INFOSCI *INFORMATION SCIENCE PUBLISHING* IGR5197
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.igi-global.com

This paper appears in the publication,
Handbook of Research on Global Information Technology Management in the Digital Economy
edited by Mahesh S. Raisinghani © 2008, IGI Global

Chapter XXI
Sampling Approaches on Collecting Internet Statistics in the Digital Economy

Song Xing
California State University, Los Angeles, USA

Bernd-Peter Paris
George Mason University, USA

Xiannong Meng
Bucknell University, USA

ABSTRACT

The Internet's complexity restricts analysis or simulation to assess its parameters. Instead, actual measurements provide a reality check. Many statistical measurements of the Internet estimate rare event probabilities. Collection of such statistics renders sampling methods as a primary substitute. Within the context of this inquiry, we have presented the conventional Monte Carlo approach to estimate the Internet event probability. As a variance reduction technique, Importance Sampling is introduced which is a modified Monte Carlo approach resulting in a significant reduction of effort to obtain an accurate estimate. This method works particularly well when estimating the probability of rare events. It has great appeal to use as an efficient sampling scheme for estimating the information server density on the Internet. In this chapter, we have proposed the Importance Sampling approaches to track the prevalence and growth of Web service, where an improved Importance Sampling scheme is introduced. We present a thorough analysis of the sampling approaches. Based on the periodic measurement of the number of active Web servers conducted over the past five years, an exponential growth of the Web is observed and modeled. Also discussed in this chapter is the increasing security concerns on Web servers.

INTRODUCTION

The Internet has grown tremendously from an early research prototype in 1969 connecting four computers to today's global communication system reaching all countries of the world. Businesses, educational institutions, government organizations, and individuals have become heavily dependent on its capability for rapid data communications and information exchange. One issue related to

this continuous growth is the evident increase in numbers of hosts connected to the Internet, and also worth noting are the numerous public IP addresses being consumed by these computers.

Theoretically, the current IPv4 address space can identify 2^{32}, or 4.3 billion hosts. However, the two-level address structure (consisting of a network and a host) categorized into five classes imposes constraints that make the use of the address space inefficient. Subnetting and supernetting (or CIDR, Classless Inter-Domain Routing) approaches allow more efficient allocation of IP addresses than classful addressing, but these strategies make routing more complicated. Actually, since the release of IPv4, the Internet population grew to over 400 million hosts by the end of 2006 (ISC–Survey, n.d.), increasing far faster than anticipated. As the space of available addresses decreases, it becomes increasingly difficult to obtain new public IPv4 addresses. Furthermore, the pace of this growth is expected to continue for years to come.

In the short term, Dynamic Host Configuration (DHCP) and Network Address Translation (NAT) relieve the pressure for additional address space. By using private addresses that are reserved for local usage, NAT allows network administrators to hide large communities of users behind firewalls and NAT boxes. Since different multi-corporate networks can each reuse the same local private addresses, NAT reduces the need for new unique public IP addresses. Unfortunately, NAT is not a permanent solution. It addresses the needs of large communities of client systems, but it does not help the servers on the Internet as each requires a unique public address. Nor does it work for peer-to-peer communications for the same reason. What happens when we run out of public IPv4 address?

To improve IPv4's scalability, as well as its security, ease-of-configuration, and network management, the next-generation Internet Protocol (IPv6) has been proposed and is now a standard. As a long-term solution, IPv6 fixed the problem of the shortage of IPv4 addresses by increasing the IP address size from 32 bits to 128 bits. And IPv6 is expected to gradually replace IPv4 with the two addresses coexisting during a transition period. Therefore, it would be more realistic and helpful to predict when IPv4 address will eventually run out, and when IPv6 will need to be widely implemented. Being able to map the growth of the Internet or take snapshots of its current size is certainly beneficial in planning the future evolution of IPv4 and the implementation of IPv6.

In addition, recent studies have stated that a historical analysis shows the phenomenal growth of Internet usage was slowed in recent times (Devezas, Linstone, & Santos, 2005; Modis, 2005). Specifically, Devezas et al. (2005) report the growth of Internet users is coming to the end of the fourth Kondratieff cycles downswing and will then embark on the fifth Kondratieff cycles upswing. Modis (2005) points out that the population trends and Internet-user trends have indicated that the percentage of the population using the Internet is decreasing everywhere despite large discrepancies in different regions in the world, and the boom years of Internet explosion are over. Similarly, Nielsen's (2006) report states that the early Web's explosive growth rate has slowed and the Web has experienced a "maturing" growth in the last five years. Whether the decreased Internet usage or the Web maturation, both arguments provide us some useful insights. On the one hand, it shows the Internet or the Web is no longer a marvel of innovation. On the other hand, it may implicate a nearly exhausted IPv4 address space. Hence having a reliable and accurate estimate of the present size of the Internet and of its growth rate would be very important and of interest both to network operators/engineers and to market analysts.

Be aware that the Internet is a decentralized and dynamic compilation of global networks. To evaluate the parameters or performance of such a complex system, analytical techniques may be applied. For example, Balchi and Mukhopadhyay (2004) have introduced several soft models—such as genetic algorithm, neural network, and fuzzy regression—to study and predict the Internet growth in several OECD[1] nations. However, the analytical techniques are usually very expensive, time consuming, and relatively inflexible. In addition, such techniques often require over-simplification of the system model, leading to uncertain and inaccurate estimates. Simulation is another powerful technology that plays a key role in exploring the scenarios that are difficult or impossible to analyze. However, it is difficult to generate simulation scenarios to map