P2P网络文件分发过程及影响传播因素研究

陈宝钢[1,2] 许 勇[2] 胡金龙[2]
(河南农业大学信息与管理科学学院 郑州 450002)[1]
(华南理工大学广东省计算机网络重点实验室 广州 510641)[2]

摘 要 在P2P网络中,流行文件的下载行为类似于流行病的传播过程,因此可以通过传染病动力学来描述P2P文件共享系统中文件的分发过程。利用传染病动力学模型,建立了一种P2P文件分发模型,并推导了确定文件传播能力的基本再生数计算公式。实验证明,得到的基本再生数公式能够很好地描述文件分发过程的行为和反映不同参数对文件传播带来的不同影响。

关键词 P2P文件共享系统,文件分发,传播,基本再生数,影响因素
中图法分类号 TP393 **文献标识码** A

Research on Process of File Diffusion and Influence Factors of Propagation in P2P Networks

CHEN Bao-gang[1,2] XU Yong[2] HU Jin-long[2]
(College of Information and Management Science, Henna Agriculture University, Zhengzhou 450002, China)[1]
(Communication and Computer Network Laboratory of Guangdong Province, South China University of Technology, Guangzhou 510641, China)[2]

Abstract In P2P networks, the behaviors of popular file downloaded are similar to the spread process of the epidemic diseases, which can be described by dynamics of infectious diseases spread. The diffusion process of P2P file was modeled in accordance with the epidemic dynamic model. And also the basic reproductive number was derived, which can determine the ability of file propagation. Further experiments are proved that the basic reproductive number obtained can describe the behavior of file propagation and reflect of the differential impact the different parameters on file diffusion fairly well.

Keywords P2P file sharing system, File diffusion, Propagation, Basic reproductive number, Influence factors

1 引言

在P2P文件共享系统中,一个文件会被多个用户下载。然而,影响一个文件下载分发的范围和广度的因素有很多,除了受用户主观意愿希望下载这个文件外,还受到用户的共享时间、在线时间等因素的影响。从已有的文献资料来看,文件分发过程中,对文件传播的影响因素的研究被忽视了。文献[1]只研究了在稳态情况下系统的平均性能,而没有分析不同因素对文件复本数量变化的影响。文献[2]分析了共享概率对用户的影响,而没有考虑其他方面的因素。文献[3,4]则根本没有考虑文件传播的效率。

为了分析各种因素在文件分发过程中的作用,需要了解文件的分发受到哪些因素的制约。在本文的研究中,我们使用传染病动力学中表征疾病能否流行的基本再生数来分析文件分发过程中相应的制约因素。首先通过引入传染病动力学模型建立了文件分发的方程系统,然后在此基础上推导出决定文件传播效率的基本再生数公式,并通过实验确定和验证了在此公式中影响文件传播效率的条件和因素。

2 文件分发过程模型

P2P文件共享系统中,不同的用户会下载同一个感兴趣的文件。并且,当一些P2P用户节点完成文件下载后还会共享和传播已下载的文件,使得拥有这个文件的用户群不断地扩大。因此,可以考虑使用传染病动力学模型来研究文件分发的模型。

不考虑文件分发系统中在文件下载尚未完成就可以上传文件这样的机制,只考虑基本的P2P文件共享系统中文件的下载过程。流行文件大量下载这样的过程极类似于传染病的传播过程,可以利用SEIR模型来描述,并把节点划分到4个不同的仓室:W、D、S和I仓室,以P_W、P_D、P_S、P_I分别表示各个仓室中的节点数量。可以认为,用户节点在提出下载请求前在文件共享系统中搜索和查找文件时的状态为易感状态,并记这些节点为需要下载节点W;当这些用户节点提出下载请求后,会进入下载队列,等待下载完成,这些用户节点被认为处于潜伏期状态,我们称这些节点为下载点D;当这些节点完成下载后,可能会共享这些下载的文件,以供别的节点完成下载而自身处于感染态,称其为共享点S;当用户对该文

到稿日期:2011-08-11 返修日期:2011-11-15 本文受国家973计划项目(2003CB314805,2009CB320505),河南省科技攻关计划项目(112102210197)资助。
陈宝钢(1973—),男,博士,讲师,主要研究方向为P2P网络、网络测量、无线传感器网络,E-mail:bgchen@scut.edu.cn。

