基于时间特征的网络流量预测模型*)

叶新铭 王 斌
(内蒙古大学计算机学院 呼和浩特 010021)

摘 要 本文设计一种基于时间特征的网络流量预测模型,并采用该流量模型预测网络流量。文章提出网络流量预测误差的数学定义,根据测试实验表明,我们的流量模型具有更高的可用性,并适用实际运行的网络环境。
关键词 流量模型,流量预测,预测误差

A Network Traffic Forecast Model Based on Time Character

YE Xin-Ming WANG Bin
(Computer College of Inner Mongolia University, Huhehaote 010021)

Abstract This paper proposes a forecast traffic model correlative to time character. Network traffic is forecasted under the traffic model. We also present a mathematics definition of network traffic forecast error. The data of experiment of network measurement shows that the traffic model has the advantage of more forecasting precision which is defined strictly. The traffic model is very suitable to practical operational networks.
Keywords Traffic model, Traffic forecasting, Forecast error

1 引言

在通信网络技术发展的30年里流量模型研究一直备受人们关注。20世纪70年代和80年代的早期人们主要借鉴PSTN的流量模型,用Poisson模型来描述数据网络的流量模型[1],一般称其为经典流量模型,基本假设为:

1) 外部数据源产生流量的时间间隔为指数分布,即数据源到达过程为一Poisson过程,令$\{G(i)|I=1,2,\cdots,N\}$,$G(i)$为数据包i和$i+1$的间隔时间;

2) 数据源一次产生流量的长度服从指数分布,令$\{H(i)|I=1,2,\cdots,N\}$,$H(i)$为数据包i的数据长度;

3) $G(i)$和$H(i)$相互独立。

但根据这个模型的实验测试结果并不是人们所期望的[2]。Leland等人通过对LAN的流量分析[3]和Klivansky等人对WAN流量的测试分析[4]独立发现流量的自相似性。20世纪90年代以来网络节点数的指数式增加和新的应用(例如VoD, VoIP等)出现增加了网络流量特征化的困难,特别是不同的网络应用具有不同的流量特征,WWW、FTP、VoD、VoIP等流量特征和QoS需求的差异以及不同比例流量迭加使得传统PSTN统计流量特征不再适用分析数据通信网络流量。在数据通信网络技术发展的20多年以来,网络研究者意识到统计模型越来越不适合表示数据通信网流量特征,但是还是未能理解数据通信的流量行为及其造成这种行为的原因与影响因素[5]。

后来人们应用排队论方法进行容量规划和性能预测。排队分析的有效性依赖于数据通信量的泊松性质。人们发现排队分析所得到的预测结果和实际的情况相差很多。最近几年有几项研究表明在某些环境中,通信量分布是自相似的而不是泊松的。自相似分析方法的理论基础是分形和混沌[6]。图1为实际网络通信量与应用泊松模型和自相似模型预测的网络通信量的对比,采集时间在27.7个小时内以100秒为单位。从图1可以看出泊松模型和自相似模型预测的结果和实际情况相差很多。关于这两种方法的比较见第4节。我们认为以上这几种都没有考虑到通信量的时间特征因素。

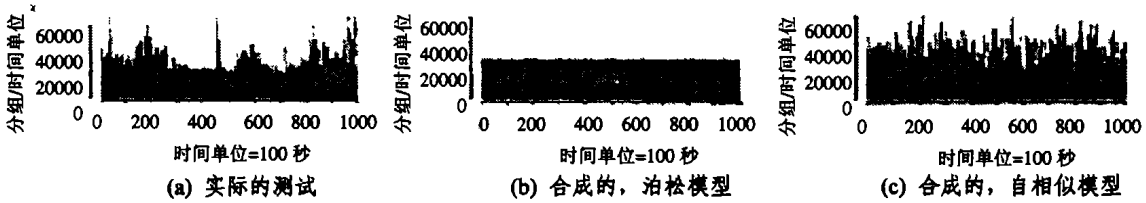

图1 实际的网络流量与应用泊松模型和自相似模型预测的网络通信量

我们试图对网络流量加上时间特征,这样可以解决以下两个问题:1) 通信量的周期性;2) 任意时刻的网络流量预测。具体的模型建立见第3部分。

2 流量数据的采集

我们采集的数据主要来自设备上运行的SNMP协议MIB的Interfaces组。数据采集主要由三部分组成:管理者、代理和MIB。MIB遵从SMI(Structure of Management Information)[7],存放设备或者网络运行状态的信息。管理者通过GetRequest, GetNextRequest, SetRequest等操作通过代理获得和设置MIB的参数值,基本模型如图2所示。

