一种基于索引的 TCP 数据流存储模型及其应用

戴宣 丁伟

东南大学计算机系 江苏南京 210096
江苏省计算机网络技术重点实验室
xdai@njnet.edu.cn wding@njnet.edu.cn

摘 要：本文提出了一种基于索引的数据流存储模型，它可以在不存储报文所有信息的前提下，完成面向数据流的有关测度和特性的计算，而这些参数如果采用传统方式需要在重复读取数据流信息的条件下才能获得。文章首先介绍了模型的原理和实现方案，然后完成了与传统方式的比较，最后是基于实际环境的测试结果。这些结果可以用于与网络行为学有关的研究。

关键词：数据流；存储模型；聚类；索引；网络行为学[1]

An Index-Based Storage Model for TCP Flow and Its Application

DAI Xuan， DING Wei，

(Dept. of Computer Science & Engineering, Southeast University, Nanjing 210096)

Key Lab of Computer Network Technology in Jiangsu Province

xdai@njnet.edu.cn wding@njnet.edu.cn

Abstract: An index-based storage model for flow is proposed. With the model, the calculation of data flow metrics and its characteristic can be completed, instead of storing the whole information of packets. In traditional way, all the result should be obtained by re-extracting data flow information again and again. First, the paper gives the description on principle and scheme of the model; then, a comparison between the model and traditional way is made; finally, a trace is analyzed with this model. Some interesting characters appeared, which can be used for network behavior research.

Keywords: Data Flow; Storage Model; Aggregation; Index; Network Behavior

1．引言

网络测量和分析是网络行为学研究的基础。随着网络应用的发展，网络带宽增加，在现有网络特别是主干网中，单纯基于报文的测量已经远远不能满足实际的需求，基于数据流的测量方法应运而生。所谓数据流，是指符合特定的流规范（specification）和超时（timeout）约束的一系列数据报文的集合[2]。本文定义数据流规范是将 IP 报文头按（源 IP，宿 IP，源端口，宿端口，协议号）五元组分类，在满足超时约束的条件下，每一类就是一个数据流。此外，也可以按照五元组中的若干元进行分类，便于观测端到端的网络流量[3]，如 MPLS 中的流可以视为仅仅按源宿 IP 的分类。这可以视为五元组数据流的聚类。

数据流信息的获取过程可以简单归纳如下：首先采集报文，再对报文进行数据流提取，然后经过统计计算，再得到数据流信息。采集的报文信息决定了可以获得的数据流信息。数据流信息可分为两类。第一类数据流信息只需要对相关报文进行一次扫描即可完成,例如数据流所含的报文个数、字节数以及数据流的持续时间等。这可以在数据流提取过程的同时，完成对这些信息的统计[4];第

二类数据流信息必须要对相关报文进行多次扫描才能获得，如计算给定五元组的所有报文的字节数方差等。这就需要记录所有相关报文信息，便于再统计。但是，因为系统资源的限制，将报文的所有信息都记录在内存中已不现实，内存中只能进行诸如报文个数之类的简单统计[5]。例如，CERNET 主干双向平均流量在 600Mbps 以上，处于高峰时超过 1.1Gbps，如果在测量过程中，还要记录报文的所有信息，那么即使是处理离线报文，测量系统的资源也将十分紧张。因此，为了获取第二类数据流信息，就要多次扫描报文，重复数据流提取过程。这又将产生大量的时间耗费。

针对获取第二类数据流信息存在的问题，本文提出一种基于索引的数据流信息存储模型，在不存储报文所有信息的前提下，避免重复的数据流提取过程。应用该存储模型，本文还对 CERNET 省网边界采集的报文进行了实验，并对实验结果进行了分析，得出了相关结论。

2．存储模型介绍

针对获取第二类数据流信息存在的问题，本文希望寻找一种存储模型，可以尽可能避免数据流提取的次数，从而节省计算时间。为此，本文提出为报文和数据流建立索引来避免多次扫描报文。一旦在数据流提取完成之后，就可以直接定位所要查找的报文。存储模型的建立过程可以分为两个步骤，首先遍历所有报文，完成数据流提取，同时建立报文索引；其次，遍历提取出的数据流，建立相应的数据流索引。

模型设计的主要思路如下：存储数据流时，存放尽可能少的流信息，而是将内存空间用于存放每个属于该数据流的报文索引；此外，还要为提取出的数据流建立一个数据流索引，方便查找给定五元组的所有数据流。因此，存储模型本质上是一个二级索引。