件并不感兴趣或者用户节点下载完文件后并不共享该文件及共享该文件一段时间后删除该文件的状态称为恢复态，记这些节点为空转点 I。

用户节点处于在线状态时，可以经历上述状态转换。而用户节点位于这些状态时，也可以从在线状态进入离线状态。同样，也可以从离线状态返回到在线状态。

所以，每个仓室由当前在线的节点和当前离线的节点组成。例如 S_{on}、S_{off} 节点分别指的是在线共享文件以及离线共享文件的节点，P_{Son}、P_{Soff} 分别指节点数量值。因此整个系统中所有仓室的节点数量就是一个常数 N_P 且有 $N_P = P_{Won} + P_{Woff} + P_{Don} + P_{Doff} + P_{Son} + P_{Soff} + P_{Ion} + P_{Ioff}$。

根据图 1 所示各类节点的状态变化，我们通过分析各类用户节点数量变化来获得文件下载分发行为的方程。

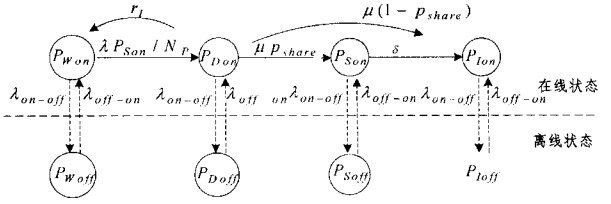


图 1　P2P 文件共享系统中文件分发过程节点状态变化图

以在仓室 W 中的节点为例。由于流行文件是共享者和下载者都很多的文件，因此对某一文件进行查询的速率和当前该文件在系统中在线共享节点的数量与总的用户节点数量的比值应该成正比，即与 P_{Son}/N_P 成正比。表 1 为 P2P 共享系统文件分发模型的参数定义。

表 1　P2P 文件共享系统文件分发模型的参数定义

$1/\lambda_{on\text{-}off}$	节点在线的平均持续时间
$1/\lambda_{off\text{-}on}$	节点离线的平均持续时间
$1/\lambda$	在线点提出下载文件请求的平均时间间隔
$1/\mu$	共享文件的平均下载完成时间，包括排队时间和下载时间
r_1	节点终止正在进行的下载文件任务的速率
$1/\delta$	一个节点共享一个文件的平均时间
p_{share}	下载完文件后准备共享该文件供其它节点来下载的概率

假设用户发出文件查询和下载请求的平均速率是 λ，根据传染病标准发生率 $\beta SI/N$ 的形式[5]，用户发出的针对该文件的查询和下载请求速率即是 $\lambda k P_{Son}/N_P$，k 是比例系数。假设 $k=1$，那么用户节点会以 $\lambda P_{Won} P_{Son}/N_P$ 的速度离开该类。同时当节点离线时，在线节点的数量会减少，规定离线发生的速率是 $\lambda_{on\text{-}off}$，因此从 W_{on} 类转到其他类的速率为 $\lambda P_{Won} P_{Son}/N_P + \lambda_{on\text{-}off} P_{Won}$。

下载节点在下载过程中可能由于共享该文件节点的离开而失去下载源，因此那些正在下载的点会继续寻求新的下载源，这些点会由 D_{on} 类再进入 W_{on} 类，假设这个转移发生的速率是 r_1。当节点从离线状态转换到在线状态时，W_{on} 类节点的数量也会增加，这种转换发生的速率为 $\lambda_{off\text{-}on}$。单位时间内进入到 P_{Won} 类的速率为 $r_1 P_{Don} + \lambda_{off\text{-}on} P_{Woff}$。

那么，W_{on} 类节点成员变化速率就表示为：

$$\frac{dP_{Won}}{dt} = -\lambda P_{Won} P_{Son}/N_P - \lambda_{on\text{-}off} P_{Won} + r_1 P_{Don} + \lambda_{off\text{-}on} P_{Woff}$$

令各类节点离线状态和在线状态转化的速率是一致的，那么所有节点都以 $\lambda_{off\text{-}on}$ 的速率从离线进入在线，而以 $\lambda_{on\text{-}off}$ 的速率从在线进入离线。