*)基金项目:国家自然基金项目(60263002),内蒙古科技攻关项目(2002061002)。叶新铭 教授,博士生导师,主要研究方向:计算机网络;王 斌 硕士研究生,主要研究方向:计算机网络。

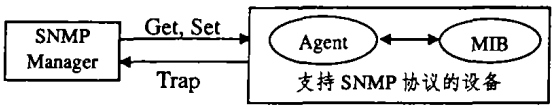

图 2　数据采集模型

每一项数据在 MIB 库中都对应一个唯一标识 ID(mibObj oid)，比如我们要察看设备的制造商则需要用到 system 组的 sysObjectID，它的 mibObj oid 是"1.3.1.2.1.1.2"。通过发送 SNMP 报文可以接收到设备返回的设备制造商数据。流量相关主要采集参数为 Interfaces 组中的下列对象：

· ifInOctets：该接口接收的总字节数，累计数目，在达到最大值时自动清零；

· ifOutOctets：该接口发送的总字节数，累计数目，在达到最大值时自动清零；

· ifInUcastPkts：该接口接收的单播数据包数，累计数目，在达到最大值时自动清零，可以由此参数统计出在某时间段该接口接收到的单播数据包数。

· ifOutUcastPkts：该接口发送的单播数据包数，累计数目，在达到最大值时自动清零，可以由此参数统计出在某时间段该接口发送出的单播数据包数。

· ifInNUcastPkts：该接口接收的非单播数据包数(组播和广播)，累计数目，在达到最大值时自动清零，可以由此参数统计出在某时间段该接口接收的非单播数据包数；

· ifOutNUcastPkts：该接口发送的非单播数据包数(组播和广播)，累计数目，在达到最大值时自动清零，可以由此参数统计出在某时间段该接口发送的非单播数据包数。

采集到上述数据我们可以计算接口的数据传输速率和数据包传输速率。

数据传输速率 Tr(bps)为 ΔifInOctets (t 时刻与 $t+\Delta t$ 时刻 ifInOctets 值之差)与 ΔifOutOctets (t 时刻与 $t+\Delta t$ 时刻 ifOutOctets 值之差)的和除以 Δt，如式(1)所示

$$Tr=(\Delta\text{ifInOctets}+\Delta\text{ifOutOctets})*8/\Delta t \quad (1)$$

数据包传输速率 Tp (pps) 为 ΔifInUcastPkts、ΔifInNUcastPkts、ΔifOutU-castPkts、ΔifOutNUcastPkts 的和除以 Δt，如式(2)所示，其中 ΔifInUcastPkts 为 t 时刻与 $t+\Delta t$ 时刻 ifInUcastPkts 值之差，ΔifInNUcastPkts、ΔifOutUcastPkts、ΔifOutNUcastPkts 同理。

$$Tp=(\Delta\text{ifInUcastPkts}+\Delta\text{ifInNUcastPkts}+\Delta\text{ifOutUcastPkts}+\Delta\text{ifOutUcastPkts})/\Delta t \quad (2)$$

3　模型建立

我们通过观察设备每个接口每天的流量，发现存在着特定的规律。比如这个星期的星期一和上个星期的星期一，接口流量图呈现出一个大致相同的波形。我们认为这个波形其实是接口流量的周期性流量。这个波形可以反映出流量的大致趋势。但在任意时刻流量的陡增与这个波形具有较大的差别。所以我们需要一个分量来校正周期性流量以得到一个比较准确的预测值。这个分量我们称作变化趋势分量。

在对具体网络进行测试的基础上，得出了一个网络流量预测的方法，以预测某个时刻的网络流量。我们认为预测流量 $\overline{T}_{k,l}(i,j)$ 是由变化趋势分量 $V_{k,l}(i,j)$ 和周期性分量 $P_{k,l}(i,j)$ 之和决定，如(1)式所示：

$$\overline{T}_{k,l}(i,j)=V_{k,l}(i,j)+P_{k,l}(i,j) \quad (1)$$

其中 $\overline{T}_{k,l}(i,j)$ 表示链路 l 第 j 个星期的星期 k 的 i 时刻的预测流量，$V_{k,l}(i,j)$ 第 j 个星期的星期 k 的 i 时刻流量变化趋势因素，$P_{k,l}(i,j)$ 表示去除流量变化趋势因素外的周期性流量。

