差异大的原始集群中划分时，必须降低 α 的值。根据 α 的计算公式可知，只有降低误差容忍向量 E 的值就可以降低 α 的值，这也非常符合现实情况的。E 的分量值的调整，可以根据计算任务对各类资源的需求的刚性程度来确定。如任务 T 对第 i 类资源需求非常多，而其他类资源要求一般，当$\sum_{j=1}^{m}e_j$的值被约束为恒定值时，可以优先保证 e_i 的值，相应减少其它资源的误差容忍度值。

综上所述，该模型中 R^* 的构建反映了 n 个计算机之间在每个资源能力约束下的相似性。阈值 α 既反映了任务对 m 类资源的需求程度，又通过误差容忍度向量 E 反映了用户对资源差异的一个适应度。该模型可以有效地表示资源差异和任务需求之间的矛盾。所以该模型是有效的。

4 模型算法测试

构建一个 30 个计算机组成的物理集群，表 1 是这 30 台计算机的 5 类资源属性描述。这里 CPU 度量为频率系数；内存度量为容量系数；磁盘为 I/O 存取系数；网络适配器为通信速度系数；网络为带宽系数，网络带宽一致。利用表 1 根据第 2 节的描述方法来构建 R^*，进而对其实施划分。

表 1 30 个计算机的 5 类资源特征描述

计算机号	CPU(G)	MEM(G)	NA(M/S)	I/O(M/S)	NET(M/S)
1	2.4	1	100	300	100
2	1.6	1	100	300	100
3	1.6	1	100	300	100
4	2.2	1	100	300	100
5	4.2	1	100	300	100
6	4.2	1	100	300	100
7	1.6	0.5	100	300	100
8	1.8	0.5	100	300	100
9	1.8	0.5	100	300	100
10	4.2	1	100	300	100
11	7.2	3.6	100	300	100
12	7.2	3.6	100	300	100
13	4.4	1	100	300	100
14	4.4	1	100	300	100
15	4.4	0.5	1000	300	100
16	4.6	1	1000	150	100
17	4.6	1	1000	150	100
18	4.6	1	1000	150	100
19	1.6	0.5	100	150	100
20	4.2	1	1000	150	100
21	4	0.25	100	150	100
22	3	0.25	100	150	100
23	3	0.25	100	150	100
24	2.8	0.5	100	150	100
25	4.6	1	100	150	100
26	3	0.5	100	150	100
27	3	0.5	100	150	100
28	2.8	1	100	150	100
29	3	0.5	100	150	100
30	3	0.5	100	150	100

试验 1 不同资源需求权值下聚类分析检验

本试验对 5 种不同任务的资源权值分布特征进行聚类测试，这 5 个资源权值分布是：

(1)$d_1=0.8,d_2=d_3=d_4=d_5=0.05$

(2)$d_1=0.6,d_2=d_3=d_4=d_5=0.1$

(3)$d_1=0.4,d_2=d_3=d_4=d_5=0.15$

(4)$d_1=0.2,d_2=d_3=d_4=d_5=0.2$

(5)$d_1=0.1,d_2=0.5,d_3=0.1,d_4=0.2,d_5=0.1$

对于第(1)类表示计算密集型任务需求，如数值计算、数据挖掘分析等；第(2)类是对 CPU 需求强而对其它 4 类资源

需求均衡的计算任务；第(3)、(4)类为均衡需求计算任务的资源需求；第(5)类为对内存要求较高的任务。这里设定最低的误差容忍度向量为(0.8,0.8,0.8,0.8,0.8)，对5类任务类型分别计算 α，然后进行聚类集群划分。聚类划分的结果如表2所列。横向表示5种权值分布下的聚类成果，纵向表示分类情况。表中 self-group 为无聚集的计算机。

表2　5种不同权值分布下的聚类结果

权值分布类别	(1)	(2)	(3)	(4)	(5)
1	1,4	1,4	1—4 7—9	1—6 10,13,14	1—6 10,13,14
2	2,3 7—9 19	2,3 7—9,19	5,6 10,13 14,25	7—9	7—9
3	5,6,10,13 14,21,25	5,6,10, 13,14,25	11,12	11,12	11,12
4	11,12	11,12	16—18 20	16—18 20	16—18 20
5	15—18 20	15—18 20	22—24 26—30	19, 21—24 26,27,29,30	19 21—24 26,27 29,30
6	22—24 26—30	22—24 26—30			25,28
self-group		21	15,19 21	15,25 28	15

根据分类结果可知，随着对 cpu 需求的权重逐渐减小，对于 cpu 的关注程度也越来越小，例如在第一类中，当 cpu 的权重为 0.8 时，只有 1,4；当权重减小到 0.4 时，2,3,7,8,9 又加了进来，可以看出它们即使在 cpu 上存在一定差异，但是综合考虑其他因素，还是比较相似的。可见该模型的划分与实际是相符的。

