Desarrollo de una metodología para la determinación de indicadores de la utilización de servicios en una red IP

AUTORES:
Andrés Leiva, *aleiva@inictel-uni.edu.pe*
Fredy Chalco, *fchalco@inictel-uni.edu.pe*

ASESORES:
José Quiroz, *jquiroz@inictel-uni.edu.pe*

Área de conmutación y Transmisión del Inictel-UNI, Universidad Nacional de Ingeniería, Perú.

Resumen: Las redes en la actualidad son cada vez más extensas y complejas que realizar la planificación y el dimensionamiento mediante el estudio de su tráfico no es una tarea sencilla. Tanto el número de usuarios conectados como la cantidad de servicios ofrecidos está en aumento debido al incremento en la velocidad de los enlaces producto del despliegue de la banda ancha y también debido a las nuevas tecnologías de hardware y a la innovación constante de servicios tales como el de correo electrónico, almacenamiento de archivos, IPTV, etc. Cada día aparecen nuevos usuarios de estos servicios y esto hace que se reporte un aumento del nivel de tráfico en las interfaces de red de los correspondientes servidores. Para un administrador de red y para una entidad que ofrece servicios de red en general es de suma importancia el conocer los destinos más relevantes de sus servicios así como el nivel de tráfico de los mismos. En este artículo se presenta el desarrollo de la metodología utilizada en [1] sobre una técnica de modelamiento estructural basado en la teoría de la información. El sistema es diseñado con la finalidad de determinar indicadores de la utilización de servicios de red que permita identificar los destinos relevantes de cada servicio así como el nivel de tráfico de cada uno. Los resultados de las pruebas realizadas muestran la utilidad de este sistema en la identificación de usuarios frecuentes y no frecuentes así como de aquellos servicios que son los más preferidos como los que requieren ser innovados para mejorar la aceptación de los usuarios. Asimismo, se muestran resultados experimentales de pruebas realizadas en el enlace troncal de una red académica.

Palabras clave: Modelamiento estructural, servicios de red, flujo de red, entropía, metodología.

1. Introducción

Los modelos clásicos del tráfico de redes, se basan en un modelo matemático que describen diversas variables cuantitativas relacionadas con el tiempo tales como el tiempo de espera, retardo, etc. Sin embargo, estos modelos no aportan suficientes valores cualitativos que describen la naturaleza de los paquetes IP tales como la dirección IP origen y destino así como los números de puerto en una comunicación entre un cliente y servidor.

En [1] se expone una metodología con la cual se obtiene un modelo estructural compacto que resume todas las características del tráfico. Dicha metodología utiliza el concepto de flujo de paquetes para realizar el proceso de modelamiento el cual consiste de básicamente tres etapas: extracción de flujos significativos, clasificación de flujos significativos en clases de comportamiento y modelamiento estructural.

Esta metodología particular se explica en [1]. Sin embargo, no se dan las pautas específicas para su programación y para la implementación en una red experimental. En este artículo, se expone los pasos seguidos para programar e implementar un sistema genérico que utiliza dicha metodología y es aplicada de forma particular para evaluar la utilización de servicios de red utilizando algunos indicadores mencionados en dicho artículo.

Este artículo se divide la siguiente manera. Primero se brinda una explicación general de la metodología empleada, luego se describen cada una de las etapas de desarrollo del sistema, después se presentan los resultados obtenidos durante las pruebas realizadas en una red académica y por último se presentan las conclusiones.

2. Conceptos Básicos

Para el buen entendimiento de la metodología es necesario tener claro dos conceptos principales: flujo de red y entropía de la información.

2.1 Definición de flujo

Un flujo se define como un conjunto de paquetes atravesando un punto de observación en una red durante un cierto intervalo de tiempo. Todos los paquetes de un flujo tienen un conjunto de propiedades comunes definidas [4]. En este artículo se considera como propiedades comunes a los siguientes cinco campos de cabeceras:

- Dirección IP origen y destino.
- Puerto origen y destino.
- Protocolo.

2.2 Definición de Entropía

La entropía mide el grado de incertidumbre de una variable aleatoria, es decir, mide la variedad de la observación de un conjunto de valores de una variable aleatoria discreta X. Suponiendo que X puede tomar N_X valores diferentes. Sea m el número de observaciones de la variable X. Se induce una distribución de probabilidad empírica de X, de la siguiente naturaleza $p(x_i)=m_i/m$, tal que $x_i \in X$, donde m_i es el número de veces que se observa la variable aleatoria X tomando el valor x_i. La entropía de la variable aleatoria X se denota como H(X) y está definida por la siguiente ecuación [3]:

$$H(X) = -\sum_{x_i \in X} p(x_i).log_2 p(x_i)$$

Por convención se tiene que $0.\log_2 0 = 0$. El rango de valores que puede tomar la entropía es el siguiente:

$$0 \le H(X) \le H_{max}(X)$$

$$H_{max}(X) = log_2[min\{N_X, m\}]$$

El máximo número de posibles valores únicos que la variable aleatoria X puede tomar en m observaciones es $2^{H_{max}(X)}$. Se asume que $m \geq 2$ y $N_X \geq 2$, de lo contrario, no tendría sentido emplear el concepto de variedad de la observación.