考虑到监测服务器的计算能力和性能要求，我们只使用链路 l 前三个星期的实际流量 $T_{k,l}(i,j-1)$、$T_{k,l}(i,j-2)$ 和 $T_{k,l}(i,j-3)$ 作为预测的依据来计算第 j 个星期的预测流量 $\overline{T}_{k,l}(i,j)$。即：$\overline{T}_{k,l}(i,j)=f(T_{k,l}(i,j-1),\ T_{k,l}(i,j-2),\ T_{k,l}(i,j-3))$。

首先我们使用时间序列平滑预测法[8]计算出设备每日的周期性流量。$P_{k,l}(i,j-1)$ 计算方法如式(2)：

$$P_{k,l}(i,j-1)=\alpha\ T_{k,l}(i-1,j-1)+(1-\alpha)\ P_{k,l}(i-1,j-1) \quad (2)$$

α 为指数平滑值，它决定计算出的 $P_{k,l}(i,j-1)$ 是否能很好地反映流量的周期性特征。利用公式(3)计算实际流量值 $T(i)$ 与周期性流量 $P(i)$ 的均方差 MSE：

$$MSE=\frac{1}{24*3600/\Delta t}\sum_{i=1}^{24*3600/\Delta t}(T(i)-P(i))^2 \quad (3)$$

我们选择使得方差 MSE 最小时的指数平滑值 α 来计算 $P_{k,l}(i,j-1)$，这样计算出来的 $P_{k,l}(i,j-1)$ 能比较好地反映流量的周期性。我们将前三周星期 k 设备的周期性流量 $P_{k,l}(i,j-1)$、$P_{k,l}(i,j-2)$ 和 $P_{k,l}(i,j-3)$ 求平均值得到 $P_{k,l}(i,j)$。如式(4)所示：

$$P_{k,l}(i,j)=(P_{k,l}(i,j-1)+P_{k,l}(i,j-2)+P_{k,l}(i,j-3))/3 \quad (4)$$

然后我们计算流量变化趋势分量 $V_{k,l}(i,j-1)$，我们认为它是实际流量 $T_{k,l}(i,j-1)$ 与通过时间序列平滑预测法计算出的周期性流量 $P_{k,l}(i,j-1)$ 的差，计算公式如式(5)所示。

$$V_{k,l}(i,j-1)=T_{k,l}(i,j-1)-P_{k,l}(i,j-1) \quad (5)$$

将计算出来的 $V_{k,l}(i,j-1)$、$V_{k,l}(i,j-2)$、$V_{k,l}(i,j-3)$ 三个值求算术平均得到 $V_{k,l}(i,j)$，如式(6)所示：

$$V_{k,l}(i,j)=(V_{k,l}(i,j-1)+V_{k,l}(i,j-2)+V_{k,l}(i,j-3))/3 \quad (6)$$

最后将式(4)和式(6)计算出的 $P_{k,l}(i,j)$ 和 $V_{k,l}(i,j)$ 相加，得到我们的预测流量 $\overline{T}_{k,l}(i,j)$。

另外我们取出 $V_{k,l}(i,j-1)$、$V_{k,l}(i,j-2)$ 和 $V_{k,l}(i,j-3)$ 中的最大值、最小值分别和周期性流量 $P_{k,l}(i,j-1)$ 求和作为流量的悲观预测流量 $p\overline{T}_{k,l}(i,j)$ 如式(7)所示和乐观的预测流量 $o\overline{T}_{k,l}(i,j)$ 如式(8)所示。悲观预测流量 $p\overline{T}_{k,l}(i,j)$ 和乐观的预测流量 $o\overline{T}_{k,l}(i,j)$ 对于我们分析设备是否能够达到实际网络的性能要求具有非常重要的意义。

$$p\overline{T}_{k,l}(i,j)=P_{k,l}(i,j)+Max(V_{k,l}(i,j-1),\ V_{k,l}(i,j-2),V_{k,l}(i,j-3)) \quad (7)$$

$$o\overline{T}_{k,l}(i,j)=P_{k,l}(i,j)+Min(V_{k,l}(i,j-1),\ V_{k,l}(i,j-2),\ V_{k,l}(i,j-3)) \quad (8)$$