2.1 相关数据结构

2.1.1 数据流存储结构

首先定义报文索引结构如下：报文所在文件编号（4 字节）+报文所在文件偏移（4 字节）。文件编号为存放该报文的文件编号；文件偏移表示该报文离文件开头处的偏移值。通过这两个值，就可以定位一个报文的位置，读出已存储报文的任何信息。

基于以上定义的结构，本文给出数据流存储结构如下：

五元组信息	报文个数	统计信息 1…n	报文索引结构 1... m

图 1

其中，统计信息是若干第一类数据流信息，如报文个数、字节数等。在第一次数据流提取过程中，可同时获得统计信息，并记录报文索引结构。

2.1.2 数据流索引结构

为了可以直接查找具有相同五元组的数据流，还需要为已经得到的数据流存储结构建立索引，称为“数据流索引结构”。数据流索引结构如下：数据流所在文件编号（4 字节）+数据流所在文件偏移（4 字节）。文件编号为该数据流存储结构所在的文件编号；文件偏移表示该数据流存储结构距文件开头处的偏移值。通过这两个值，就可以定位一个数据流存储结构的位置，读出数据流存储结构的任何信息。

在数据流提取过程中，可以根据数据流的超时约束设定超时值，保证到达超时值之后，及时将内存中已超时的数据流写入外存，避免内存溢出。而在建立索引的过程中，没有超时约束，所以为了避免内存溢出，也需要将数据流索引结构及时写入外存。但是，相同五元组的数据流可能还未被完全记录在索引结构中，如此将导致索引结构分散存储在外存的不同地方。为此，在数据流索引结构中增加一个“拉链结构”，用于相同五元组的下一个数据流索引结构被写入外存后，存放其所在的文件号和偏移值。因此，其中的拉链结构必须等待下一个数据流索引写入外存时才能重新写入，这个过程称为“回填”。基于上述目的，内存中的数据流索引结构，可以利用该拉链结构存放上一