La entropía estandarizada o incertidumbre relativa es una medida especial definida en la metodología. Esta medida proporciona un índice de la variedad o uniformidad sin tener en cuenta el número de valores diferentes "N_X" que puede tomar una variable aleatoria ni el número de observaciones "m". Esta medida se denota como RU (Relative Uncertainty) y está definida por la siguiente expresión:

$$RU(X) = \frac{H(X)}{H_{max}(X)} = \frac{H(X)}{log_2[min\{N_X, m\}]}$$

Cuando RU(X) = 0, quiere decir que todas las observaciones de X son las mismas, es decir, p(x) = 1 y por lo tanto, no existe variedad de la observación.

Generalizando, sea A un subconjunto de X perteneciente al conjunto de los diferentes valores observados de X en un experimento aleatorio. Se consideran 2 medidas de la uniformidad de los valores observados de X la entropía condicional denotada por H(X|A) y la respectiva incertidumbre relativa condicional denotada por RU(X|A), definidas por las siguientes expresiones:

$$H(X|A) = H(X)$$

$$RU(X|A) = \frac{H(X)}{log_2|A|}$$

$$H_{max}(X|A) = log_2|A|$$

El valor de |A| es la cantidad de valores del conjunto A, es decir, el número total de diferentes valores que se observaron.

Cuando $RU(X|A) \approx 1$, indica que los valores observados están cerca de estar uniformemente distribuidos y por lo tanto son casi indistinguibles unos de otros.

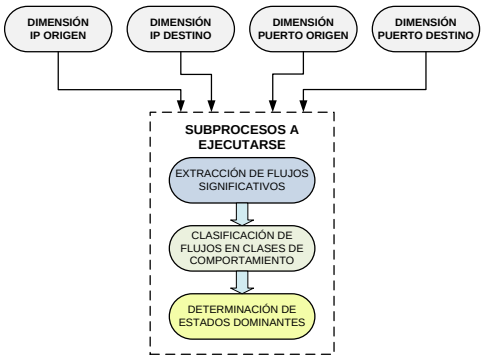
Cuando $RU(X|A) \ll 1$, indica que los valores observados está sesgado, es decir, contiene unos pocos valores que son observados con mayor frecuencia que el resto.

3. Breve descripción de la metodología

La metodología utiliza el concepto de flujo para realizar el procesamiento de la información. Asimismo, se hace un análisis del tráfico desde 4 perspectivas diferentes cada una de ellas corresponde a una dimensión las cuales son:

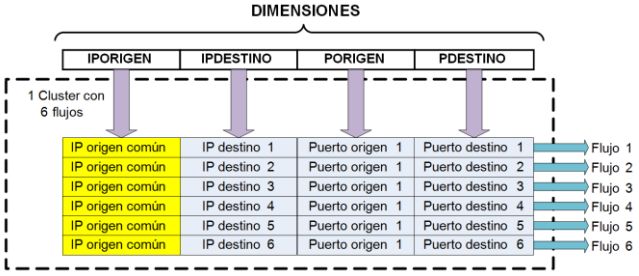
- Dimensión IP origen denotada por "IPORIGEN".
- Dimensión IP destino denotada por "IPDESTINO".
- Dimensión puerto origen denotada por "PORIGEN".
- Dimensión puerto destino denotada por "PDESTINO".

Se realiza cuatro procedimientos para cada dimensión. Estos procedimientos son la extracción de flujos significativos, la clasificación de flujos en clases de comportamiento y la determinación de estados dominantes. La siguiente figura muestra un esquema donde se representa el proceso general de la metodología.



3.1 Extracción de flujos significativos

Primero se capturan todos los flujos observados en un enlace de red y luego se hace un filtrado de aquellos flujos relevantes en comparación con el resto. Se utiliza el término cluster para referirse a un conjunto de flujos o registros de información de flujos que tienen el mismo valor en el campo de una determinada dimensión. En la siguiente figura se muestra un ejemplo particular que refleja el concepto de un cluster como un conjunto de flujos que tienen el valor de la dirección IP origen en común.