4　流量预测实验

为了验证我们网络流量模型的精度，我们以时间间隔为 $\Delta t=300$s 采集交换机的接口 MIB 参数，采集的时间为 3 月 4 日到 3 月 26 日。测试环境如图 3 所示，主要包括一台核心交换机 Catalyst 6509、两台交换机 Catalyst 3000、Catalyst 1200。Catalyst 6509 为内蒙古大学校园网的核心交换机，它提供内大网络服务，主要任务包括 WEB 服务器、FTP 服务器、DNS 服务器和各学院接入教育网的服务，是流量预测的主要对象。流量采集和预测任务由一台接在 Catalyst 1200 下的

PC机NMS来完成。

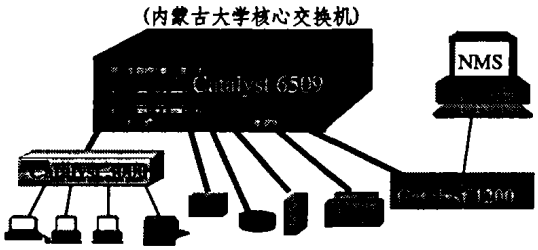

图3 流量预测模型测试环境

表1为3月4日,11日,18日三天0点至4点接口每5分钟的流量值。图4为3月4日,11日,18日三天0点至8点的接口流量图。根据我们的模型,使用上述三天0点至8点的流量数据预测3月25日0点至8点的网络流量。表2为3月25日0点至4点接口每5分钟的实际和预测流量值,图5为3月25日0点至8点的预测流量曲线,图6为3月25日0点至8点的实际流量曲线。

表1 3月4日,11日,18日三天0点至4点接口每5分钟的流量值,单位为kB

3月4日0点至4点接口流量	
0点至1点:1.7, 1.6, 1.4, 1.4, 1.5, 1.6, 1.7, 1.5, 1.5, 1.5, 1.4, 1.4	1点至2点:1.5, 1.4, 1.6, 1.3, 1.9, 1.4, 1.6, 1.8, 1.7, 1.4, 2.0, 1.8
2点至3点:1.5, 1.5, 1.7, 2.0, 1.7, 1.7, 2.1, 2.9, 2.0, 2.1, 2.9, 3.0	3点至4点:2.0, 1.9, 1.5, 1.7, 1.5, 1.5, 1.5, 1.8, 1.5, 1.6, 1.6, 1.6
3月11日0点至4点接口流量	
0点至1点:1.5, 1.5, 1.3, 1.4, 1.5, 1.7, 1.7, 1.9, 1.7, 1.4, 1.3, 1.4	1点至2点:1.4, 1.8, 1.6, 1.8, 1.3, 1.7, 1.7, 1.4, 1.5, 1.7, 1.5, 1.8
2点至3点:1.7, 2.0, 1.8, 1.4, 1.6, 2.0, 1.8, 2.4, 2.3, 2.5, 2.8, 2.5	3点至4点:2.2, 1.6, 1.5, 1.6, 1.8, 4.3, 2.3, 2.0, 1.4, 1.7, 1.4, 1.4
3月18日0点至4点接口流量	
0点至1点:1.3, 1.3, 1.5, 1.3, 1.4, 1.3, 1.4, 1.4, 1.3, 1.3, 1.7, 1.8	1点至2点:1.4, 1.5, 1.3, 1.3, 1.2, 1.3, 1.4, 1.6, 1.5, 1.7, 1.6, 1.3
2点至3点:1.5, 1.8, 1.5, 1.6, 1.9, 1.5, 1.6, 2.2, 2.4, 2.5, 4.3, 2.4	3点至4点:1.8, 1.7, 2.4, 2.1, 1.8, 2.1, 1.6, 1.6, 1.5, 1.3, 1.5, 1.4