试验2　容忍度不同的划分测试

仅对第(4)类任务进行测试，即在资源需求权值均衡的情况下($d_1=d_2=d_3=d_4=d_5=0.2$)，根据不同容忍度值(这里也假设对每类资源的容忍度是一样的)，来观察本模型划分的效果，测试结果如表3所列。

表3　不同容忍度值的划分结果

容忍度分布 / 类别	0.9—0.93	0.94	0.95	0.96	0.97—0.99
1	1—10, 13,14,19, 21—30	1—10 13,14	1—6 10,13,14	1—4	1—4
2	11,12	11,12	7—9	5,6, 10,13,14	5,6,10 13,14
3	15—18 20	16—18 20	11,12	7—9	7—9
4		19, 21—24 26,27,29,30	16—18 20	11,12	11,12
5		25,28	19, 21—24 26,27,29,30	16—18 20	16—18 20
6				21—24 26,27,29,30	22,23 26
7					24,27 29,30
self-group		15	15,25,28	15,19 25,28	15,19,21 25,28

可以看出，当聚类的最低容忍度值增大时，聚类(逻辑集群)就会增多，而且自成一类的计算节点也在增多，这样就使每台计算机在聚类时的特点更加鲜明，每个聚类(逻辑集群)的内部相似性更加接近，但聚类的粒度(逻辑集群规模)更小。相反，逻辑集群的规模就会扩大，集群内的计算节点差异也会增大。实验证明，资源需求权值能很好地控制聚类所关心的属性，最低误差容忍度的变换可以有效调整聚类划分的阈值。

5　云计算应用

云计算过程中需要大规模集群作为物理支撑环境，面对物理集群性能迥异的计算机，通过性能聚类，形成若干个逻辑集群，而每个逻辑集群具有相近的计算资源特性，适合做执行同一类并行计算任务。这种方法可以合理有效地利用计算资源，进一步达到均衡负载。我们在构建物联网存储云和通信云过程中运用本方法划分了集群计算资源，物联存储云如图1所示。图中 SG 为传感网的网管，其负责管理一个或几个物理传感网，同时负责来接受自所辖传感网络的传感器信息，并把其传给相应的 DBS，所有 SG 构成云的网关集群。DBS 为基于 Hadoop 的 HBASE 服务器，其上配置多个 HTABLE，所有 DBS 构成基于 Hadoop 的集群数据库服务系统，来自 SG 的数据被均衡地配置在 DBS 集群上，所有 DBS 构成云的数据库集群。每个 SG 中的传感设备的传感信息存有一个独享的 HTABLE 存储，这里平衡配置是基于 HTABLE 的，即 HTABLE 的创建是 HBASE 在 DBS 集群中自动创建，所有负载均衡由 Hadoop 和 HBASE 来执行。SS 为基于 Web 的物联服务集群，其用来平衡来自用户的服务请求，并且通过访问集群数据库系统来支持用户的传感数据访问，所有 SS 构成云的服务集群。

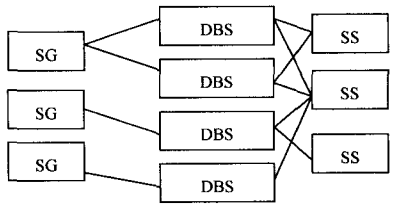

图1　一个物联网存储云模型

网关集群、数据库集群和数据服务集群都是基于 Hadoop 的。只是它们提供的服务不同，不同的服务对计算资源的各类参数的需求也不同。如果同类服务集群的计算节点的配置相近，将对集群的整体性能改进和降低负载均衡的难度有所帮助。例如可以把内存大的逻辑集群作为数据服务集群，把 CPU 强大的逻辑集群作为计算密集任务执行集群(网关集群)，把 I/O 性能好的逻辑集群用于 I/O 密集访问任务的数据库集群。

结束语　本文给出的模糊聚类划分方法能够有效地划分集群中的计算机。目前根据该划分方法，配置多个 Hadoop 集群，进行云计算技术研究，相关研究结果将陆续发表。

参 考 文 献

[1] What is the Cloud Computing [OL]. http://searchcloudcomputing. techtarget. com/definition/cloud-computing

[2] White T. Haddop: The Definitive Guide. O'Reilly Media, 2009

(下转第168页)

我们与如下 3 种方法在上述两种标准下进行比较。

1. Pagerank 方法：该方法纯粹以链接结构为基础计算博客的重要性。选取前 K 个重要度值最大的博客作为最具影响力的博客。

2. 随机取样(RS)的方法：随机选取 K 个博客作为最具影响力的博客。

3. 基于发文数排序方法：选取发文数最多的前 K 个博客作为最具影响力的博客。这是大多数网站常用的博客推荐方法。

如图 1、图 2 所示，我们分别比较了各种算法的博客直接覆盖数和间接覆盖数，我们的算法和 PageRank 算法在不同的 N 值上，覆盖面较高，且曲线越来越平滑。而随机取样算法和基于博文数统计的方法覆盖面较低，且直线陡峭，说明它们选取的点在影响力上是无序的排列，也能进一步说明发表博文数越多的博客，其影响力不一定越大，从而在大多数网站上所采用的基于博文统计数的博客推荐是不合理的。