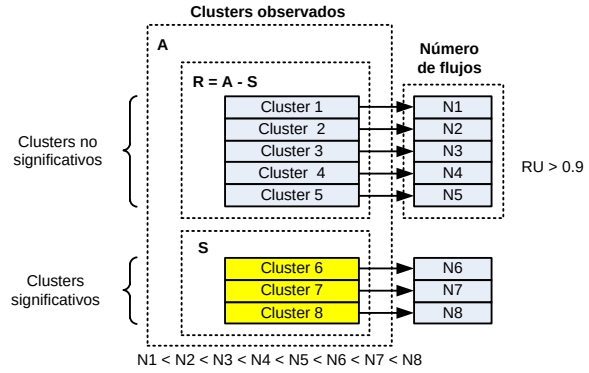
El criterio utilizado para la extracción de flujos significativos está basado en el concepto de incertidumbre relativa (RU). Primero se detectan los diferentes clusters observados en una de las 4 dimensiones. Como cada uno de los clusters tiene una cantidad de flujos fija se genera una distribución de probabilidad P_A para cada cluster a_i de la siguiente manera:

$$P_A := P(a_i) = m_i / m$$

Donde m_i es el número total de flujos en el cluster a_i y m es el número total de flujos observados.

El subconjunto de valores significativos del conjunto A es denotado por S y el resto de valores no significativos se denota como R. La probabilidad de cualquier valor en S es mayor que la de los valores restantes. La distribución de probabilidad condicional del conjunto R está cerca de ser uniformemente distribuida, es decir que $RU(P_R) > \beta = 0.9$. El subconjunto S contiene los clusters más significativos de A mientras que los valores restantes son casi indistinguibles unos de otros y por lo tanto no son relevantes para el análisis.

La siguiente figura muestra un ejemplo en donde se descartan cinco de un total de 8 clusters utilizando esta metodología.



Para obtener los clusters del subconjunto S, se ordenan los clusters en orden ascendente según sus probabilidades y se calcula la incertidumbre relativa de su distribución. Si este valor no es mayor que 0.9 se elimina el cluster con mayor probabilidad y se calcula la incertidumbre relativa del conjunto restante. El proceso continúa de manera análoga y se detiene cuando la incertidumbre relativa de un conjunto restante supera el valor de 0.9 por primera vez. Los clusters que se han eliminado corresponden a los clusters significativos de la dimensión analizada. El proceso es análogo para las demás dimensiones. La salida de este proceso es almacenado en un arreglo multidimensional el cual es utilizado como entrada en el siguiente proceso de clasificación.

3.2 Clasificación de flujos en clases de comportamiento

Este proceso consiste en la clasificación de cada cluster significativo en 27 clases de comportamiento definidas según determinadas condiciones. Cada una de estas clases define un patrón de comportamiento especial.

Cada flujo o registro de flujo dentro de un cluster significativo comparten un mismo valor en el campo de la dimensión que se está analizando (dimensión fija); sin embargo, las otras tres dimensiones (dimensiones libres) pueden tomar cualquier valor posible. Por consiguiente, cada flujo en un cluster induce una distribución de probabilidad en cada una de las tres dimensiones libres.

Para cada cluster significativo se denota a sus dimensiones libres como X, Y y Z. El orden de las notaciones se define según la convención establecida en la Tabla 5.1.

Dimensión fija	Dimensiones Libres		
	X	Y	Z
IPORIGEN	PORIGEN	PDESTINO	IPDESTINO
IPDESTINO	PORIGEN	PDESTINO	IPORIGEN
PORIGEN	PDESTINO	IPORIGEN	IPDESTINO
PDESTINO	PORIGEN	IPORIGEN	IPDESTINO

Cada cluster es caracterizado por un vector de incertidumbres relativas correspondiente a cada una de las dimensiones libres. El vector es denotado por $[RU_X, RU_Y, RU_Z]$. Cada RU se divide en 3 categorías según su valor. Cada categoría está etiquetada por los números 0 (bajo), 1 (medio) y 2 (alto) según el criterio mostrado en la siguiente tabla.

CATEGORÍA	CONDICIÓN
L=0 (bajo)	$0 \le RU \le \zeta$
L=1 (medio)	$\zeta < RU < 1 - \zeta$
L=2 (alto)	$1-\zeta \le RU \le 1$

El valor del parámetro ζ depende de la dimensión libre. Si la dimensión libre es IPORIGEN o IPDESTINO, entonces $\zeta=0.3$. Si la dimensión libre es PORIGEN o PDESTINO, entonces $\zeta=0.2$.