所以 W_{off} 类节点数量变化方程为：

$$\frac{dP_{Woff}}{dt} = \lambda_{on\text{-}off} P_{Won} - \lambda_{off\text{-}on} P_{Woff}$$

依此类推，本文的文件分发模型可以用下面的方程系统来确定：

$$\frac{dP_{Won}}{dt} = -\lambda P_{Won} P_{Son}/N_P - \lambda_{on\text{-}off} P_{Won} + r_1 P_{Don} + \lambda_{off\text{-}on} P_{Woff} \quad (1)$$

$$\frac{dP_{Don}}{dt} = \lambda P_{Won} P_{Son}/N_P - \mu P_{Don} - r_1 P_{Don} - \lambda_{on\text{-}off} P_{Don} + \lambda_{off\text{-}on} P_{Doff} \quad (2)$$

$$\frac{dP_{Son}}{dt} = \mu p_{share} P_{Don} - \delta P_{Son} - \lambda_{on\text{-}off} P_{Son} + \lambda_{off\text{-}on} P_{Soff} \quad (3)$$

$$\frac{dP_{Ion}}{dt} = \delta P_{Son} + \mu(1-p_{share}) P_{Don} - \lambda_{on\text{-}off} P_{Ion} + \lambda_{off\text{-}on} P_{Ioff} \quad (4)$$

$$\frac{dP_{Woff}}{dt} = \lambda_{on\text{-}off} P_{Won} - \lambda_{off\text{-}on} P_{Woff} \quad (5)$$

$$\frac{dP_{Doff}}{dt} = \lambda_{on\text{-}off} P_{Don} - \lambda_{off\text{-}on} P_{Doff} \quad (6)$$

$$\frac{dP_{Soff}}{dt} = \lambda_{on\text{-}off} P_{Son} - \lambda_{off\text{-}on} P_{Soff} \quad (7)$$

$$\frac{dP_{Ioff}}{dt} = \lambda_{on\text{-}off} P_{Ion} - \lambda_{off\text{-}on} P_{Ioff} \quad (8)$$

3　文件传播的基本再生数

在人口统计学和生态学中，R_0 意味着个体在整个生命期中成功再生的物种个数。在流行病学中，R_0 的意思是单个感染个体在它的整个感染阶段所能够感染的易感个体的数量[6-8]。可以看到，当 $R_0<1$ 时，每个感染个体平均产生少于一个新感染的个体，因而传染病会很快在人群中消失；当 $R_0>1$ 时，每个感染个体平均至少产生一例感染，那么在疾病爆发初期，感染者数量会逐步增多，疾病会在易感人群中流行。

根据文献[9]定义的方法分别找出各仓室增加的新感染个体的速率 $\mathcal{F}$ 以及各仓室中不属于增加新感染个体的个体数量离去的变化率 $\mathcal{V}$。按照 P_{Don}，P_{Doff}，P_{Son}，P_{Soff}，P_{Won}，P_{Woff}，P_{Ion}，P_{Ioff} 的排列顺序求出的 $\mathcal{F}$ 和 $\mathcal{V}$ 向量的值分别是：

$$\mathcal{F} = \begin{pmatrix} \lambda P_{Won} P_{Son}/N_P \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad (9)$$

$$\mathcal{V} = \begin{pmatrix} r_1 P_{Don} + \mu P_{Don} + \lambda_{on\text{-}off} P_{Don} - \lambda_{off\text{-}on} P_{Doff} \\ -\lambda_{on\text{-}off} P_{Don} + \lambda_{off\text{-}on} P_{Doff} \\ -\mu p_{share} P_{Don} - \delta P_{Son} + \lambda_{on\text{-}off} P_{Son} - \lambda_{off\text{-}on} P_{Soff} \\ -\lambda_{on\text{-}off} P_{Son} + \lambda_{off\text{-}on} P_{Soff} \\ \lambda P_{Won} S_{on}/N_P - r_1 P_{Don} + \lambda_{on\text{-}off} P_{Won} - \lambda_{off\text{-}on} P_{Woff} \\ -\lambda_{on\text{-}off} P_{Won} + \lambda_{off\text{-}on} P_{Woff} \\ -\delta P_{Son} - \mu(1-p_{share}) P_{Don} + \lambda_{on\text{-}off} P_{Ion} - \lambda_{off\text{-}on} P_{Ioff} \\ -\lambda_{on\text{-}off} P_{Ion} + \lambda_{off\text{-}on} P_{Ioff} \end{pmatrix} \quad (10)$$