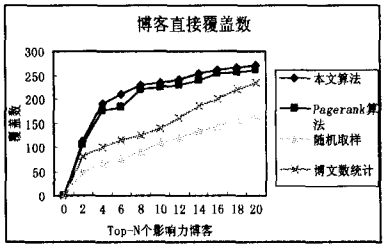

图 1　各种方法的博客直接覆盖数比较

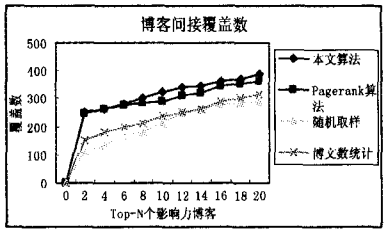

图 2　各种方法的间接覆盖数比较

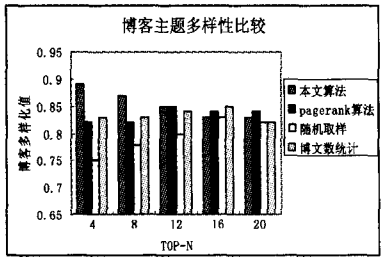

图 3　探测到影响力博客的主题多样性比较

如图 3 所示，我们选取的博客中能覆盖的主题相对较多，特别是在 N 值较小时能覆盖到更多样化的主题。当 N 值逐渐变大时，各种方法的主题覆盖率相对较为接近。

结束语　本文提出了一种探测影响力博客的模型，该模型根据博客特有的格式，不仅考虑了博客在热点主题中的博文数，还分析了引用链接、评论链接这些和影响力密切相关的博客信息，使其结果相比于通过简单统计特征识别重要博客更加合理。在以后的工作中，我们还应充分考虑链接下的语义信息，对各种评论链接、引用链接下的观点进行分析，即是赞同，是反对还是无倾向性的一般陈述，并由此对链接设定权值，计算博客的声誉度，使所提出的模型更加完善。

参 考 文 献

[1] Lin Y R, Sundaram H, Chi Y, et al. Discovery of blog communities based on mutual awareness[C]// Proc. of the World Wide Web 2006 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. Edinburgh, 2006

[2] Lin Y R, Sundaram H, Chi Y, et al. Blog community discovery and evolution based on mutual awareness expansion[C]// Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence. 2007: 48-56

[3] Kritikopoulos A, Sideri M, Varlamis I. Blogrank: ranking Weblogs based on connectivity and similarity features[C]// AAA-IDEA'06: Proceedings of the 2nd international workshop on Advanced architectures and algorithms for internet delivery and applications. 2006: 8

[4] Kleinberg J M. Authoritative sources in a hyper-linked environment[C]// SODA '98: Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms. 1998: 668-677

[5] Page L, Brin S, Motwani R, et al. The pagerank citation ranking: Bringing order to the Web[R]. Stanford Digital Library Technologies Project, 1998

[6] Durant K T, Smith M. Mining sentiment classification from political Weblogs[C]// Proc. of WebKDD workshop inconj with ACM SIGKDD. Philadelphia, PA, August 2006

[7] Gruhl D, Guha R, Liben-Nowell D, et al. Information Diffusion through Blogspace[C]// Proceedings of the 13th International Conference on World Wide Web. 2004: 491-501

[8] Mei Q, Liu C, Su H, et al. A Probabilistic Approach to Spatio-temporal Theme Pattern Mining on Weblogs[C]// Proceedings of the 15th International Conference on World Wide Web. 2006: 533-542

[9] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022

[10] Lu Lu, Zhu Fu-xi, Hu Bin. A novel method to detect latent community in blogspace[J]. Journal of Computational Information Systems, 2010, 7: 2151-2157

(上接第 160 页)

[3] Buyya R. High Performance Cluster Computing: Architectures and Systems, Volume1[M]. 郑纬民，石威，汪东升，等译. 北京：电子工业出版社，2002

[4] Buyya R. High Performance Cluster Computing: Programming and Applications, Volume2[M]. 郑纬民，石威，汪东升，等译. 北京：电子工业出版社，2002

[5] 陈庆奎，那丽春. 基于强化学习的多机群网格资源调度模型[J]. 计算机科学，2007(11)

[6] Chen Qing-kui, Wang Hai-feng, Wang Wei. Continuance parallel computation grid composed of multi-clusters[J]. Journal of Networks, 2010, 5(1): 3-10

[7] 王国俊. 计算智能-词语计算与 Fuzzy 集[M]. 北京：高等教育出版社，2006

[8] 陈水利，李敬功，王向松. 模糊集理论及其应用[M]. 北京：科学出版社，2005