Se clasifican los clusters en 27 clases de comportamiento representados por un vector de etiquetas $[L(RU_X), L(RU_Y), L(RU_Z)]$. Este vector es tratado como un número entero llamado identificador de clase "N" según:

$$N = 3^2.L(RU_X) + 3.L(RU_Y) + L(RU_Z) \in \{0,1,2,3,4...,26\}$$

Cada clase de comportamiento se refiere como BCN, (Behavior Class N) donde "N" es el identificador de clase. Los clusters extraídos utilizando las diferentes dimensiones fijas tienen sus propias clases de comportamiento y cada uno tiene significado e interpretación diferente. Se le denota como IPORIGEN BCN, IPDESTINO BCN, PORIGEN BCN y PDESTINO BCN a las clases de comportamiento en las respectivas dimensiones fijas.

El proceso de clasificación se repite para cada uno de los clusters de la dimensión fija en cuestión. El proceso es análogo cuando se toma como dimensión fija a otras dimensiones.

Se definen tres métricas que describen las características temporales de las 27 clases de comportamiento. Tales métricas se muestran en la siguiente tabla:

Métrica	Descripción	Definición matemática
Popularidad Π_i	Número de intervalos de tiempo en promedio en donde se observa la aparición de una clase de comportamiento.	$\Pi_i=\frac{O_i}{T}$ $0 \le \Pi_i \le 1$
Tamaño promedio Σ_i	Número promedio de clusters que pertenecen a una clase de comportamiento particular en cada intervalo de tiempo donde se observa la clase.	$\Sigma_i=\sum_{j=1}^{j=T}\frac{C_{ij}}{O_i}=\frac{1}{O_i}\sum_{j=1}^{j=T}C_{ij}$
Volatilidad Ψ_i	Mide la tendencia de que una clase de comportamiento contenga diferentes clusters todo el tiempo, es decir, si los mismos clusters vuelven a aparecer con el tiempo o si tienden a aparecer nuevos clusters.	$\Psi_i=\frac{U_i}{\sum_{j=1}^{j=T}C_{ij}}=\frac{U_i}{\Sigma_i O_i}$ $0 \le \Psi_i \le 1$

Donde:
i.- identificador de clase de comportamiento.
T.- Número de intervalos de tiempo de 5 minutos.
C_{ij}.- Número de clusters observados de la clase BC_i en el intervalo de tiempo j.
O_i.- Número de intervalos de tiempo donde $C_{ij} \neq 0$ para la clase BC_i. ($O_i \leq T$).
U_i.- Número de clusters únicos (que no se repiten) de la clase BC_i observados en todo el periodo de tiempo. Estos clusters aparecen una sola vez en cualquier y sólo en un único intervalo de tiempo j de todos los "T" intervalos de tiempo del análisis.

3.3 Determinación de estados dominantes

Inmediatamente después de clasificar un cluster de una dimensión específica, se determinan los estados dominantes del mismo. Se construye un modelo del tráfico observado utilizando la técnica de análisis de estados dominantes que consiste en la caracterización de las interacciones entre clusters de una dimensión específica. Esta técnica se basa en conceptos de modelamiento estructural [2] y análisis de reconstrucción [5] de la teoría de sistemas.

El objetivo del análisis de estados dominantes es explorar la interacción o dependencia entre las dimensiones libres identificando subconjuntos de valores más simples llamados estados dominantes de un cluster o modelos estructurales según [2] los cuales representan o aproximan la naturaleza de los flujos que constituyen un cluster.

Los estados dominantes presentan las siguientes características:

a. Reproducen la distribución de probabilidad original de las dimensiones libres de un cluster con exactitud razonable.
b. Proporciona un resumen compacto de cada cluster de una dimensión.
c. Permiten predecir, con una exactitud razonable, el tipo habitual de transacciones existentes en la red y cómo se comportan cada una de ellas.

Para extraer los estados dominantes de un cluster, se determinan los valores sustanciales de cada dimensión libre según el siguiente procedimiento:

Se realiza un reordenamiento de las dimensiones libres del cluster según su incertidumbre relativa de menor a mayor. En caso de empate en las incertidumbres relativas, la dimensión X precede a Y ó Z y la dimensión Y precede a Z. Las dimensiones reordenadas se denotan por "A", "B" y "C" tal que $RU(A) \le RU(B) \le RU(C)$.

Se determinan los valores sustanciales de la dimensión "A". Se calcula la probabilidad marginal de cada objeto "a" de la dimensión "A". Si la probabilidad marginal de cierto objeto "a" supera un umbral, se considera a "a" como valor sustancial, es decir:

$$p(a)=\sum_{b\in B}\sum_{c\in C}p(a,b,c) \geq \delta=0.2$$

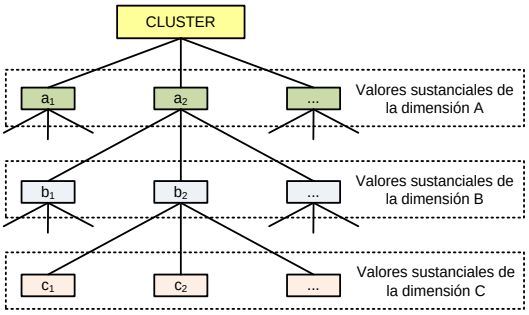
El valor δ es un valor umbral para la selección de valores sustanciales, se utiliza el valor 0.2. En caso de que no exista ningún valor sustancial el proceso se detiene.