个数据流索引结构的拉链结构位置，以备回填。

数据流索引结构如下：

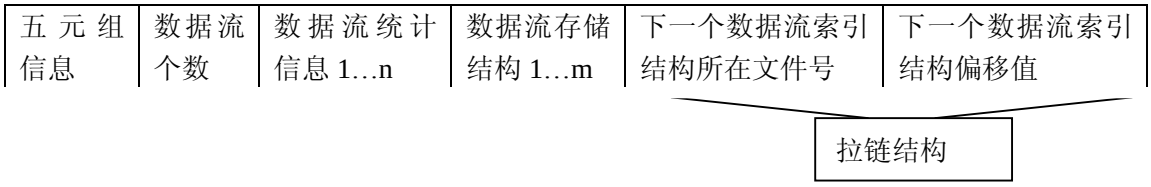

图 2

2.1.3 五元组索引文件

对所有流进行按照五元组聚类，记录每个数据流的索引，建立索引文件。便于查找特定五元组产生的数据流。五元组索引文件中索引项结构如下：

五元组信息	数据流索引结构个数	五元组统计信息1…n	第一个数据流索引结构所在文件号	第一个数据流索引结构偏移值

图 3

综上，上述三种数据结构形成了两级索引，其结构关系如图 4 所示。

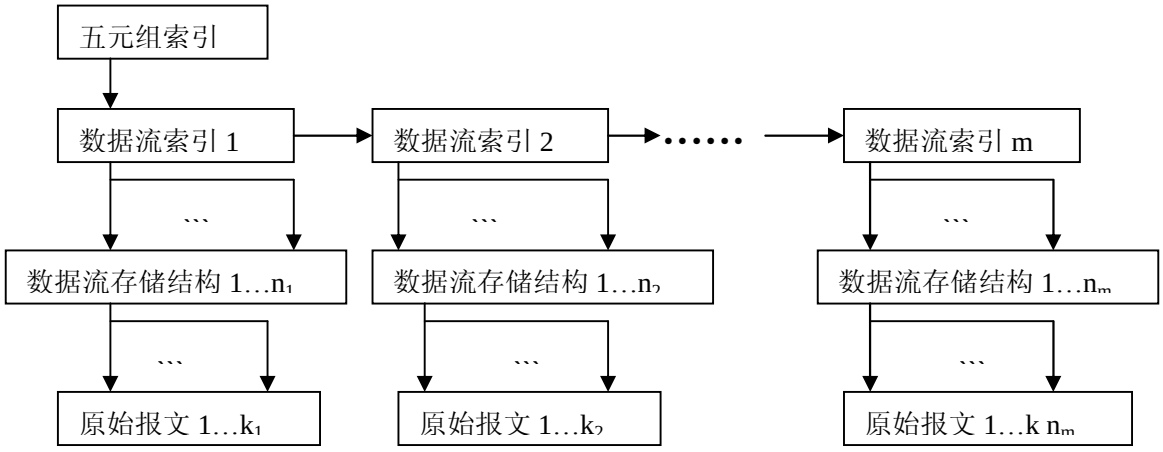

图 4

对应于上述数据流存储模型，下文将阐述相关的操作流程。

2.2 操作流程

整个模型的相关操作可分为三个部分：数据流提取过程、数据流索引建立过程、数据流查找过程。提取过程扫描所有报文，完成数据流的提取工作，同时建立报文索引并统计第一类数据流信息。索引建立过程扫描所有已经提取的数据流，同时建立数据流索引并生成五元组索引文件。查找过程根据给定的五元组直接查找该数据流的报文。其中，索引建立过程需要使用“回填”机制。

2.2.1 数据流提取过程

内存中，每条数据流按图 1 结构存储。“报文索引结构指针”指向一个报文索引结构链表；“当前最后一个报文到达时刻”就是指报文索引结构链表中，最近被插入报文的到达时刻。整个数据流的提取过程，实际上就是该内存模型的维护过程。

具体流程如下：

1. 报文是否读完，是转 7；否则转 2
2. 读取一个新报文，判断所属流的存储模型是否存在，是转 4；否则转 3
3. 为该数据流创建一个如图 1 的结构，转 6

4. 判断数据流是否应该结束（如超时或者报文是否为 FIN/RST 等）。若是，转 5；否则，转 6
5. 流结束。将数据流按照存储模型格式写入外存，并释放所占用的所有内存。转 3
6. 将该报文所在的文件号与偏移值插入“报文索引结构指针”所指向的链表中，并更新“当前最后一个报文到达时刻”及所有统计信息字段，报文个数字段加 1。转 1
7. 将内存中所有的数据流存储结构，按照存储模型写入外存。数据流提取过程结束。

2.2.2 数据流索引建立过程

在此过程中，前一个数据流索引结构的拉链结构位置，被存放在当前数据流索引结构中的拉链结构中，以备回填。此外，为了避免内存溢出，必须设定一个阈值。当可用内存少于该阈值时，则强制将内存中当前的索引结构写入外存。该阈值记为 ML。ML 可根据系统资源确定。

具体流程如下：

1. 数据流存储结构是否读完，是转 7；否则转 2
2. 读取一个新数据流存储结构，判断其索引模型是否存在，若是转 4；否则转 3。
3. 为该流创建一个如图 2 的结构，并置下一个数据流索引结构的文件号偏移值都为 0。转 6
4. 当前可用内存是否小于 ML。若是，转 5；否则，转 6
5. 将内存中所有数据流索引结构按照存储模型格式写入外存，并保留每个索引结构的五元组信息。之后，用已写入外存的拉链结构所在的文件号及偏移值，更新内存中的拉链结构。转 1
6. 将该数据流存储结构所在文件号与偏移值插入“数据流存储结构指针”所指向的链表中。数据流个数字段加 1。转 1
7. 将内存中所有的数据流存储结构写入外存，并完成“回填”。数据流索引建立过程结束。

2.2.3 数据流查找过程

当数据流索引建立完成之后，就可以查找任意给定五元组的数据流报文，过程如下：

1. 给定数据流的五元组信息。
2. 查询索引文件，找出第一个数据流所在文件的编号及其偏移，作为当前数据流所属的文件编号和偏移。若找到，则转 3；否则，该数据流不存在，转 9
3. 与文件编号相对应的文件是否打开?是转 5；否则，转 4
4. 如果有数据流文件打开，则关闭之。
5. 如果文件编号和偏移值都是 0，则转 9。
6. 根据偏移值，找到文件的指定位置，读取数据流报文个数，记为 N。
7. 根据 N，在文件中连续读取 N 个报文索引；可以根据需要，根据报文索引，读取报文信息
8. 读取最后的拉链结构，将其文件编号和偏移值，作为当前数据流的文件编号和偏移值。转 3
9. 数据流查找结束。

对采集得到的原始报文，在完成了数据流提取过程与数据流索引建立过程之后，就建立了如图 4 的索引结构。之后，就可以利用该索引结构，进行数据流查找。其中，数据流提取与索引建立最为耗时，本文将就所有过程的时间耗费进行分析比较。