根据在传染病动力学中当模型处于无病平衡点时导数为 0 且各个仓室不存在染病的个体[9]，有以下条件存在：

$$\frac{dP_{Won}}{dt}=\frac{dP_{Woff}}{dt}=\frac{dP_{Don}}{dt}=\frac{dP_{Doff}}{dt}=\frac{dP_{Son}}{dt}=\frac{dP_{Soff}}{dt}=\frac{dP_{Ion}}{dt}=\frac{dP_{Ioff}}{dt}=0$$

和 $P_{Don}=P_{Doff}=P_{Son}=P_{Soff}=0$。因此，在平衡点有 $N_P=P_{Won}+P_{Woff}$。而在平衡点 x_0 可以表示为 $x_0=\{\hat{P}_{Won},\hat{P}_{Woff},0,0,0,0,0,0\}$，并且 $\hat{P}_{Won}$ 和 $\hat{P}_{Woff}$ 可以分别表示为 $\hat{P}_{Won}=\frac{\lambda_{off\text{-}on}}{\lambda_{on\text{-}off}+\lambda_{off\text{-}on}}N_P$ 和 $\hat{P}_{Woff}=\frac{\lambda_{on\text{-}off}}{\lambda_{on\text{-}off}+\lambda_{off\text{-}on}}N_P$。

为求出文件传播基本再生数，按照文献[9]，需求出向量 $\mathcal{F}$ 和 $\mathcal{V}$ 在无病平衡点的导数，也即求出在无病平衡点时前 m 个感染仓室的导数值 F 和 V。按照文献[10]中的方法来求 F 和 V。因为我们只有用户节点一个类型，所以“物种数量”s 应等于 1，而在线状态和离线状态使得“空间区域数量”k 设置为 2，所以 F 和 V 的值表示如下：

$$F=\left[\begin{array}{c|c}0 & G\\ \hline 0 & 0\end{array}\right]\quad V=\left[\begin{array}{c|c}A & 0\\ \hline -C & B\end{array}\right]$$

这里每个分块矩阵都是 2×2 的矩阵。而且有

$$G=\begin{bmatrix}\lambda P_{Won}/N_P & 0\\ 0 & 0\end{bmatrix},A=\begin{bmatrix}r_1+\mu+\lambda_{on\text{-}off} & 0\\ 0 & \lambda_{off\text{-}on}\end{bmatrix}-\widetilde{M}$$

$$B=\begin{bmatrix}\delta+\lambda_{on\text{-}off} & 0\\ 0 & \lambda_{off\text{-}on}\end{bmatrix}-\widetilde{M},C=\begin{bmatrix}\mu P_{share} & 0\\ 0 & 0\end{bmatrix}$$

$$\widetilde{M}=\begin{bmatrix}0 & \lambda_{off\text{-}on}\\ \lambda_{on\text{-}off} & 0\end{bmatrix}$$

我们假设，当用户处于下载状态和共享状态时并不离线，也即使得在仓室 P_{Don}、P_{Doff}、P_{Son} 和 P_{Soff} 时的 $\lambda_{on\text{-}off}=0$ 而 $\lambda_{off\text{-}on}=1$。那么矩阵：

$$GB^{-1}CA^{-1}=\begin{bmatrix}\lambda P_{Won}/N_P & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\frac{1}{\delta} & \frac{1}{\delta}\\ 0 & 1\end{bmatrix}\begin{bmatrix}\mu P_{share} & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\frac{1}{r_1+\mu} & \frac{1}{r_1+\mu}\\ 0 & 1\end{bmatrix}$$

$$=\begin{bmatrix}\frac{\lambda\mu P_{share}P_{Won}}{\delta(r_1+\mu)N_P} & \frac{\lambda\mu P_{share}P_{Won}}{\delta(r_1+\mu)N_P}\\ 0 & 0\end{bmatrix}$$