Se explora la dependencia entre las dimensiones "A" y "B" calculando la probabilidad condicional de observación de un valor "b_j" en la dimensión "B" dado cada valor sustancial "a_i" de la dimensión "A" previamente determinado, es decir:

$$p(b_j/a_i)=\sum_{b\in B}\frac{p(a_i,b_j,c)}{a} \geq \delta=0.2$$

En caso de que no exista ningún valor sustancial "b_j" el proceso se detiene. Por último, se calcula la probabilidad condicional de observación de un valor "c_k" en la dimensión "C" dado cada combinación de valores sustanciales "a_i" de la dimensión "A" y "b_j" de la dimensión "B" previamente determinados, es decir $p(c_k \mid a_i, b_j) \geq \delta$. En caso de que exista o no exista algún valor sustancial "c_k" el proceso se detiene. El proceso total termina cuando se han determinado todos los posibles valores sustanciales a_i de la dimensión "A", todos los posibles valores sustanciales b_j de la dimensión "B" dado cada a_i y todos los posibles valores sustanciales c_k de la dimensión "C" dado a_i y b_j.

La siguiente figura muestra un diagrama que representa de manera general el proceso de determinación de valores sustanciales de un cluster de cualquier dimensión.



Los estados dominantes se determinan por las secuencias de valores sustanciales que se generaron durante el proceso.

Las cuatro formas posibles de estados dominantes que pueden existir, dependiendo de la existencia de valores sustanciales en las dimensiones A, B y C se muestran en la siguiente tabla. El símbolo "*" en la tabla indica valores arbitrarios en la dimensión libre respectiva.

Estado dominante	Situación
$a_i \rightarrow (* , *)$	Existe un valor sustancial "a_i" en la dimensión "A". No existen valores sustanciales en la dimensión "B" ni en la "C".
$a_i \rightarrow b_j \rightarrow *$	Existe un valor sustancial "b_j" de la dimensión "B" dado el valor sustancial "a_i" de la dimensión "A". No existen valores sustanciales en la dimensión "C".
$a_i \rightarrow b_j \rightarrow c_k$	Existe un valor sustancial "c_k" de la dimensión "C" dado el valor sustancial "a_i" de la dimensión "A" y "b_j" de la dimensión "B".

Al final del proceso, se obtiene una tabla con los estados dominantes o modelos estructurales de cada cluster junto con la clase de comportamiento a la que pertenece. En una tabla de estados dominantes los clusters dentro de una clase de comportamiento tienen formas casi idénticas de modelos estructurales. El modelo estructural de un cluster es un resumen compacto de todos los flujos que lo constituyen y revela sólo la información esencial del cluster.

4. Indicadores de utilización de servicios

Se escogieron los clusters relacionados con la entrega o recepción de un servicio de red según la dimensión fija y la clase de comportamiento asociados a cada uno de los modelos estructurales. Se identifica, se filtra y se contabiliza aquellos flujos relevantes del tráfico que están relacionados con transacciones de tipo cliente-servidor.

La mayoría de clusters significativos presentan 3 perfiles de comportamiento canónicos según las clases de comportamiento a los que pertenecen, las características temporales de clusters individuales, sus estados dominantes y el tamaño promedio de un flujo en términos de cantidad de bytes y de paquetes y sus variaciones.

Los 3 perfiles de comportamiento canónicos se muestran en la siguiente tabla. Se utilizan únicamente las 8 clases de comportamiento que tienen el perfil canónico de servidores o servicios y se utilizan las métricas relacionados a estas clases como indicadores de utilización de servicios.

Perfil canónico	Dimensión	Clases de comportamiento	Ejemplos
Servidores o Servicios	IPORIGEN	$BC_{6,7,8}$	Web, DNS, email.
	IPDESTINO	$BC_{18,19,20}$	
	PORIGEN	BC_{23}	Tráfico de servicio agregado.
	PDESTINO	BC_{25}	
Host predominantes	IPORIGEN	$BC_{18,19}$	Rastreadores, proxies web y NAT boxes.
	IPDESTINO	$BC_{6,7}$	
Escaneos o exploits	IPORIGEN	$BC_{2,20}$	Exploits y escaners.
	IPDESTINO	$BC_{2,8}$	Blancos de escaneos.
	PDESTINO	$BC_{2,5,20,23}$	Tráfico de exploits agregados.

