利用熵检测 DNS 异常①

丁森林 吴 军 毛 伟 (中国科学院 计算机网络信息中心 北京 100190)

摘 要: 利用 DNS 查询数据中出现的查询类型作为随机条件计算熵值，根据熵值的突发性变化，检测 DNS 查询中是否出现异常。利用该算法对 2009 年 5 月 19 日发生大面积断网事件时的 DNS 查询原始数据进行分析，证明这种方法可以及时检测到 DNS 查询中的异常情况的发生，并具有较好的检测效果。

关键词: DNS；熵；DNS 异常；DNS 查询类型

Entropy Method for DNS Abnormal Detection

DING Sen-Lin, WU Jun, MAO Wei

(Computer Network Information Center, Chinese Academy of Sciences (CNIC), Beijing 100190, China)

Abstract: This paper propose a method of detecting DNS abnormalities by calculating the entropies of the DNS query types observed in consecutive windows of fixed-size. Applied to the DNS query data targeting .CN on May 19th 2009 when there was a major DNS accident happened, this method demonstrates ability of detecting the abnormal behavior towards DNS before the event was observed and reported.

Keywords: DNS; entropy; DNS abnormal; query type

1 引言

域名系统(Domain Name System，DNS)是整个互联网络的基础设施，其主要功能是实现域名地址和 IP 地址之间的转换[1,2]。DNS 系统的正常运行，是 Web 服务、电子邮件服务等众多网络服务正常运行的基础。

DNS 系统是目前全球最大最复杂的分布式层次数据库系统，由于其开放、庞大、复杂的特性以及设计之初对于安全性的考虑不足，使得 DNS 系统中存在着诸多的潜在错误和威胁，同时针对 DNS 系统的人为攻击和破坏也时有发生，这就对 DNS 系统的安全稳定运行提出了严峻挑战。对 DNS 查询流量进行主动监测，及时发现系统中突发的异常行为，对于保证 DNS 系统的正常运行，提升 DNS 服务质量，具有重要意义。

长期以来，由于缺乏有效的检测手段，使得对于 DNS 系统的攻击很难被及时发现和处理。本文将信息论中熵的理论运用到 DNS 异常检测实例中，通过监测并分析熵值的变化，及时发现 DNS 使用中的突发异常状况。通过对真实的包含异常的 DNS 查询数据进行分析，证实该方法确实可以检测到 DNS 查询数据中的异常，并且具有较好的检测准确率和实效性。

2 信息论中的熵

“熵”(entropy)是德国物理学家克劳修斯(Rudolf Clausius, 1822－1888)在 1850 年提出的一个术语，用来表示任何一种能量在空间中分布的均匀程度[3]。信息论的创始人香农在 1948 年将熵的概念引入到信息论中，在其著作《通信的数学理论》中提出了建立在概率统计模型上的信息度量，他把信息定义为“用来消除不确定性的东西”。熵在信息论中的定义如下[4,5]：

如果在一个系统 S 中存在一个事件集合 $E=\{E_1, E_2,\cdots, E_n\}$，每个事件的概率分布 $P=\{P_1, P_2,\cdots, P_n\}$，则每个事件本身的信息量可由公式(1)表示如下：

$$I_i = -\log_2 P_i \quad (1)$$

① 基金项目:国家发改委 2009CNGI 项目(CNGI-09-03-04)
收稿时间:2010-04-02;收到修改稿时间:2010-05-18

熵是整个系统 S 的平均信息量，其计算方法如公式(2)所示：

$$H_s = \sum_{i=1}^{n} p_i I_i = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad (2)$$

在信息论中，熵表示的是信息的不确定性，高信息度的信息熵是很低的，而低信息度的熵则很高。具体说来，凡是导致随机事件集合的肯定性，组织性，法则性或有序性等增加或减少的活动过程，都可以用信息熵的改变量这个统一的标尺来度量。

熵值表示了系统的稳定情况，熵值越小，表示系统越稳定，反之，当系统中出现不确定因素较多时，会引起熵值升高[5]。如果某个随机变量的取值与系统的异常情况具有相关性，那么系统异常时刻该随机变量的平均信息量就会与系统稳定时刻不同。如果某一时刻该异常情况大量出现，则系统的熵值会出现较大幅度的变化，这就使我们有可能通过熵值的变化情况，来检测到系统中异常现象的发生，而且这种强相关性使得检测具有较高的准确度。