3．性能分析

索引的优点主要在于查找的便利，特别是对于无序数据的查找。在没有索引的情况下，如果查找特定五元组的所有报文，需要对所有报文进行一遍扫描，同时判断每个报文是否为所要查询的报文；借助上文介绍的索引结构，则可以精确定位同属一个五元组的所有报文位置。将无索引查找方法的时间耗费记为 T，索引查找方法的时间耗费记为 T’，遍历全部报文的时间耗费记为 $t_{travelall}$，判断报文是否为所要查询报文的时间耗费记为 $t_{verdict}$，建立数据流索引的时间耗费记为 t_{index}，根据索引遍历报文的平均时间耗费为 t_{travel} 则利用两种方法查找 N 个五元组的时间耗费如下表示：

$$T=(t_{travelall}+t_{verdict})\times N \quad ① \qquad T'=t_{travelall}+t_{index}+t_{trave}\times N \quad ②$$

对于②式而言，$t_{travelall}$ 就表示一次数据流提取的时间耗费。需要注意的是：$t_{travelall}+t_{index}$ 的时间消耗是预先完成的，在每次查找过程中，索引查找方法实际耗费的时间只有 $t_{trave}\times N$；而无索引查找方法则仍旧是$(t_{travelall}+t_{verdict})\times N$。

下面以实际采集的报文作为实验数据，根据五元组所含报文数大小分类，分别利用两种方法，对其报文进行扫描。实验机器配置为 CPU Xeron2.4×2，内存 2G，磁盘阵列 raid5，硬盘总容量 1.6T。

实验结果如表 1 所示：

五元组所含报文个数	五元组个数	无索引查找方法		索引查找方法	
		总耗时	单个五元组平均耗时	总耗时	单个五元组平均耗时
90-110	30	59 分钟	1.96 分钟	47 秒	1.56 秒
990-1010	30	62 分钟	2.06 分钟	90 秒	3 秒
9090-10010	30	63 分钟	2.1 分钟	370 秒	12.3 秒
90090-100010	30	65 分钟	2.16 分钟	40 分钟	1.3 分钟

表 1

另外，本次实验中，遍历所有报文需要耗费时间 60 分钟，建立索引需要的时间为 60 分钟。其中，索引一旦建立就可以被反复使用，相比每次查找都需要遍历所有报文，将节省大量时间。综上所述，在对数据流所含报文进行统计时，带索引的数据流存储模型效率更高。

4．存储模型的应用

依照上述描述的存储模型，只要给定五元组，就可以在较短时间内查找出该五元组所含的所有报文。考虑到网络中的 TCP 流数量占有绝对优势[6],本文将使用上述存储模型，对 TCP 报文进行分析，观察五元组所产生 TCP 数据流的个数分布和最大 TCP 流长度的变化趋势。

本文参考[7]，为了保护数据流的完整信息，选取 64 秒作为数据流的超时限制。实验中使用了 CERNET 一个省网边界采集的 10 分钟原始报文，信道为 3 根双向光纤，每根光纤带宽容量为 2G。采集时间为 2005-11-10 11:00:00 到 2005-11-11 11:10:00，采集长度 60 个字节，采集报文总数为 370085640。

4.1 四元组（源宿 IP 及源宿端口）产生 TCP 数据流的数量分布

对 trace 中出现的所有四元组，统计每个四元组产生的所有数据流数量。具有相同四元组的 TCP 数据流数量的累计分布图形见图 5。

图中 99%的四元组所含的数据流个数都在 10 个流以下，而产生了较多数据流的四元组只占四元组总数的 1%。可以推断，大量使用临时端口，导致了四元组数量较多，但是单个四元组产生的数据流数量有限，所以很多四元组产生的数据流数量都较少。而产生数据流较多的四元组，可能是因为两端指定了端口，使得每次通信产生的数据流都有相同的四元组。由此说明，给定四元组，产生较多数据流的可能性较小。