矩阵 $GB^{-1}CA^{-1}$ 的谱半径为：

$$R_0=\frac{\lambda\mu P_{share}P_{Won}}{\delta(r_1+\mu)N_P}\tag{11}$$

又由于在平衡点状态时只有 P_{Won} 和 P_{Woff} 不为 0，根据前面对 $\hat{P}_{Won}$ 和 $\hat{P}_{Woff}$ 的定义，因此有：

$$R_0=\frac{\lambda\mu P_{share}\lambda_{off\text{-}on}}{\delta(r_1+\mu)(\lambda_{on\text{-}off}+\lambda_{off\text{-}on})}\tag{12}$$

从公式中可以看到，描述文件传播效率的参数 R_0 和用户提出请求下载的速率、文件的下载完成速率、用户对文件的共享程度、用户从离线到在线的转换速率成正比，同用户不再共享文件的速率、用户终止正在进行的下载文件任务的速率与文件的下载完成速率之和以及用户从离线到在线的转换速率和用户从在线到离线的转换速率的和成反比。

4 文件传播影响因素分析

设 P2P 文件共享系统中对某一种文件感兴趣的用户群数量为 $N_P=20050$。这些用户对某一文件都有下载的需求，而其中有 50 个用户拥有并共享该文件。设该文件的长度为 700MB，即普通的 AVI 视频格式文件的大小。设置模拟时间的粒度为 min，文件分发时间为 21600min，即 15 天。并假设用户每周对单个文件提出的文件下载请求个数是 70 个，平均下载请求速率为每分钟 0.007。

4.1 文件下载完成速率

设平均下载完成速率从 200kB/s 到 2MB/s 发生变化，设置其他参数如表 2 所列。

表 2 文件下载完成速率模拟中使用的参数值

λ	δ	r_1	P_{share}	λ_{on_off}	λ_{off_on}
0.007	6.94e−4	6.94e−4	0.3	0.0042	0.0042

结果如表 3 所列。在不同的速率下，共享节点数量随时间变化的过程是很相近的。经过计算看出，在这些情况下的基本再生数都大于 1，并且数值相差不大。

表 3 共享时间为 1440min 时文件下载完成速率影响的模拟效果

下载完成速率	0.017	0.034	0.051	0.068	0.085	0.102	0.119	0.136	0.153	0.17
R_0	1.453	1.482	1.492	1.497	1.501	1.502	1.504	1.505	1.506	1.507
完成下载节点数	10704	11212	11374	11453	11501	11532	11555	11571	11584	11595

4.2 文件下载后共享时间

设 δ 从 0.00556（平均共享时间为 3h）变化到 5.56e−4（平均共享时间为 30h）。r_1 的变化和 δ 一致，平均下载完成速率设置为 500kB/s。其他参数如表 4 所列。

表 4 共享时间影响模拟中使用的参数值

λ	μ	P_{share}	λ_{on_off}	λ_{off_on}
0.007	0.0428	0.3	0.0042	0.0042

可以从表 5 看到，在不同的共享时间取值下，基本再生数的取值及用户节点的下载完成数量相差较大。这表明共享时间对用户节点下载完成数的影响很大。

表 5 文件下载后共享时间影响的模拟结果

离开速率	5.56e−3	2.78e−3	1.85e−3	1.39e−3	1.11e−3	9.25e−4	7.93e−4	6.94e−4	6.17e−4	5.56e−4
R_0	0.167	0.354	0.544	0.731	0.922	1.111	1.299	1.488	1.677	1.864
完成下载节点数	83	140	245	481	1318	4174	8205	11312	13461	14968

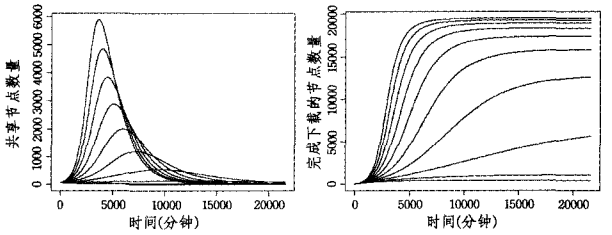
4.3 文件下载后共享概率

假设 P_{share} 在 0.08 和 0.8 之间取值，其他参数的取值情况设置如表 6 所列。

表 6 共享概率影响模拟中使用的参数值

λ	μ	δ	r_1	$\lambda_{on\text{-}off}$	$\lambda_{off\text{-}on}$
0.007	0.0428	5.94e−4	5.94e−4	0.0042	0.0042