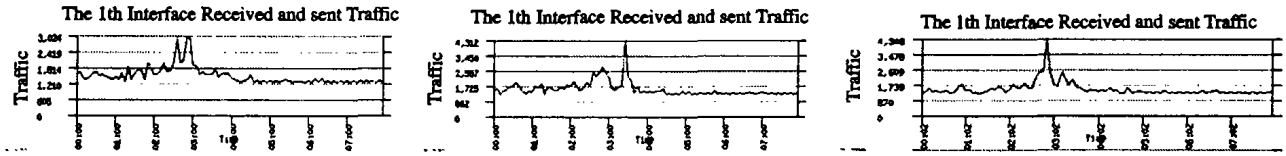

图4 3月4日,11日,18日三天的接口流量图

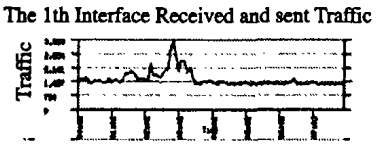

图5 预测流量变化曲线

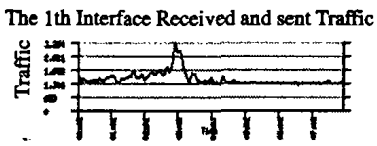

图6 实际流量变化曲线

表2 3月25日0点至4点接口每5分钟的实际和预测流量值,单位为kB

3月25日0点至4点接口预测流量	
0点至1点:1.3, 1.5, 1.5, 1.6, 1.4, 1.3, 1.5, 1.4, 1.5, 1.3, 1.4, 1.6	1点至2点:1.4, 1.5, 1.5, 1.3, 1.5, 1.9, 1.9, 1.9, 1.8, 1.5, 1.5, 1.5
2点至3点:1.5, 1.4, 2.3, 1.7, 1.8, 1.6, 1.8, 2.0, 2.1, 3.2, 3.6, 2.8	3点至4点:1.9, 2.5, 2.4, 1.8, 2.2, 1.6, 1.3, 1.3, 1.3, 1.4, 1.4, 1.5
3月25日0点至4点接口实际流量	
0点至1点:1.8, 1.3, 1.5, 1.4, 1.3, 1.4, 1.4, 1.4, 1.5, 1.4, 1.5, 1.6	1点至2点:1.6, 1.6, 1.3, 1.4, 1.5, 1.5, 1.6, 1.7, 1.9, 1.6, 1.5, 1.7
2点至3点:1.4, 1.6, 1.8, 1.8, 1.7, 1.9, 1.8, 1.7, 2.1, 1.8, 2.2, 3.3	3点至4点:2.8, 2.8, 2.0, 1.6, 1.3, 1.8, 1.7, 1.5, 1.3, 1.4, 1.4, 1.4

为了衡量模型的精确性,我们定义网络流量预测误差。

定义:网络流量预测误差:$E=\frac{1}{24*3600/\Delta t}\sum_{i=1}^{24*3600/\Delta t}|T_{k,l}(i,j)-T_{k,l}(i,j)|/T_{k,l}(i,j)*100\%$,其中$T_{k,l}(i,j)$为实际网络流量,$T_{k,l}(i,j)$为预测网络流量。

经过计算我们的网络预测流量与实际流量的误差为23.38%。图1中提到的泊松模型和自相似模型预测的误差分别为30.2%和48.9%。表3为图1中实际流量值、泊松模型

和自相似模型预测的网络流量值。

表 3 实际流量值、泊松模型和自相似模型预测的网络流量值

27.7 个小时内实际每 100 秒的通信量
4.2,3.8,4,5,7,3.8,3,4,3.5,3.9,4,3,3.7,3.9,4,3.7,3,3.7,4.2,4.8,5,6.1,5,4.3,3.8,4.3,3.7,3.9,2.1,3.7,2.2,2.4,2.5,2.4…
2.44,2.4,6.1,4.8,2.4,2.4,2.4,4.1,3.8,4.2,4.5,4.8,4.8,2.6,4,2.4,4.1,3.1,4.8,5.8,4.5,4.1,2.4,4.1,4.3,4,6,4.2,6.1,
27.7 个小时内泊松模型每 100 秒的预测流量值
2.8…….2.8
27.7 个小时内自相似模型每 100 秒的预测流量值
3.4,4.7,3.1,4,3.7,3.9,4.1,3.1,5,4.1,3.8,3.2,3.8,3,5,5.8,5.2,3.6,5.7,6,2.8,4.1,3.7,4.5,5.8,5,4.2,4.1,3.7,5,5.1,4,3.1,
4.5,3.8,5.1,5,4.2,3,5.8,4.1,3.7,4.1,4.1,5,4.3,4.8,3.7,5.7,5.3,37,3.7,4.2,5.2,5.2,3.4,3.5,3.4,4.2,4.8,4,2.1,3.7,4.1,4.2

使用这种方法预测的优点如下：

1. 计算量少，当从设备上取回第 2 个时间点的数据时就可以开始使用移动平均法来计算实际流量的周期性流量。一台核心交换机可能有上百个千兆接口，性能数据量是很大的。而我们的监测服务器可以由一台普通的计算机来完成。在大型的网络环境中如果实施分布式监测系统，那么可能需要多台网络监测服务器。这样采用我们的方法可以减少网络管理成本。