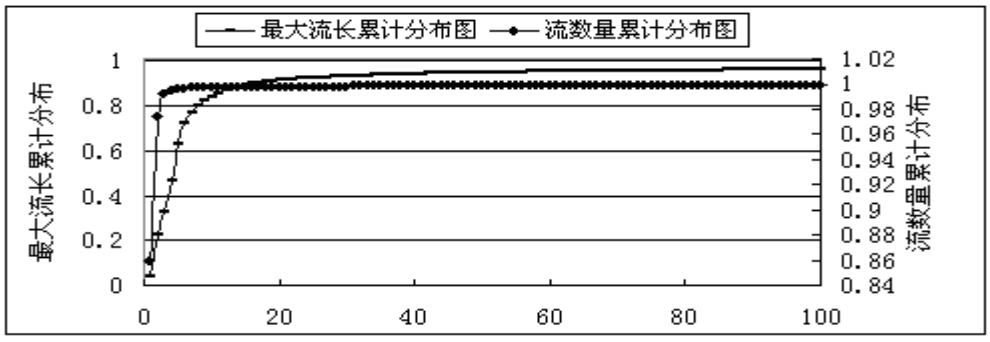

图 5

4.2 四元组产生的最大流长分布

对于上图中产生流个数在 10 个以下的所有四元组，再统计每个四元组产生的数据流中，最长流的长度分布。最大流长累计函数图形如图 5 所示。

根据文献[2]，根据所含报文的数量，可以将数据流分为短流（报文数小于 10）、长流（报文数大于 10 而小于等于 1000）、超长流（报文数大于 1000）。根据图 5，在产生流个数在 10 个以下的所有四元组中，只产生短流的四元组占据了 84.1%，而产生了长流与超长流的四元组占据了 15.9%。由此可见，特定两端之间产生的数据流数量虽然少，但是仍然可能产生长流与超长流。实验中，TCP 数据流总数为 4579828，其中长流和超长流个数为 548644。另经统计得到，在产生流个数在 10 个以下的所有四元组中，产生长流和超长流个数达到了 546468，占长流和超长流总数的 99.6%。可以推断，给定四元组，虽然固定了端口，产生较多个数的数据流，但是每次连接产生的报文个数有限。例如，可能是因为每次连接不成功，导致多次重试所致；或者是多次重复扫描所致。由此可说明，四元组产生长流与超长流的可能性，与其产生的数据流数量无必然关联；产生数据流少的四元组仍然可能产生长流与超长流。

5．结论

针对网络数据流分析的需要和系统资源的限制，本文提出了一种 TCP 数据流存储模型，便于查找特定五元组的数据流信息。叙述了该模型的相关数据结构以及操作流程，并通过实验，将该模型与传统的存储模型进行了比较。实验证明，本文的数据流存储模型所需查找时间更短。

本文还应用该模型对实际的 Trace 进行了观测，并得到以下结论：

1) 绝对多数的四元组都只产生个数有限的数据流，能产生大量数据流的四元组很少。
2) 在 TCP 数据流中，给定四元组，虽然产生的数据流个数少，但是仍有可能产生长流与超长流。从观测中发现，长流与超长流的产生与数据流产生的数量无必然联系。

本文提出的存储模型十分有助于基于数据流的离线测量。虽然该模型在查找上具有较好性能，但主要针对离线报文。对实时测量而言，该模型有待进一步的改进。

参考文献

[1]“Measuring the Internet” KC Claffy IEEE Internet Computing Online, v4n1 January 2000

[2]“Internet Flow Characterization: Adaptive Timeout Strategy and Statistical Modeling”. B.Ryu, D.Cheney, H.W.Braun. In Workshop on Passive and Active Measurement(PAM), Apr, 2001. pp. 94-105

[3].Their share: diversity and disparity in ip traffic. Broido, Y.Hyun, R.Gao, and kc claffy. In PAM, 2004.

[4]“A New Algorithm for Long Flows Statistics: MGCBF” Zhou,M.Z. Gong. J Ding W Cheng G Journal of Southeast University 2006 May Vol.36,No.3, p472-476

[5]“A Robust System for Accurate Real-time Summaries of Internet Traffic” Ken Keys, David Moore, Cristian Estan SIGMETRICS’05 June 6-10,2005

[6] “Longitudinal study of Internet traffic in 1998-2003” Marina Fomenkov, Ken Keys, David Moore, KC Claffy Winter International Symposium on Information and Communication Technologies (WISICT) on January 5-8th, 2004 in Cancun, Mexico page: 1-6

[7]”A parameterizable methodology for Internet traffic flow profiling” Kimberly C.Claffy,Hans-Werner Braun,George C.Polyzos IEEE Journal on Selected Areas in Communications, 13(8):1481--1494, October 1995.

第一作者介绍

戴宣，男，1982 年，东南大学计算机科学与工程学院,硕士生，研究方向：网络行为学；

联系方式:手机:13645164653 固定电话: (025)83794000-211 Email: xdai@njnet.edu.cn

通信地址：南京市东南大学华东北地区网络中心 邮编:210096