Determinación de Perfiles de Tráfico de Nodos de Red usando Clustering

Santiago Pérez, Higinio Facchini, G. Mercado, Luis Bisaro
Grupo de Investigación y Desarrollo en TICs (GRID TICs)
Universidad Tecnológica Nacional, Facultad Regional Mendoza
santiagocp@frm.utn.edu.ar

Resumen. La comunidad dedicada al análisis de tráfico de redes LAN destina constantemente un gran esfuerzo para incorporar nuevas metodologías para facilitar y acelerar las especificaciones y dimensionamiento de los dispositivos activos y pasivos de nuevas redes, asegurando la mejor relación costo-performance y estabilidad de la infraestructura. Un punto de partida es obtener conocimiento desde la colección de muestras del tráfico de diferentes nodos. Esto ayuda a comprender el comportamiento de cada nodo de red, y a dimensionar su tráfico asociado. Para una red grande, este tiempo de ingeniería se vuelve apreciable, o simplemente imposible de costear por su volumen. Es deseable reducir las horas de ingeniería de la etapa de diseño o rediseño, y acotar los errores que se producen por su omisión y/o simplificación. En este trabajo, se propone la determinación de tráficos característicos de redes LAN usando clustering, como recurso para orientar y/o sistematizar en las especificaciones de las nuevas redes o rediseño de las existentes.

Palabras claves: análisis de tráfico, Ethernet, Clustering

1 Introducción

Hay un interés creciente en entender la conducta de los nodos de red (usuarios y dispositivos activos), y extender tal entendimiento para mejorar la performance y productividad de las redes. En diversos trabajos se ha planteado parcialmente el problema, con la recolección y estudio de trazas, para encontrar tendencias comunes de grupos de nodos de red, caracterizando cada nodo de red por la conducta común mostrada en sus patrones de asociación a través de los días. Estos trabajos se han abocado a proponer y medir índices o parámetros para caracterizar su comportamiento. Sin embargo, dichos estudios no están dirigidos a un planteo integral, excluyendo las similitudes de comportamiento a nivel de los protocolos de red ó de longitud de los paquetes.
El entendimiento que se desarrolle sobre la conducta de los nodos de red, a partir del estudio profundo de la trazas (incluyendo los protocolos involucrados) en LANs, puede ser aplicado para generar un modelo de patrones de tráfico característicos LAN realista, basado en la similitud de los comportamientos de subgrupos de nodos de red.

Dicha modelación podrá ser usado en diversas aplicaciones y líneas de estudio, y profundizarse en diversos temas de investigación.
En este trabajo, el modelo resultante será aplicado para proponer una nueva metodología sistemática de relevamiento, para reducir las horas de ingeniería de la etapa de diseño o rediseño, acotando sensiblemente los errores que se producen por su omisión y/o simplificación. