计算机研究与发展 ISSN 1000-1239/CN 11-1777/TP
Journal of Computer Research and Development 48(1): 125-132, 2011

一种基于被动路标的网络距离预测方法

吴国福 窦 强 班冬松 窦文华 宋 磊
(国防科学技术大学计算机学院 长沙 410073)
(gfwu@nudt.edu.cn)

A Novel Passive-Landmark Based Network Distance Prediction Method

Wu Guofu, Dou Qiang, Ban Dongsong, Dou Wenhua, and Song Lei
(*College of Computer, National University of Defense Technology, Changsha* 410073)

Abstract With the direction of network topology information, the performance of large scale distributed applications could be enhanced greatly. However, if the topology information between nodes is obtained by directly measure, the cost of the probing packets may be more than the gain from the performance improvement. This paper proposes a novel passive landmark based network distance prediction method-PLNDP. The vector of transmission delay from normal node to landmarks is embedded into the metric space $\boldsymbol{R}^n$ by the Lipschitz transformation. After getting the network coordinates, normal nodes use the distance function to compute the distance between coordinates. Then the network distances between nodes is predicted by the distance between nodes' coordinates. Unlike other network coordinates system, landmarks in PLNDP only need to respond to probes passively, while not measuring distances to other landmarks actively. Existing high performance public servers, such as DNS servers and Web servers, can be used as landmarks. So the cost of deployment can be reduced greatly. In order to improve the prediction accuracy, valid landmarks and correctional factor are used in the distance function. Experiment results show that, for several different accuracy metrics, PLNDP is better than classical network distance prediction methods GNP and Vivaldi, especially when some landmarks have been failed.

Key words network distance; passive-landmark; embedding; network coordinates system; distributed applications

摘 要 网络拓扑信息的引导能够显著提高大规模分布式应用程序的性能,然而直接测量节点之间拓扑信息产生的开销远大于其收益.提出一种新的基于被动路标的节点间网络距离预估方法PLNDP,使用Lipschitz变换将普通节点到路标节点的网络延迟映射到度量空间$\boldsymbol{R}^n$,再利用距离函数计算映射后的网络坐标之间的距离,从而预测节点之间的网络距离.PLNDP中路标节点不需要主动探测,可利用Internet上已部署的高性能服务器为之,极大降低部署成本.引入有效路标和修正因子,提高了预测的准确性.实验结果表明,与经典方法GNP和Vivaldi相比,PLNDP在多个性能参数方面具有明显的优势.

关键词 网络距离;被动路标;空间嵌入;网络坐标;分布式应用
中图法分类号 TP393

收稿日期:2009-10-12;修回日期:2010-05-10
基金项目:国家自然科学基金项目(60603061)

在大规模分布式应用中,如果能够获取关键的网络性能(如网络带宽、传输延迟)信息来指导服务节点的选择而不是随机选取服务节点,将显著提高QoS敏感的应用程序的服务质量.网络性能估计技术具有代价小、可扩展性高的特点,具备很高的实用价值.网络传输延迟是重要的网络性能参数,本文主要研究网络传输延迟参数的估计技术.我们使用网络距离描述正常情况下报文在节点之间的网络传输延迟.

网络坐标系统将节点嵌入到有限维度量空间,使用节点在度量空间的距离估算节点间的网络距离.该机制具有以下优点:1)极大降低探测报文的数量,可扩展性好,支持大规模分布式系统;2)支持第三方预测,只需获得节点网络坐标即可预测节点的网络距离.本文提出一种新的基于被动路标的网络距离预测方法PLNDP,利用Internet上已部署的服务器作为路标节点,准确地预测节点之间的网络距离.该方法克服了类似方案中路标必须主动发送探测报文的缺点,利用现有的服务器而不需要部署专门的服务器,降低部署成本.PLNDP计算网络距离快速、准确,对改善大规模分布式系统有重要参考意义.

1 相关工作

网络距离估计技术早在1996年2月IETF相关工作组就提出简单的网络距离估计服务SONAR[1],并在1997年形成更一般的服务HOPS[2].分布式应用的迅速发展使得网络性能估计凸现价值,近几年不断有研究人员提出基于网络坐标系统的网络距离预测方法.

GNP[3]最先采用网络坐标概念预测网络距离,将节点映射到n维向量空间中的元素,使用元素间的欧氏距离来代替节点间的网络距离.GNP首先在网络中布置一些路标节点,路标节点测量相互之间的网络距离,然后使用Simplex Downhill[4]方法计算路标节点的坐标值,使得路标之间的欧氏距离与实际测量的网络距离误差最小.普通节点测量与路标节点之间的网络距离使用Simplex Downhill方法求解自己的坐标,使得节点与路标之间的计算距离与网络距离的误差最小.GNP验证了通过网络坐标系统可以准确估计节点之间的网络距离,但其存在缺点:预测的准确度与路标节点的数量有关系,需要在Internet中专门部署路标节点;使用Simplex Downhill方法计算坐标值收敛速度慢;最终结果因初始值的不同而会产生较大差异.PIC[5]采用与GNP相同的网络距离预测方法,但任何已经计算出坐标的普通节点都可担当路标节点.

Lighthouses[6]采用全局坐标和相对坐标来避免通信瓶颈和单节点失效的问题.节点随机选择已经加入系统的$k+1$个节点,采用GramrSchmidt Process方法计算以某个节点为原点的新坐标系,节点采用类似GNP方法计算在新坐标系的坐标,然后通过坐标变换将新坐标转换为全局坐标,系统需要维护坐标变换矩阵$\boldsymbol{P}$.

ICS[7]和Virtual Landmarks[8]将到其他路标的延迟向量作为路标坐标,对所有路标坐标构成的延迟矩阵进行PCA分析提取主成分矩阵,原始坐标与主成分矩阵相乘获得新坐标.普通节点先测量与路标节点的网络延迟获得延迟向量,然后与主成分矩阵相乘获得其坐标.IDES[9]通过矩阵分解的方法来预测网络距离.每个路标获得两个坐标——输出坐标和输入坐标.A到B的距离使用A的输出坐标与B的输入坐标内积计算.普通节点通过使得估算延迟和测量延迟误差和最小来获得近似坐标.

Vivaldi[10]克服GNP的缺点,没有引入路标节点,而是采用逐步迭代的方法来确定节点坐标.普通节点获取远程节点坐标及到其网络距离后,通过比较预测距离与测量距离决定自身移动方向和移动幅度,节点重复迭代过程直到达到理想的位置.Pharos[11]是在Vivaldi基础上层次式网络距离估算技术.节点加入到全局和局部两个覆盖网络,全局坐标和局部坐标分别使用Vivaldi方法获得.节点位于同一个局部覆盖网络中,则使用局部坐标来计算距离,否则使用全局坐标来计算距离.

文献[12]针对结构化覆盖网络提出分布式节点聚集算法RANRA,利用网络坐标算法确定节点的二维平面坐标,对坐标平面进行等面积的划分,每个子区域均与DHT中的多层命名空间的某区间一一映射,使得在物理拓扑相邻的节点经过映射后在逻辑拓扑也保持相邻.

2 空间嵌入及其评价参数

网络坐标系统实质上是将有限的网络空间嵌入到其他度量空间中,本节介绍空间嵌入的概念,并提出相关评价参数.

2.1 空间嵌入定义

下面先引入度量空间的概念.

度量空间[13] 是一个二元组 $\boldsymbol{M}=(X, d)$, 其中 X 是元素的集合, d 是距离函数, 即 $d: X \rightarrow R^{+}$, 对于任意的元素 $a, b \in X$, a, b 之间的距离由函数 $d(a, b)$ 确定. 对于度量空间, 要求满足以下 3 个性质. 对任意的元素 $a, b, c \in X$:

1) 正定性. $d(a, b) \geqslant 0$, $d(a, b)=0$, 当且仅当 $a=b$.

2) 对称性. $d(a, b)=d(b, a)$.

3) 三角不等式. $d(a, c) \leqslant d(a, b)+d(b, c)$.

Internet 中网络传输延迟存在违背对称性和三角不等式情况, 使得 Internet 不能被视为度量空间. 但是这种情况并不常见, 如果容忍违背度量空间性质的少量情况, 仍可将 Internet 视为度量空间 $\boldsymbol{M}=(X, d)$, 其中 X 是网络中所有节点的集合, $d(a, b)$ 是节点 a, b 之间的网络传输延迟.

定义 **1.** 空间嵌入[13]. 假定 $\boldsymbol{M}_1=(X_1, d_1)$ 和 $\boldsymbol{M}_2=(X_2, d_2)$ 均为度量空间, φ 为从 X_1 到 X_2 的单射函数, 定义 $(\varphi(X_1), d_2)$ 为 $\boldsymbol{M}_1$ 在映射 φ 下到 $\boldsymbol{M}_2$ 的空间嵌入, 简记为 $\varphi(\boldsymbol{M}_1)$.

在不混淆距离函数的情况下, 我们也称映射 φ 将 X_1 嵌入到 X_2 中. 对于任意的 $x, y \in X_1$, 如果都有 $d_1(x, y)=d_2(\varphi(x), \varphi(y))$ 则称映射 φ 下的嵌入是等距的. 如果 X_1 嵌入到 X_2 中, Z 是 X_1 的子集, 映射 φ 同样定义了 Z 到 X_2 的嵌入. 一个简单的例子是 X_1 为有限集, 其上的距离函数任意, X_2 为 n 维向量空间 $\boldsymbol{R}^n$, X_2 上的距离函数为向量空间中的标准欧氏距离, 即

$$d_2(x, y)=\|x-y\|=\sqrt{\sum_{1 \leqslant i \leqslant n}(x_i-y_i)^2}. \quad (1)$$

将 X_1 嵌入到 X_2 中就是寻找合适的单射函数, 从而在空间 X_2 中研究空间 X_1 中的性质.

2.2 空间嵌入质量的评价参数

存在多种评价参数反映嵌入的质量, 其优劣取决于应用需求. 在网络应用中, 有的要求能准确预测原始空间的距离, 有的要求能准确反映出原始空间的相对距离即可. 在后面叙述中, 我们约定 $\boldsymbol{M}_1=(X_1, d_1)$ 为原始度量空间, X_1 为有限集, $\boldsymbol{M}_2=(X_2, d_2)$ 为嵌入度量空间, φ 为单射函数.

1) 失真度[13] (distortion)

在关于嵌入的理论文献[14-15]中, 最常用的评价参数是失真度, 该参数具有向量乘不变性, 即 $distortion(\varphi)=distortion(\alpha\varphi)$ $(\alpha \neq 0)$, 它是嵌入质量最坏情况的描述. 首先定义比值:

$$r(\varphi, x, y)=\frac{d_2(\varphi(x)-\varphi(y))}{d_1(x, y)}. \quad (2)$$

φ 的膨胀, $expansion(\varphi)$ 是比值 $r(\varphi, x, y)$ 的最大值 $(x, y \in X_1$ 且 $x \neq y)$; φ 的紧缩, $contraction(\varphi)$ 是比值 $r(\varphi, x, y)$ 的最小值 $(x, y \in X_1$ 且 $x \neq y)$. 失真度定义为

$$distortion(\varphi)=\frac{expansion(\varphi)}{contraction(\varphi)}. \quad (3)$$

$distortion(\varphi)=1$ 当且仅当存在常数 θ, 使得 $r(\varphi, x, y)=\theta$ 对任意的 x, y 均成立 $(x, y \in X_1$ 且 $x \neq y)$.

2) 重度[13] (stress)

嵌入整体质量的体现则用重度参数, 其表达式为

$$stress(\varphi)=\frac{\sum_{x \neq y \in X_1}(d_2(\varphi(x), \varphi(y))-d_1(x, y))^2}{\sum_{x \neq y \in X_1} d_1(x, y)^2}. \quad (4)$$

如果重度值为 0, 嵌入完全保留原始空间的距离, 即该嵌入是等距的.

3) 相对误差[13] (relative error)

相对误差描述使用嵌入空间中的距离代替原始空间中的距离产生的相对误差. 其表达式为

$$RE(\varphi, x, y)=\frac{|d_2(\varphi(x), \varphi(y))-d_1(x, y)|}{d_1(x, y)} \quad (x \neq y \in X_1). \quad (5)$$

相对误差 $RE=0$ 表示嵌入操作保持距离不变. RE 表征的是特定两个元素间距离在嵌入操作中产生的误差, 要衡量元素 x 到其他所有元素距离在嵌入操作中产生的误差使用平均相对误差 $average\_RE$ 表示, 其表达式为

$$average\_RE(\varphi, x)=\frac{\sum_{y \neq x \in X_1} RE(\varphi, x, y)}{|X_1|-1}. \quad (6)$$

4) k-最近保持率 (k-closest preserve, k-CP)

我们引入 k-最近保持率描述嵌入操作对最近距离节点的保持能力, 假设在原始空间中与节点 x 距离最近的 k 个节点为 $S_1=\{q_1, q_2, \cdots, q_k\}$, 在嵌入空间中与节点 $\varphi(x)$ 距离最近的 k 个节点为 $S_2=\{m_1, m_2, \cdots, m_k, m_i \in \varphi(X_1)\}$, $\varphi(S_1)$ 与 S_2 的交集反映了 k-最近保持率, 其表达式为

$$k-CL(\varphi, x)=\frac{|\varphi(S_1) \cap S_2|}{k}. \quad (7)$$

5) 保序率(order-keep rate)

我们引入保序率反映嵌入操作对相对距离的保持能力,如果 $d_1(x,y) \geqslant d_1(x,z)$ 并且 $d_2(\varphi(x),\varphi(y)) \geqslant d_2(\varphi(x),\varphi(z))$,或者 $d_1(x,y) < d_1(x,z)$ 并且 $d_2(\varphi(x),\varphi(y)) < d_2(\varphi(x),\varphi(z))$,则称 y,z 相对 x 是保序的.给定元素 x,任取元素 y,z,定义 x 相对 y,z 的保序值如下:

$$R(x,y,z)=\begin{cases}1, & \text{如果 } y,z \text{ 相对 } x \text{ 是保序的;}\\ 0, & \text{否则.}\end{cases} \tag{8}$$

则元素 x 的保序率计算如下:

$$order\text{-}keep(\varphi,x)=\frac{\sum_{y,z\in X_1,x\neq y\neq z}R(x,y,z)}{(|X_1|-1)\times(|X_1|-2)}. \tag{9}$$

由于 $x\neq y\neq z$ 且均为 X_1 中的元素,当 x 固定时,由排列知识得到式(9)中分子部分共有 $(|X_1|-1)\times(|X_1|-2)$ 项,归一化得到元素 x 的保序率计算公式.

$order\text{-}keep(\varphi,x)$ 计算的是其他元素相对于 x 的保序率,则用平均保序率来衡量嵌入操作对所有元素的保序能力,

$$average\text{-}order\text{-}keep(\varphi)=\frac{\sum_{x\in X_1}order\text{-}keep(\varphi,x)}{|X_1|}. \tag{10}$$

在网络坐标系统中,原度量空间已经确定,其中 X_1 是全体参与预测的网络节点,d_1 由网络节点之间的网络传输延迟决定.$\boldsymbol{M}_2$ 多采用 n 维向量空间 $\boldsymbol{R}^n$,常用的距离函数为 L_p,即

$$L_p(d_i,d_j)=\left(\sum_{k=1}^{n}|d_{ik}-d_{jk}|^p\right)^{\frac{1}{p}}. \tag{11}$$

给定 $\boldsymbol{M}_1$,$\boldsymbol{M}_2$ 后,空间嵌入问题就是寻找合适的映射函数 φ,将空间 $\boldsymbol{M}_1$ 中的元素映射到空间 $\boldsymbol{M}_2$.映射函数 φ 在很多程度上决定嵌入的质量,φ 的选择应该根据实际应用的需求来优化某些评价参数,例如最小化失真度、相对误差等.

3 基于被动路标的网络距离预测技术 PLNDP

在基于网络坐标系统的预测方案中大都引入路标节点,路标节点或者全局固定不变,或者随机选择普通节点担当.Internet 上已部署大量高性能服务器,如 DNS 服务器、Web 服务器等,这些节点服务能力强,响应端节点的探测报文.如果利用这些节点充当路标节点,既解决全局不变方案中部署成本问题,又解决随机选择方案中误差累积和局部陷入问题.但是这些服务器只被动地响应请求,无法主动对外探测,所以需要特殊方法来利用这些服务器的路标功能.

Internet 上的节点是按域结构组织的,域内节点之间延迟小,域间节点之间延迟大.网络距离小的两个节点到其他节点的网络距离较为接近;网络距离大的两个节点到其他节点的网络距离差异较大.如果直接使用节点到路标的网络延迟作为节点的网络坐标,坐标之间的欧氏距离可以反应出节点之间的网络距离.

鉴于上述两方面认识,我们提出一种新的基于被动路标的网络距离预测方法 PLNDP(passive-landmark based network distance prediction).PLNDP 整体思路是:在 Internet 上搜索已经部署的高性能服务器作为路标节点;普通节点测量到各个路标节点的网络延迟,使用 Lipschitz 变换将延迟向量 $\boldsymbol{r}$ 映射到向量空间 $\boldsymbol{R}^n$,得到自身网络坐标;交换得到其他节点网络坐标,使用距离函数计算坐标间的距离,预测节点间网络距离.下面详细描述 PLNDP 的执行过程.

1) 路标节点的选择.Internet 上已经部署大量的服务器,选择哪些服务器作为路标节点需要解决下面 3 个问题:①路标节点的数量 l.路标数量过多会产生太多的探测报文同时造成信息冗余;路标数量太少使得预测结果偏差过大.后面的实验结果表明,路标数量在 10 个左右就可获得较高的预测准确性.为了应对路标节点失效,我们认为路标节点数量在 16 个比较合适.②路标节点的分布范围.路标节点应该处在合适的位置才能发挥参考作用.路标的位置范围取决于应用程序的使用范围,如果面向整个 Internet,则路标节点应该分布在整个 Internet;如果是面向区域性的,比如面向中国大陆地区,则路标节点应该限制在区域以内.③路标节点的相对位置.距离太近的两个路标提供的信息存在冗余,因此路标节点应该尽可能分散.综合上述考虑在特定范围内搜集性能稳定的大型服务器,从中选择数量合适、均匀分散的服务器作为路标节点,其他留作备份使用.

2) 普通节点坐标获取.普通节点 i 通过 ping 或者 tracert 工具发送 echo request 探测报文测量路标节点的传输延迟.普通节点多次测量 RTT,选择

网络传输延迟的最小值或者平均值来规避网络暂时拥塞带来的影响. 路标节点序号和 IP 地址预先发送给普通节点. 普通节点测得网络延迟后, 使用 Lipschitz 变换将网络延迟嵌入到度量空间 $\boldsymbol{R}^n$, 即按照路标顺序获得的延迟向量作为网络坐标, 失效路标对应的元素为 0. 获取网络坐标后向中心服务器注册自身网络坐标. 为了应对底层网络状况变化, 普通节点定时测量到路标的延迟并更新中心服务器上的数据.

3) 网络距离预测. 普通节点通过中心服务器或者其他节点获取感兴趣节点的网络坐标, 使用距离函数计算网络距离. 假设路标节点的数量为 l, 普通节点 a, b 的坐标分别为$(a_1, a_2, \cdots, a_l)$, $(b_1, b_2, \cdots, b_l)$. 原始坐标进行预处理获取有效坐标: 首先剔除失效路标, 去除失效路标对应的元素得到新坐标, 由于路标节点是性能稳定的服务器, 失效概率极低, 此法完全可应对失效问题; 其次去除冗余路标, 路标仅具有相对参考作用, 如果普通节点 a, b 到路标节点 A, B 的网络传输延时差相当, 则认为 A, B 对于 a, b 的参考作用是相同的, 即 A, B 相对于 a, b 是冗余的, 保留一个即可. 将失效路标和冗余路标对应的坐标元素去除后, a, b 的新坐标变为$(a'_1, a'_2, \cdots, a'_m)$, $(b'_1, b'_2, \cdots, b'_m)$, m 为有效路标节点数目, 则 a, b 之间的距离函数为

$$L_{ab} = \sqrt{\frac{3}{\max(m, 3)} \sum_{i=1}^{m} (a'_i - b'_i)^2}. \quad (12)$$

L_{ab} 距离函数与欧氏距离不同之处是增加了修正因子 $\sqrt{3/\max(m, 3)}$, 如果仅采用欧氏距离, 随着有效路标数量的增多, 预测距离会逐步增大, 添加修正因子后消除该问题.