当 P_{share} 在 0.08 和 0.8 之间取值时，用户节点的下载完成数量的范围在 452 和 19481 之间变化。共享节点数量变化和完成下载的节点数量变化如图 2 所示。而基本再生数的取值也在较大范围内变化。这表明用户的共享程度在文件传播过程中起到了重要作用，决定了文件传播的速度和范围。文件下载后共享概率影响模拟结果如表 7 所列。



(a) 共享节点数量变化　　(b) 完成下载的节点数量变化

图 2　共享概率影响下共享节点和完成下载节点的数量变化

表 7　文件下载后共享概率影响的模拟结果

共享概率	0.08	0.16	0.24	0.32	0.40	0.48	0.56	0.64	0.72	0.80
R_0	0.397	0.794	1.191	1.588	1.985	2.382	2.779	3.176	3.573	3.97
完成下载节点数	452	1072	5652	12594	15760	17384	18315	18883	19244	19481

4.4　用户在线时间和离线时间

设在线时间和离线时间的取值在 1h 和 10h 之间变化，分析在线时间和离线时间的变化给用户下载效率带来的影响。其他参数如表 8 所列。

表 8　在线时间和离线时间影响使用的参数值

λ	P_{share}	μ	δ	r_1
0.007	0.3	0.0428	6.94e−4	6.94e−4

(a) $\lambda_{on\text{-}off}=\lambda_{off\text{-}on}$

在线时间和离线时间相等时，基本再生数是相等的。但是下载完成的节点数会有一些小的区别。因为基本再生数是相等的，但是共享文件的用户节点在一定时间内会退出共享，而转换间隔短的用户节点会有更多的机会得到下载源，所以完成下载的用户节点数会多于在线时间和离线时间间隔较长的情况。结果如表 9 所列。

(b) $\lambda_{on\text{-}off}\geqslant\lambda_{off\text{-}on}$

设用户节点处于离线状态的平均时间是 10h，而处于在线状态的平均时间在 1h 和 10h 之间变化。在表 10 中可以看到，当离线时间和在线时间相差较大时，很少有用户在模拟的时间内得到下载。而且在经过一段时间之后，系统中提供共享文件的用户节点数量也在减少，从而影响了完成下载的用户节点数。

(c) $\lambda_{on\text{-}off}\leqslant\lambda_{off\text{-}on}$

当在线时间多于离线时间时，用户从离线状态转换到在线状态的速率就会高于从在线状态到离线状态的速率。设节点处于在线状态的平均时间是 10h，处于离线状态的平均时间在 1h 和 10h 之间变化。从表 11 中可以看到，当离线时间越短时，完成下载的用户数量也越多；当离线时间较长时，完成下载的用户数量随之减少。

表 9　在线时间和离线时间相等条件影响下的模拟结果

$\lambda_{on\text{-}off}$	0.017	0.0083	0.0056	0.0042	0.0033	0.0028	0.0024	0.0021	0.00186	0.0017
$\lambda_{off\text{-}on}$	0.017	0.0083	0.0056	0.0042	0.0033	0.0028	0.0024	0.0021	0.00186	0.0017
R_0	1.488	1.488	1.488	1.488	1.488	1.488	1.488	1.488	1.488	1.488
完成下载节点数	11409	11375	11344	11312	11277	11249	11218	11187	11155	11129

表 10　在线时间不大于离线时间条件影响下的模拟结果

$\lambda_{on\text{-}off}$	0.017	0.0083	0.0056	0.0042	0.0033	0.0028	0.0024	0.0021	0.00186	0.0017
$\lambda_{off\text{-}on}$	0.0017	0.0017	0.0017	0.0017	0.0017	0.0017	0.0017	0.0017	0.0017	0.0017
R_0	0.271	0.506	0.693	0.858	1.012	1.124	1.234	1.332	1.422	1.488
完成下载节点数	111	218	408	839	1973	3726	6136	8323	10053	11129

表 11　在线时间不小于离线时间条件影响下的模拟结果

$\lambda_{on\text{-}off}$	0.0017	0.0017	0.0017	0.0017	0.0017	0.0017	0.0017	0.0017	0.0017	0.0017
$\lambda_{off\text{-}on}$	0.017	0.0083	0.0056	0.0042	0.0033	0.0028	0.0024	0.0021	0.00186	0.0017
R_0	2.706	2.471	2.284	2.119	1.965	1.852	1.743	1.645	1.555	1.488
完成下载节点数	18400	17833	17224	16531	15703	14958	14079	13131	12070	11129