2. 预测流量对于性能报警和峰值监测具用非常有用的参考价值。

3. 这种方法可以反映出流量的时间趋势及其变化，而不只是使用统计方法来反映流量的变化趋势。

结束语 高性能的网络协议设计、网络设备的设计和制造、网络性能评价等必须依靠精确的网络流量模型。我们采用实际网络测试的方法建立基于时间相关的网络流量模型，并应用该模型进行运行网络的流量预测。

参 考 文 献

1 Fuchs E, Jackson P E. Estimates of Distributions of Random Variables for Certain Computer Communication Traffic Models. Comm. of ACM, Dec. 1970, 13(12): 752～767

2 Duffy D E, Mcintosh A A, Rosenstain M, Willinger W. Statistical analysis fo CCSN/SS7 Traffic Data from Working CCS Subnetworks. IEEE Journal of Selected Areas in Communication, 1994, 12(3): 544～551

3 Leland W E, Taqqu M S, Willinger W, Wilson D V. On the Self-Similar Nature of Ethernet Traffic. IEEE/ACM Transactions on Networking, Feb. 1994, 2(1): 1～15

4 Klivansky S K, Mukherjee A, Song C. On Long Range Dependence in NSFNET Traffic: [Technical Report GIT-CC-94/61]. Geogia Institute of Technology, Altlanta, GA 30332, USA, Dec. 1994

5 Jain R. ATM Networking: Issues and Challenges Ahead. Networld + InterOp'95 Engineer Conf. Las Vegas, Nevada, March 1995. 27～31

6 郑大钟，赵千川. 离散事件动态系统. 清华大学出版社，2001. 163～167

7 Case J, Fedor M, Schoffstall M, Davin J. A Simple Network Management Protocol (SNMP). RFC1157, IETF, May 1990

8 徐国祥，胡清友. 统计预测和决策. 上海财经大学出版社，1998. 113～118

(上接第 19 页)

4 次网络操作、从一个客户端到另一个客户端的数据发送过程、1 毫秒初始化工作)。而在新的通信模型中，只需要 2X 毫秒。因为系统已经事先进行了相关的初始化工作。

Internet 上，来回时延(Round-Trip Delay)大概在几毫秒到几百毫秒之间[6]。表 1 对两种模型进行了比较：

表 1 效果比较

单程时延	1ms	50ms	200ms
传统模型	7ms	301ms	1201ms
新模型	2ms	100ms	400ms

结论与展望 文章基于传统的虚拟环境的多服务器 C/S 模型及兴趣区管理技术提出了一个面向 Internet 的 2.5 维虚拟环境的通信模型。该模型在 2003 年浙江省网上旅游交易会中得到了应用，并且取得了很好的效果。

目前的模型只能用于 AOI 相对固定的虚拟环境，下一步工作应该考虑适合于不同 AOI 的系统中的应用；另外，多个服务器间的协调，并进行动态的划分问题也需要进行相应的研究。

参 考 文 献

1 史美林，向勇，杨光信，等. 计算机支持的协同工作理论与应用[M]. 北京：电子工业出版社，2000

2 Shimamura J, Takemura H, Yokoya N, Yamazawa K. Construction and Presentation of a Virtual Environment Using Panoramic Stereo Images of a Real Scene and Computer Graphics Models [A]. In: Proc. 15th Int. Conf. on Pattern Recognition (15ICPR), Barcelona Spain, Vol. 4, Sep. 2000. 463～467

3 Fox D. Tabula Rasa A Multi-scale User Intl. System. http://www.foxthompson.net/dsf/diss/diss.html

4 Breiteneder C. Lookmark: A 2.5D Web Information Visualization System. In: Proc. of EURASIA-ICT Conf., Teheran, Iran, 2002

5 Funkhouser T A. Network Topologies for Scalable Multi-User Virtual Environments [A]. In: Proc. 1996 IEEE Virtual Reality Annual International Symposium (VRAIS96), San Jose, CA, April, 1996

6 Funkhouser T A. RING: A Client-Server System for Multi-User Virtual Environments. ACM SIGGRAPH Special Issue on 1995 Symposium on Interactive 3D Graphics, New York, 1995. 85～92

7 Morse Katherine L, et al. Interest management in large-scale distributed simulations [R]. Irvine: University of California: [ICS-TR-96-27]. 1996

8 Hagsand O, Marsh I, Hanson K. Sicsophone: a low-delay internet telephony tool [A]. In: Proc. of the 29th Euromicro Conference, Sept 1-6, 2003. 189～195