2.1 利用熵检测 DNS 异常

将熵的理论运用到 DNS 系统的异常检测中来，就是通过测量 DNS 数据包的某些特定属性的统计特性(熵)，来判断系统中是否有异常状况产生。这里熵值提供了一种对 DNS 的查询数据属性如查询名字、查询类型、各种错误查询的分布等特性的描述。熵值越大，表示属性的分布越随机。相反，熵值越小，表示属性分布范围越小，某些属性值出现的概率高。在正常稳定运行的 DNS 系统中，如果把查询数据作为信息流，以每条 DNS 查询请求中的某种查询类型的出现作为随机事件，那么在一段时间之内，查询类型这个随机变量的熵应该是一个比较稳定的值，当 DNS 查询数据中出现异常状况时，它的分布一定会发生变化，导致计算的熵值变化。

例如当利用 DNS 查询发起的分布式拒绝服务攻击(Distributed Denial of Service，DDoS)发生时，网络中会出现大量的攻击数据包，势必引起与查询类型、查询源地址等相关的统计特性发生变化[6]，另一方面，如果某些 DNS 服务器处于非正常状态，也会导致与查询类型、查询错误分布相关的统计特性发生变化。即便是黑客在发动攻击时，对于发送的查询请求的类型和数量进行过精心设计，可以使从攻击者到目标服务器之间某一路径上的熵值维持在稳定的水平，但绝不可能在所有的路径上都做到这一点。因此通过检测熵值的变化情况来检测 DNS 系统中异常状况的发生，不仅是一种简便可行的方案，而且还可以具有很好的检测效果。

在本文中，我们以查询类型为例研究熵值变化与系统异常的关系。DNS 系统是通过资源记录(Resource Record，RR)来记录域名和 IP 地址信息的，每个资源记录都有一个记录类型(QType)，用来标识资源记录所包含的信息种类，如 A 记录表示该资源记录是域名到 IP 地址的映射，PTR 记录 IP 地址到域名的映射，NS 记录表示域名的授权信息等，用户查询 DNS 相关信息时，需要指定相应的查询类型。

2.2 检测算法综述

按照 1.1 节所描述的思想，我们采用 DNS 查询数据中查询类型的出现情况作为随机事件来计算熵的变化情况，算法的具体实现如下：

1) 设定一个查询量窗口，大小为 W，表示窗口覆盖了 W 条记录。

2) 统计窗口中出现的所有查询类型及其在所属窗口中出现的概率，根据公式(2)计算出该窗口的熵 H_1。

3) 获取当前窗口中第一条查询记录所属的查询类型出现的概率 P_f，求出该类型所对应的增量 $T_f = -P_f \log_2 P_f$。

4) 将窗口向后滑动一条记录，此时新窗口中的第一条记录为窗口滑动前的第二条记录。

5) 获得窗口移动过程中加入的最后一条记录所代表的查询类型在原窗口中出现的概率 P_l 以及对应的增量 $T_l = -P_l \log_2 P_l$。

6) 计算新窗口中第一条记录所对应的查询类型出现在新窗口中出现的概率 $P_f^{'}$，以及对应的增量 $T_f^{'} = -P_f^{'} * \log_2 P_f^{'}$

7) 计算新窗口中最后一条记录所属的查询类型

在当前窗口出现的概率 $P_l^{'}$ 以及对应的增量 $T_l^{'}=-P_l^{'}*\log_2 P_l^{'}$。

8) 根据前面的结果计算窗口移动后的熵：

$$H_2=H_1-T_f-T_l+T_f^{'}+T_l^{'}$$

9) 重复步骤 2 至步骤 8 的过程，得到一系列的熵值，观察熵值的变化曲线，当熵值曲线出现剧烈波动时，可以断定此时的 DNS 查询中出现了异常。

窗口的设定是影响检测算法的一个重要因素，窗口越大，熵值的变化越平缓，能够有效降低误检测的情况发生，但同时也降低了对异常的敏感度，漏检率上升；反之，能够增加检测的灵敏度，但准确性相应的会降低。因此，窗口大小的选择，需要根据实际中查询速率的大小进行调整。