Según la referencia [1], los clusters o grupos de flujos de cada clase de comportamiento tienen ciertas similitudes y diferencias entre sí. Por tal motivo, se definen indicadores que son utilizados para distinguir subclases dentro de una clase de comportamiento. Para ello, se considera el tamaño promedio de los clusters tanto en bytes como en cantidad de paquetes y su variabilidad.

Cada flujo de un cluster se denota por f_i tal que $1 \leq i \leq m$, donde el valor de "m" el número total de flujos en el cluster. Sea PKT_i el número total de paquetes y BT_i el número total de bytes dentro del flujo f_i. Se definen 4 métricas para la diferenciación de los clusters dentro de las clases de comportamiento.

Métrica	Definición matemática
Número promedio de paquetes	$\mu(PKT)=\sum_{i=1}^{i=m}\frac{PKT_i}{m}$
Número promedio de bytes	$\mu(BT)=\sum_{i=1}^{i=m}\frac{BT_i}{m}$
Coeficiente de variación de paquetes	$CV(PKT)=\frac{\sigma(PKT)}{\mu(PKT)}$ Donde: $\sigma(PKT)=\sqrt{\frac{\sum_{i=1}^{i=m}(PKT_i-\mu(PKT))^2}{m}}$ $\sigma(PKT)$: Desviación estándar del número de paquetes.
Coeficiente de variación de bytes	$CV(BT)=\frac{\sigma(BT)}{\mu(BT)}$ Donde: $\sigma(BT)=\sqrt{\frac{\sum_{i=1}^{i=m}(BT_i-\mu(BT))^2}{m}}$ $\sigma(BT)$: Desviación estándar del número de bytes.

La popularidad de una clase de comportamiento indica la frecuencia de aparición de la clase de comportamiento. Un valor alto (cercano a 1) indica que se están observando clusters de la clase la mayor parte del tiempo.

El tamaño promedio de una clase de comportamiento indica la cantidad promedio de clusters que pertenecen a la clase. Un valor alto indica que aparecen un mayor número de clusters pertenecientes a la clase en los intervalos de tiempo donde aparece la clase.

La volatilidad de una clase de comportamiento indica la tendencia de la clase a contener diferentes clusters todo el

tiempo. Un valor alto (cercano a 1) indica que aparecen nuevos clusters de la clase la mayor parte del tiempo. Un valor bajo (cercano a 0) indica que aparecen los mismos clusters de la clase la mayor parte del tiempo.

El número promedio de paquetes/bytes de un cluster específico indica la cantidad promedio de paquetes/bytes en cada uno de los flujos del cluster. Un valor alto indica que fluye mucha información en la dirección indicada por los flujos del cluster.

El coeficiente de variación de paquetes/bytes de un cluster específico indica la dispersión del número de paquetes/bytes respecto a su valor promedio. Un valor alto indica que la cantidad de paquetes en los flujos del cluster son diferentes y varían demasiado. Un valor bajo indica que las cantidades de paquetes se asemejan y están cercanas a su valor promedio.

Se observa simetría entre las clases IPORIGEN BC_6, BC_7 y BC_8 con las clases IPDESTINO BC_{18}, BC_{19} y BC_{20} respectivamente y la clase PORIGEN BC_{23} con la clase PDESTINO BC_{25}. Debido a esta simetría se toma en consideración únicamente a los clusters de las clases IPORIGEN BC_6, BC_7 y BC_8 y PORIGEN BC_{23} con las cuales se obtienen los indicadores de utilización de servicios previamente mencionados.

Cada cluster de las clases IPORIGEN BC_6, BC_7 y BC_8 representa el comportamiento de un servidor comunicándose con pocos, regular número y muchos usuarios clientes respectivamente. La dirección IP del servidor es identificada por la dirección IP que identifica al cluster de la clase.

Cada cluster de las clases PORIGEN BC_{23}, representa el comportamiento agregado de un número regular de servidores alojando el mismo servicio (mismo número de puerto) y comunicándose con un número mucho mayor de clientes. El servicio específico es identificado por el puerto origen que identifica al cluster de la clase.