下面举例说明 PLNDP 的具体操作. 在图 1 所示的网络拓扑中, 选择节点 $ACDEFH$ 作为路标节点, abc 为普通节点, 边上的数字表示传输延迟. 不妨假设路标 E 对普通节点 a 是失效的, 则节点 a, b, c 的网络坐标分别为$(3, 3, 6, 0, 9, 13)$, $(13, 9, 6, 6, 3, 3)$, $(8, 6, 1, 7, 4, 8)$. 计算 a, b 之间的距离时, 剔除失效路标 E 对应的元素得到坐标$(3, 3, 6, 9, 13)$, $(13, 9, 6, 3, 3)$, 无冗余路标, 计算得到 $L_{ab} = 12.78$. 预测 b, c 之间的距离时, b, c 到路标 A, D 的距离差都是 5, 所以 A, D 是相互冗余的, 去掉路标 D 的参考作用, 同样路标 E, F 也是相互冗余的, 去掉路标 F, 得到新坐标$(13, 9, 6, 3)$, $(8, 6, 7, 8)$, 计算得到 $L_{bc} = 6.71$. 计算 a, c 之间的距离时, D, F, H 互为冗余节点, 去除失效和冗余节点得到新坐标为$(3, 3, 6)$, $(8, 6, 1)$, 计算可得 $L_{ac} = 7.7$. a, b, c 之间的实际距离为 $M_{ab} = 12$, $M_{bc} = 7$, $M_{ac} = 7$, 预测误差分别为 6.5%, 4.1%, 10%, 预测距离与实际距离较为接近.

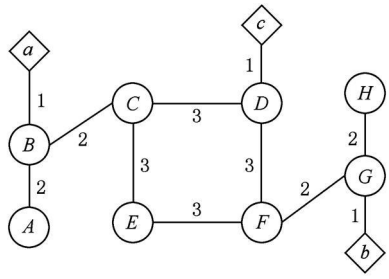


Fig. 1 Random connected network.
图 1 随机连接网络

4 PLNDP 性能验证

为了检验 PLNDP 方法的性能, 本节通过实验比较其与 GNP, Vivaldi 的差异.

4.1 实验设置

我们采用两种不同的数据集检测预测方案的准确性: 一是使用文献[16] 中的方法生成随机拓扑网络; 二是使用 Internet 采集的数据集 King[10]. 默认情况下路标节点数量为 16, 普通节点数量为 500, 路标和普通节点都随机选择; 不同实验场景路标和普通节点数量的变化在下面有详细说明. GNP 采用 Simplex Downhill 方法获取节点的网络坐标, 设置最大的迭代次数为 10^5, 迭代停止条件为绝对误差平方和小于 100. Vivaldi 中设置迭代次数为 200 次, 关键参数 c_e, c_c 分别取 0.25, 0.25.

4.2 实验结果

相对误差描述预测距离与实际距离的偏离是常用的预测准确度评价参数. 图 2 显示三者在仿真生成数据集上相对误差的累积分布, 可以看出 PLNDP 明显优于 GNP, Vivaldi. PLNDP 中 24.7% 的节点, 预测相对误差在 5% 以内, 而 GNP 中只达到 11%; PLNDP 中 90% 的节点预测相对误差在 30% 以内, 而 GNP 中最好的 90% 节点预测相对误差在 55% 以内. 相对误差在 15% 以内 Vivaldi 与 PLNDP 预测精确度相当, 但是前者有相当一部分节点预测精确度很差. 图 3 描述了三者在数据集 King 上的相对误差累积分布, 由于数据集 King 中大部分节点对之间的网络延迟无法测得, 相当于部分路标节点失效, 因此预测准确度没有生成数据集上高. 由于 PLNDP 中采用有效路标计算网络距离, 其效果明显好于 GNP 和 Vivaldi.

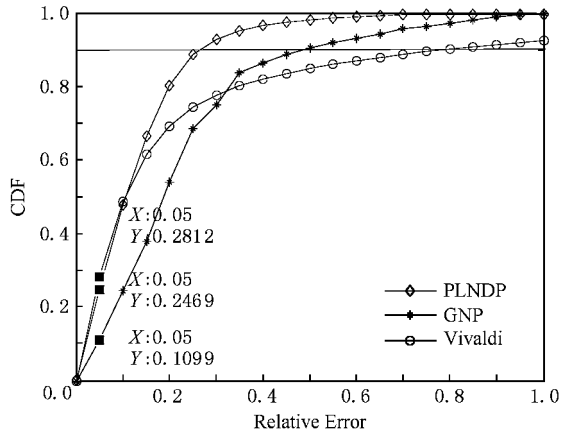


Fig. 2 CDF of relative error on simulated data set.