3 检测算法的验证

2009 年 5 月 19 日，多省市的递归服务器由于收到超负荷的 DNS 查询而失效，中国互联网出现了大范围的网络瘫痪事故，这起事故可以看作是一起典型的利用 DNS 查询发起的分布式拒绝服务攻击，这种突发的大量异常查询混入到正常的 DNS 查询中，必然会使 DNS 查询中查询类型的组成发生变化。我们利用从某顶级结点的 DNS 权威服务器上采集到的 2009 年 5 月 19 日 9:00-24:00 之间的查询日志，来检验算法是否能够对 DNS 中的异常行为做出反应。图(1)和图(2)分别是窗口大小为 1,000 和 10,000 时所得到的熵变化曲线，图(3)是该节点查询率曲线，处于保密的需要，图中数据均进行过处理。

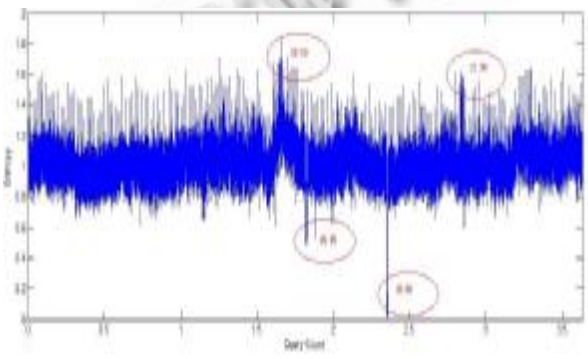

图 1 窗口大小为 1,000 时熵的变化情况

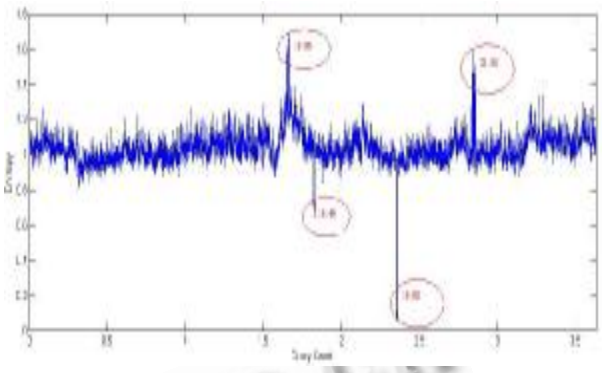

图 2 窗口大小为 10,000 时熵的变化情况

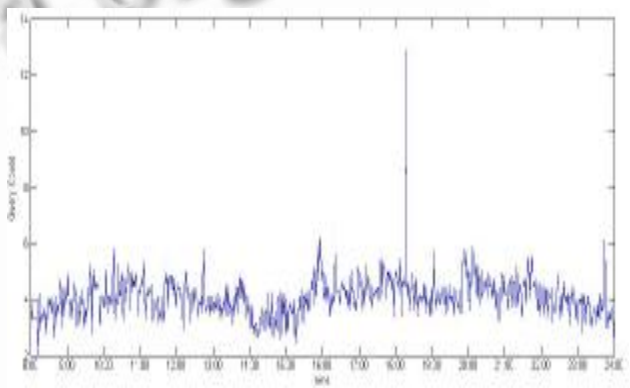

图 3 查询率曲线

4 结果分析

根据 5．19 事件的调查报告，黑客于 5 月 18 日对 DNSPOD 的服务器发起攻击，DNSPOD 是一个免费的域名服务提供商，承担着包括暴风影音的域名 BAOFENG.COM 在内的约十万个域名的解析工作，攻击当晚 DNSPOD 的电信主力服务器被迫离线，导致大批的域名无法解析，但是由于域名系统中递归服务器对查询结果的缓存作用，攻击的效果到 19 日才趋于严重。暴风影音在国内拥有数以亿计的用户，每台安装暴风影音的机器上都被加装了开机自动运行的与 BAOFENG.COM 进行通信的后台程序，由于负责为 BAOFENG.COM 提供解析的 DNSPOD 服务器已经当机，数量庞大的查询请求在得不到响应的情况下，疯狂地向递归服务器重复发送请求，19 日晚间开始，多省的 DNS 递归服务器因此崩溃，出现大面积网络瘫痪。