5. Desarrollo del sistema

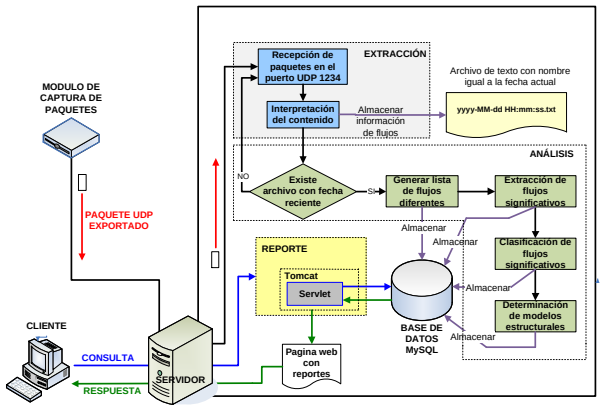
El sistema se implementó en un servidor en el cual se almacenan 3 proyectos principales escritos en lenguaje JAVA los cuales son:

Extracción: Implementa el proceso de recolección de registros de información de flujos. Se encarga de recibir los datos provenientes de un módulo de captura de paquetes y almacenarlos en archivos de texto simples.

Análisis: Implementa el proceso de modelamiento estructural basado en la teoría de la información. Contiene los subprocesos de extracción de flujos significativos, clasificación de los flujos significativos en clases de comportamiento y determinación de estados dominantes.

Reporte: Implementa la interfaz de visualización de reportes que contienen la información de indicadores de producción de servicios.
Los programas de estos proyectos interactúan con una base de datos MySQL, la cual se encarga de almacenar los resultados del proceso de modelamiento estructural basado en la teoría de la información. Luego, estos resultados son utilizados por la interfaz de reportes la cual realiza un procesamiento online para mostrar los resultados en una interfaz web según la fecha ingresada por el usuario.

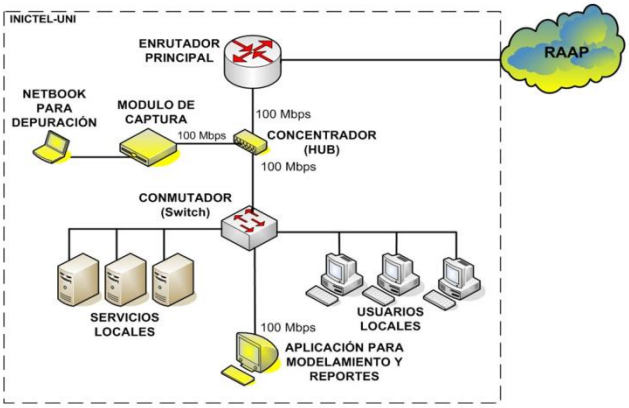
La siguiente figura muestra un esquema conceptual del sistema donde se muestran los procesos implementados en el servidor y el flujo de la información desde el módulo de captura de paquetes hasta la página web de reportes.



La siguiente figura es un extracto de la interfaz web de reportes, en la cual, el usuario debe ingresar la fecha inicial y final del análisis. Asimismo, debe escoger el tipo de reporte que desea observar y el tipo de servicio según el número de puerto de la aplicación o servicio que se desea analizar. Igualmente, se debe indicar si se desean analizar servicios que están dentro de la red de pruebas como fuera de ésta.

6. Resultados y discusión

El sistema fue probado en un nodo troncal de la Red Académica Peruana ubicado en el INICTEL-UNI con un tráfico máximo de 2 Mbps. Se utilizó un concentrador o hub para derivar el tráfico, un módulo de captura de paquetes y el servidor en una ubicación remota. El módulo envía paquetes UDP con registros de información de flujos hacia el servidor. La siguiente figura muestra la ubicación de dichos componentes en la red de pruebas.



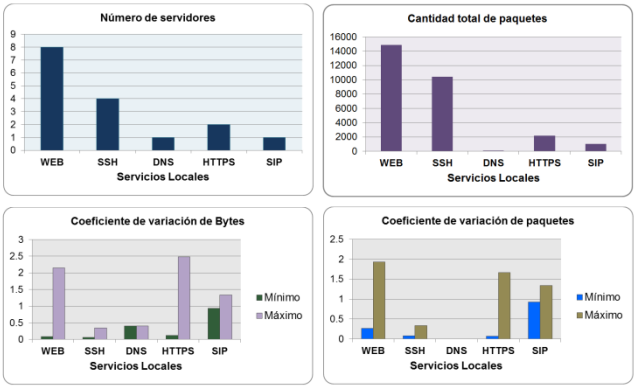
Las tablas siguientes muestran los resultados de una semana de análisis de servicios web locales correspondiente a la dimensión IPORIGEN BC_6, es decir, el comportamiento de un servidor comunicándose con pocos clientes. Se muestran los modelos estructurales y los clientes potenciales del servidor web local que presenta la mayor cantidad de bytes. Se observa además las 8 direcciones IP de los servidores web más relevantes de la red así como los indicadores de cada uno de éstos. La última tabla muestra un resumen del servicio web en general.