图 2 相对误差累积分布(仿真生成数据集)

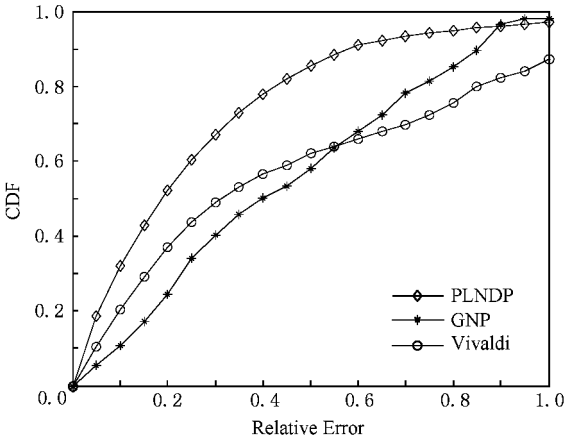


Fig. 3 CDF of relative error on King data set.

图 3 相对误差累积分布(数据集 King)

我们改变系统规模,查看平均相对误差随系统规模的变化. 图 4 显示了三者在生成数据集上平均相对误差的变化. PLNDP 和 GNP 相对误差不受系统规模影响,具备良好的可扩展性; Vivaldi 随系统规模的增大误差逐渐放大,因此其要保证准确性必须随系统规模增加迭代次数. 图 5 显示三者在数据集 King 上相对误差随系统规模的变化,与图 4 具有相似的曲线走势. 图 4、图 5 都显示 PLNDP 在预测准确性方面优于 GNP 和 Vivaldi.

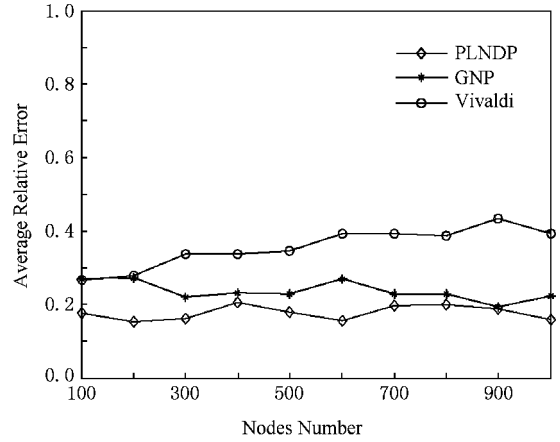


Fig. 4 Average relative error on simulated data set.

图 4 平均相对误差的变化(仿真生成数据集)

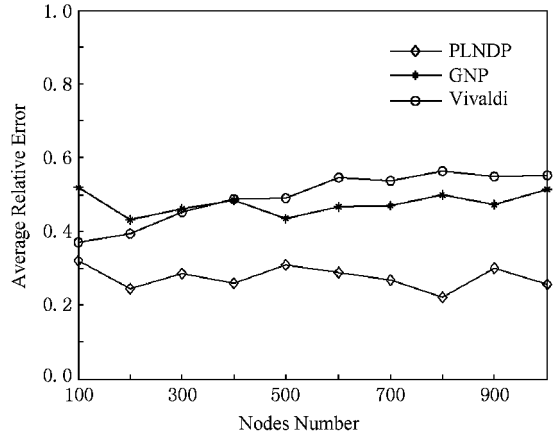


Fig. 5 Average relative error on King data set.

图 5 平均相对误差的变化(数据集 King)

图 6、图 7 显示了三者在两种不同数据集上 k-最近保持率上的差别,我们统计了 $k=20$ 时每个节点的 k-最近保持率. 在合成数据集上, PLNDP 和 GNP 都保持很高的预测准确度, PLNDP 略胜于 GNP,但 Vivaldi 在该参数上表现一般. 在合成数据集上 PLNDP 平均 20-最近保持率为 86.1%, GNP 为 84.3%, 而 Vivaldi 仅有 67.2%. 这意味着在前 20 个最近的节点中, PLNDP 和 GNP 预测命中 18 个,而 Vivaldi 预测命中 13 个. 在数据集 King 上, PLNDP 表现更为优秀,显著好于其他两者. GNP 受到失效路标的影响,精确度明显下降. 三者的平均 20-最近保持率分别为 83.4%, 37.2%, 41.4%.

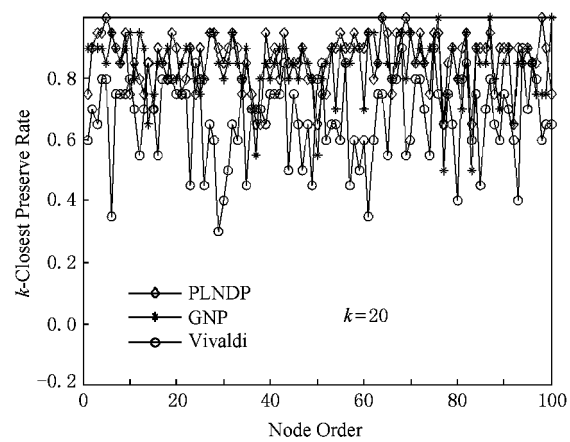


Fig. 6 k-closest preserve rate on simulated data set.