结束语　P2P 文件共享系统中，流行文件的下载行为类似于传染病在人群中的传播过程。基本再生数在传染病动力学研究中用于表示已感染个体能够感染新个体的能力。本文利用传染病动力学中基本再生数的概念以及提出的一种 P2P 文件分发模型，推导了描述影响文件传播效率的基本再生数公式。通过实验分析发现，提高用户节点的下载完成速率、用户节点完成下载后的共享概率、用户节点的共享时间、用户节点的在线时间都会对文件下载的完成数量产生积极的影响。但在其他条件不变的情况下，用户节点下载完成速率的提高并没有带来相应的下载完成用户节点数量的明显提高。而增加用户节点的共享概率、延长用户节点的共享时间和用户节点的在线时间都可以使下载完成的效率获得明显的提高，特别是提高用户的共享概率所起的作用最为明显。

参 考 文 献

[1] Qiu D Y, Srikant R. Modeling and performance analysis of BitTorrent-like peer-to-peer networks [A]// Proceedings of the Conference on Computer Communications [C]. New York: Association for Computing Machinery, 2004: 367-377

[2] Leibnitz K, Hossfeld T, Wakamiya N. et al. Modeling of epidemic diffusion in peer-to-peer file-sharing networks [A]// Proceedings of the 2nd International Workshop on Biologically Inspired Approaches for Advanced Informatin Technology [C]. Berlin: Springer-Verlag, 2006: 322-329

[3] 杨巍，常桂然，朱志良，等. P2P 网络中文件分发过程的动力学模型研究[EB/OL]. http://www.paper.edu.cn/paper.ph p? serial_number=200611-66, 2006

[4] Lo P F, Giovanni N, Giuseppe B. The effect of heterogeneous link capacities in BitTorrent-like file sharing systems [A]// Proceedings of the First International Workshop on Hot Topics in Peer-to-Peer Systems [C]. New York: Institute of Electrical and Electronics Engineers Inc., 2004: 40-47

(下转第 134 页)

发送到指定端口，从而最终完成基于 Fuzz 数据的故障注入。

4.3 故障注入模块设计

故障注入模块作为整个系统的核心部分，通过分析协议交互内容来完成分布式软件的交互测试。其主要功能是向目标系统中注入故障，要求其能保持注入故障的类型和位置等信息，并包含适当的软件或硬件逻辑，以确保故障在正确的时刻注入到正确的位置。

故障注入模块基于协议扁平化思想，将协议解析为字符长串，自底向上从不同层次支持包括 TCP/IP、HTTP、SOAP 及私有协议在内的各类通信协议，通过网络客户与服务端的通信监听，实现客户端通信报文的动态捕获。故障注入类解析通信过程中所捕获的消息数据包，获取消息格式，结合协议规范中所描述的消息单元语法，构造完整的消息单元。对于所构造的消息单元，依据语法和语义信息选取特定字段作为故障注入点，结合 Fuzz 测试数据生成的故障数据消息，并通过绑定服务端通信进程，利用故障消息模拟客户与服务器通信。

同时，故障注入模块还负责 Fuzz 测试数据的生成。它根据目标程序中可能存在的特定故障而产生数据，这些故障数据很有可能引发软件故障。故障注入模块使用 Fuzz 测试验证应用的可靠性，其基本思想是故意将随机的坏数据插入应用程序，然后等待并观察哪里遭到了破坏。Fuzz 测试的技巧在于，它是不符合逻辑的：不去猜测哪个数据会导致破坏，而是将尽可能多的杂乱数据投入程序中。

Fuzz 测试的输入数据主要强调的是数据的随机性，包括消息传输过程的通信故障：延时、丢包、乱序；接口调用过程的数据包语法故障和 API 参数故障，可概括为：在命令行模式下随机输入的随机 ASCII 字符流（如非法数据、乱序数据、上下文不一致数据、非法分隔符、不一致数据、空数据等），以及在视窗模式下随机输入的有效键盘和鼠标输入序列。这些输入完全不考虑系统逻辑，从某种程度上说是一种破坏性的测试。具体实现界面如图 3 所示。