在大面积断网发生以前，大约从 5 月 19 日 16 点开始，ISP 的递归服务器缓存相继失效，大量的重复

查询请求开始涌入网络，一些递归服务器的重定向配置导致大量的查询请求被引到该顶级域的根服务器来，导致顶级域的根服务器所收到的 DNS 查询流量组成情况发生变化，针对 DNSPOD 所服务的域名的 A 类型和 NS 类型的查询大量涌入到正常查询流量中来，引起熵值的剧烈变化。从 21 点开始，随着查询量的持续增大并最终达到递归服务器所能承受的极限，多省的递归服务器相继失效，DNS 查询流量无法到达该顶级域的根服务器，在经历过剧烈波动后，DNS 查询的熵值也逐步回复正常。

从图(1)和图(2)中可以发现，大约从 16:00 时开始，熵值剧烈上升，这是由于此时系统中查询类型为 A 和 NS 的查询请求大量涌入，打破了系统原有的稳定态势，在经历较大的波动之后，又回复到一个稳定值。随着系统中缓存失效的递归服务器不断增多，该根服务器收到的异常数据量逐渐增大，在 16:45 左右熵值达到一个较低点，此时系统中已经混入了大量的异常查询数据。由于各省递归服务器的缓存设置的不一致，不断的有递归服务器崩溃，同时不断缓存失效的递归服务器加入，一直到 18:00 左右，这种异常查询量到达峰值，表现为熵值到达一个极低的位置，随着大批递归服务器在巨大的压力下瘫痪，查询数据的组成再次发生剧烈波动，接下来随着大面积断网的发生，异常查询无法到达该根服务器，熵值在经历波动之后又重新回到较稳定的状态，图(3)中的流量变化也证实了这一点。

图(1)和图(2)分别将查询窗口设为 1,000 和 10,000，对比两图可以看出，图(1)中的熵值变化较为频繁，反映出对 DNS 异常更加敏感，但同时误检测的几率也较高，图(2)中熵值的变化相对平缓，对异常情况敏感程度较低，同时误检率也相对较低。

图(3)中给出的是该时间段内的查询速率(查询次数/分钟)，从查询速率的变化来看，在 21:00 左右，查询流量出现了显著的异常，递归服务器缓存失效到崩溃这段时间，单纯的依据网络流量的变化无法判定异常的出现，到 21:00 点发现异常时，大面积网络瘫痪已经发生，已经错过了对 DNS 系统异常进行处理的宝贵时间。

此外，DNS 系统作为一个全球的、开放的系统，单纯依据查询流量的大小变化来判定系统是否出现异常，是不科学的，一方面是这种方式只能对流量累计到一定程度的异常做出反应，往往具有滞后性，不能起到异常预警的作用。另一方面，在网络范围很大的情况下，异常流量相对于总的流量而言往往只占很小一部分，这种方式很难起到检测效果。而基于熵的变化，则能比较客观的反映出查询流量中的各种随机出现元素的动态平衡情况，及早地发现异常情况的出现，从而及时采取相应的措施，避免服务质量下降甚至更严重的后果出现。

5　结束语

实验结果表明该方法能够及时发现 DNS 查询中的异常情况，提供预警信息。将该算法应用到 DNS 查询流量的实时监测中，可以做到准实时的发现 DNS 异常从而能够及早采取应对措施。此外，结合使用错误查询类型或者源 IP 地址等其他属性的分布来计算熵，或是采用时间窗口划分流量等，可以进一步提高异常检测的准确率，这也是下一步工作的重点。

鉴于 DNS 系统在互联网中的重要作用以及 DNS 系统崩溃所带来的严重后果，DNS 的安全问题越来越受到重视。DNS 系统异常的实时检测，结合国际上以 IETF 为首的一些组织提出的协议等改进措施，未来将对 DNS 系统的安全性和服务质量做出重大提升。

参考文献

1 Mockapetris P. RFC1034 Domain Names - Concepts and Facilitilies. 1987.

2 Mockapetris P. RFC1035 Domain Names - implementation and specification.1987.

3 维基百科. 熵. http://zh.wikipedia.org/zh-cn/熵

4 Shannon CE, A Mathematical Theory of Communi cation. The Bell System Technical Journal, 1948, 127:379－423,623－656.

5 维基百科. 熵(信息论). http://zh.wikipedia.org/wiki/熵_(信息论).

6 Feinstein L, Schnackenberg D, Balupari R, Kindred D. Statistical Approaches to DDOS Attack Detection and Response. Proc. DARPA Information Survivability Conf. and Exposition, 2003, IEEE CS Press,303－314.