图 6 k-最近保持率(仿真生成数据集)

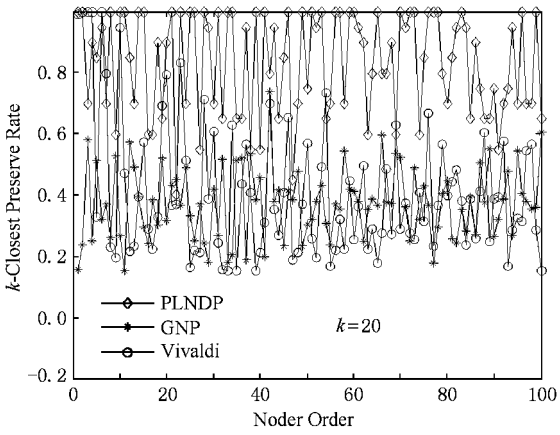


Fig. 7 k closest preserve rate on King data set.

图 7 k 最近保持率(数据集 King)

保序率描述对相对距离的保持能力, 由于数据集 King 中大部分节点对之间没有网络延迟数据, 因此我们仅比较生成数据集上三者的保序率. 图 8 和图 9 描述三者保序率的差别, 三者都表现了很强的保序性, 保序率分别达到 92.7%, 95.9%, 89.4%.

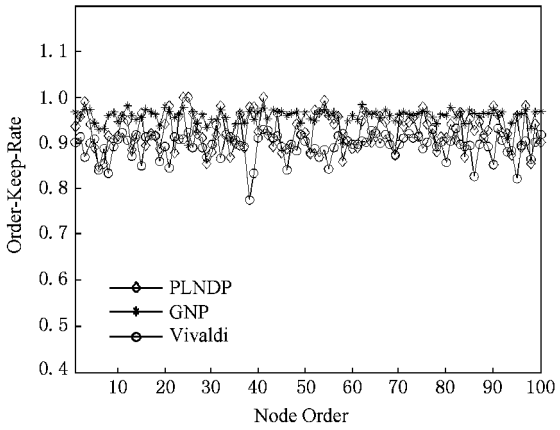


Fig. 8 Order-keep rate on simulated data set.

图 8 不同节点的保序率(仿真生成数据集)

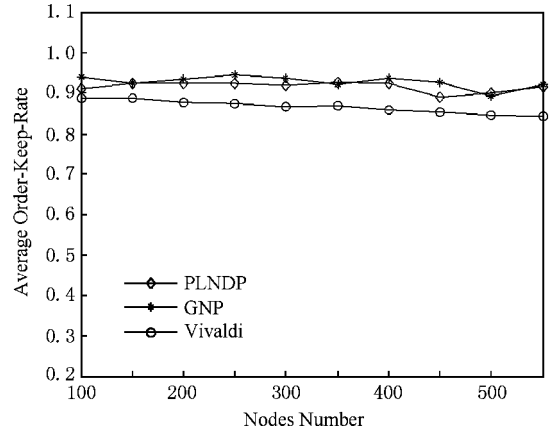


Fig. 9 Average order-keep rate on simulated data set.

图 9 不同规模下的平均保序率(仿真生成数据集)

PLNDP 略逊于 GNP, Vivaldi 相对最差. 图 8 显示了每个节点的保序率, 而图 9 显示了平均保序率随系统规模的变化, 保序率随系统规模变化保持不变, 都具有很高的预测准确性.

4.3 补充说明

4.2 节的实验结果是在网络拓扑结构和传输延迟不变的情况下获得, 实际 Internet 中底层网络结构和传输延迟都随时间变化. 在动态网络环境情况下, PLNDP 的预测性能会产生怎样的变化呢, 我们将动态网络看作由一系列网络快照组成, 快照的持续时间为 τ, 在时间 τ 内, 网络变化可忽略. PLNDP 每隔周期 T 更新节点的网络坐标, 如果限制 $T<\tau/2$, 则保证测量和预测在同一网络快照内完成, 仍然可以看作 PLDNP 在静态网络中运行. 我们认为只要合理设置更新周期 T, PLDNP 在实际网络中仍保持高的预测准确性.

在大规模网络应用中, 可扩展性是一项重要指标. 假设系统规模为 10 万个节点, 更新周期 T 为 10 min, 每次测量到路标的传输延迟发送 5 次 echo request 报文, 路标节点数量为 20, 则路标节点平均 1.2 ms 响应一次请求, 对于高性能稳定服务器而言属于极轻负载; 假设请求应答报文的长度为 32B, 则对整个网络产生的流量为 53.3 Kbps. 中心服务节点主要负责更新节点的网络坐标和响应普通节点对其他节点坐标的请求, 每隔 6 ms 收到一次更新报文, 假设报文长度为 80 B, 需要消耗的带宽为 13.3 Kbps, 服务器完全可承载. PLNDP 产生的负载随系统规模线性增加, 具备良好的可扩展性.