IP DEL SERVIDOR	MODELO ESTRUCTURAL	TOTAL FLUJOS	TOTAL BYTES	PROMEDIO BYTES	CV_BT mínimo	CV_BT máximo	TOTAL PAQUETES	PROMEDIO PAQUETES	CV_PKT mínimo	CV_PKT máximo	Frecuencia
IP LOCAL 1	[origen(80) -> (*, *)]	64	1795628	28056,69	1,28	2,15	1520	23,75	1,09	1,93	3
IP LOCAL 1	[origen(80) -> ip-destino (IP EXTERNA 1) -> *]	4	502720	125680	1,73	1,73	351	87,75	1,65	1,65	1
IP LOCAL 1	[origen(80) -> ip-destino (IP EXTERNA 2) -> *]	6	27760	4626,67	1,18	1,18	42	7	0,46	0,46	1
IP LOCAL 1	[origen(80) -> ip-destino (IP EXTERNA 3) -> *]	13	293427	22494,38	0,77	0,77	239	18,38	0,57	0,57	1

ORDEN	IP DEL SERVIDOR	NOMBRE DE DOMINIO	TOTAL FLUJOS	TOTAL BYTES	PROMEDIO BYTES	CV_BT mínimo	CV_BT máximo	TOTAL PAQUETES	PROMEDIO PAQUETES	CV_PKT mínimo	CV_PKT máximo	FRECUENCIA	CANTIDAD DE MODELOS ESTRUCTURALES
1	IP LOCAL 1		420	148082271	35257,79	0,37	2,15	12254	29,18	0,27	1,93	23	21
2	IP LOCAL 2		97	1068906	11019,65	0,3	2,01	2362	24,35	0,56	1,7	2	2
3	IP LOCAL 3		12	5801	483,42	0,15	0,29	58	4,83	0,29	0,45	2	2
4	IP LOCAL 4		6	5532	922	0,09	0,09	30	5	0,31	0,31	1	1
5	IP LOCAL 5		6	5017	836,17	0,18	0,18	30	5	0,37	0,37	1	1
6	IP LOCAL 6		6	4866	811	0,14	0,14	30	5	0,45	0,45	1	1
7	IP LOCAL 7		6	4500	750	0,15	0,15	30	5	0,45	0,45	1	1
8	IP LOCAL 8		6	3552	592	0,2	0,2	30	5	0,45	0,45	1	1

CLASE DE COMPORTAMIENTO	SERVICIO	TOTAL SERVIDORES	TOTAL FLUJOS	TOTAL BYTES	PROMEDIO BYTES	CV_BT mínimo	CV_BT máximo	TOTAL PAQUETES	PROMEDIO PAQUETES	CV_PKT mínimo	CV_PKT máximo	FRECUENCIA	CANTIDAD DE MODELOS
IPORIGEN BC6	WEB	8	559	15966445	28455,18	0,09	2,15	14824	26,52	0,27	1,93	32	30

Las siguientes figuras muestran comparaciones en la cantidad de servidores de cada servicio, cantidad total de paquetes y los coeficientes de variación de paquetes y de bytes de los servicios más relevantes extraídos de todo el tráfico. Los servicios web HTTP y HTTPS son los servicios más preferidos y también los que presentan mayores variaciones en la cantidad de paquetes y de bytes.



En las pruebas realizadas no se observaron paquetes relevantes dentro de otra clase de comportamiento, por lo que se puede concluir que en el periodo de pruebas la red presentaba poca demanda de los servicios mostrados en las figuras.

7. Conclusiones

Las redes actuales presentan una cantidad de tráfico tanto relevante como irrelevante. El filtrado de tráfico relevante según técnicas estadísticas permite simplificar el análisis del tráfico de servicios manteniendo una exactitud razonable. Con la herramienta diseñada se podrían generar otros indicadores más precisos utilizando los aquí mostrados. Esto permitiría hacer comparaciones entre varios tipos de servicios sin importar la naturaleza del mismo y poder determinar con una mayor exactitud cuál es el que rinde mejor y cuál es el que tiene una mayor aceptación por parte de los clientes.

Referencias

[1] Profiling Internet Backbone Traffic: Behavior Models and Applications. Kuai Xu, Zhi-Li Zhang, Supratik Bhattacharyya.
[2] Information Theory Structural Models for Qualitative Data. Klaus Krippendorff.
[3] Elements of Information Theory. Thomas M. Cover, Joy A. Thomas.
[4] Requirements for IP Flow Information Export (IPFIX). Network Working Group. RFC 3917.
[5] R. Cavallo and G. Klir, “Reconstructability analysis of multi-dimensional relations: A theoretical basis for computer-aided determination of aceptable systems models”, International Journal of General Systems, vol. 5, pp. 143-171, 1979.