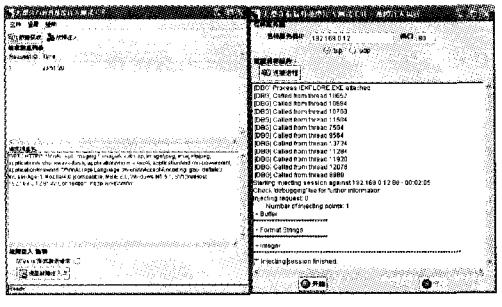


图 3 故障注入模块解界面图

4.4 XML 数据处理模块设计

XML 数据处理类主要针对 XML 文件的处理，提供了诸如读写 XML 元素、生成 XML 文档对象模式（DOM）、处理 XSL 文件等功能。

4.5 数据收集与分析模块设计

数据收集与分析包负责跟踪工作的任务，并在适当的时刻采集数据，同时对采集到的数据进行分析和处理。数据分析工作采用离线分析模式，通过分析计算可以判定目标系统中是否存在特定故障，为进一步进行故障诊断、隔离和获取提供数据基础。数据收集与分析包还通过日志记录器将工具运行期间的信息保存至本地硬盘。同时，日志记录器还保存所有接收和发送的数据包。

结束语 可靠性是衡量大型分布式软件质量的关键属性。但长期以来，在分布式软件可靠性测评领域缺乏易操作、高效的测试方法。本文分析了分布式软件系统中的典型通信协议故障，研究了基于 API Hook 的分布式软件可靠性测试方法，其包括分布式软件故障模型构建、故障数据构造等；提出了基于通信协议故障注入的应用软件可靠性测试方法，并实现工具原型，为基于故障注入的分布式软件可靠性测试提供了技术手段。

参 考 文 献

[1] Andrews J H, Brand L C, Labiche Y. Is mutation an appropriate tool for testing experiments[C]// Proceedings of the 27th International Conference on Software Engineering (ICSE'2005). St. Louis, MO, USA, 2005: 402-411

[2] Offutt A J, Untch R. Uniting the orthogonal[C]// Proceedings of the Mutation 2000: Mutation Testing in the Twentieth and the Twenty First Centuries. San Jose, CA, USA, 2000: 45-55

[3] Delamaro M E, Maidonado J C, Mathur A P. Interface mutation: An approach for integration testing[J]. IEEE Transactions on Software Engineering, 2001, 27(3): 228-247

[4] Ma Y-S, Kwon Y-R, Offutt J. Inter-class mutation operators for Java[C]// Proceedings of the 13th International Symposium on Software Reliability Engineering (ISSRE'2002). Annapolis, MA, USA, 2002: 352-363

[5] Lee H-J, Ma Y-S, Kwon Y-R. Empirical evaluation of orthogonally of class mutation operators[C]// Proceedings of the 11th Asia-Pacific Software Engineering Conference. Busan, Korea, 2004: 512-518

[6] Aichernig B K. Mutation testing in the refinement calculus[J]. Formal Aspects of Computing, 2003, 15(2/3): 280-295

[7] Jiang Y, Hou S-S, Shan J-H, et al. Contract-based mutation for testing components[C]// Proceedings of the 21st International Conference on Software Maintenance (ICSM 2005). Budapest, Hungary, 2005: 483-492

(上接第 103 页)

[5] 马知恩. 传染病动力学的建模和研究[M]. 北京：科学出版社，2004

[6] Heffernan J M, Smith R J, Wahl L M. Perspectives on the basic reproductive ratio [J]. Journal of the Royal Society Interface, 2005, 2(4): 281-293

[7] Anderson R M, May R M. Infectious Diseases of Humans [M]. Oxford: Oxford University Press, 1991

[8] Dickmann O, Heesterbeck J A P. Mathematical epidemiology of infectious diseases [M]. New York: John Wiley & Sons, 2000

[9] Van den Driessche P, Watmough J. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission [J]. Math Biosci, 2002, 180: 29-48

[10] Baroyan O V, Rvachev L A, Basilevsky U V, et al. Computer modeling of influenza epidemics for the whole country (USSR) [J]. Advances in Applied Probability, 1971, 3: 224-226