5 结 论

网络距离信息对提高大规模分布式应用性能有显著效果, 直接测量节点之间的网络距离产生的开销远大于其收益. 基于网络坐标的网络距离预测技术较为准确地估计节点之间的网络传输延迟, 能够满足大多数分布式应用的需求. 本文提出的基于被动路标的网络距离预测方法 PLNDP 与以往方法最大的不同之处是 PLNDP 只需路标节点被动响应探测报文. 利用 Internet 上现有的高性能服务器作为路标节点可减少部署成本, 提高部署速度. 引入有效路标和距离修正因子提高了预测的准确性. 本文提出 k-最近保持率和保序率评价参数, 更好地满足某些应用的需要. 实验结果表明, PLNDP 在多个性能

评价参数上都具有较高的预测准确性,对大规模分布式应用具有积极的参考价值.

参 考 文 献

[1] Moore K, Cox J, Green S. Sonar—A network proximity service[EB/OL]. 1996 [2009-08-06]. http://www.netlib.org/utk/projects/sonar/

[2] Francis P. Host proximity service (Hops)[EB/OL]. 1998. [2009-08-06]. http://datatracker.ietf.org/wg/hops/

[3] Ng T S E, Zhang Hui. Predicting Internet network distance with coordinates-based approaches [C] //Proc of Infocom 2002. Los Alamitos, CA: IEEE Computer Society, 2002: 170-179

[4] Nelder J A, Mead R. A simplex method for function minimization[J]. Computer Journal, 1965, 7(1): 308-313

[5] Manuel C, Miguel C, Antony R, et al. PIC: Practical Internet coordinates for distance estimation [C] //Proc of ICDCS 2004. Los Alamitos, CA: IEEE Computer Society, 2004: 178-187

[6] Marcelo P, Jon C, Steve W, et al. Lighthouses for scalable distributed location [C] //Proc of IPTPS 2003. Berlin: Springer, 2003: 278-291

[7] Hyuk L, Jennifer C H, ChongHo C, et al. Constructing Internet coordinate system based on delay measurement [J]. IEEE/ACM Trans on Networking, 2005, 13(3): 513-525

[8] Tang Liying, Mark C. Virtual landmarks for the Internet [C] //Proc of IMC 2003. New York: ACM, 2003: 143-152

[9] Mao Yun, Saul L K. Modeling distance in large-scale networks by matrix factorization [C] //Proc of IMC 2004. New York: ACM, 2004: 278-287

[10] Frank D, Russ C, Frans K, et al. Vivaldi: A decentralized network coordinate system [C] //Proc of ACM SIGCOMM 2004. New York: ACM, 2004: 15-26

[11] Chen Yang, Xiong Yongqiang, Shi Xiaohui, et al. Pharos: Accurate and decentralised network coordinate system [J]. IET Communications, 2009, 3(4): 539-548

[12] Duan Hancong, Lu Xianliang, Tang Hui, et al. Topology-aware node rendezvous algorithm based on DHT [J]. Journal of Computer Research and Development, 2006, 43(10): 1790-1796 (in Chinese)
(段翰聪, 卢显良, 唐晖, 等. 基于DHT的拓扑感知节点聚集算法[J]. 计算机研究与发展, 2006, 43(10): 1790-1796)

[13] Lua E K, Griffin T, Pias M. On the accuracy of embeddings for internet coordinate systems [C] //Proc of IMC 2005. Berkeley, CA: USENIX Association, 2005: 1-11

[14] Gupta A. Embedding tree metrics into low dimensional euclidean spaces [C] //Proc of the 31st Annual ACM Symp on Theory of Computing. New York: ACM, 1999: 694-700

[15] Linial N, Saks M. The Euclidean distortion of complete binary trees [J]. Discrete and Computational Geometry, 2003, 29(1): 19-21

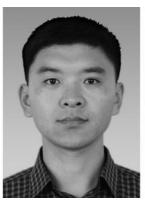
[16] Zegura E, Calvert K, Bhattacharjee S. How to model an internetwork [C] //Proc of INFOCOM96. Los Alamitos, CA: IEEE Computer Society, 1996: 594-602

Wu Guofu, born in 1980. Received his B.A's and M.A's degrees in computer science from the National University of Defense Technology, Changsha, China, in 2003 and 2005 respectively. Since 2005, he has been a PhD degree candidate in computer science from the National University of Defense Technology. His current research interests include peer to peer network and Internet routing technology.

Dou Qiang, born in 1973. PhD in the National University of Defense Technology. His main research interests include high performance computing and real-time system.

Ban Dongsong, born in 1982. Since 2006, he has been a PhD candidate in computer science from the National University of Defense Technology. His current research interests include sensor networks.

Dou Wenhua, born in 1946. Professor and PhD supervisor in the National University of Defense Technology. His main research interests include high performance computing and advanced computer network.



Song Lei, born in 1976. Since 2005, he has been a PhD candidate in computer science from the National University of Defense Technology. His current research interests include sensor networks and security of WLAN.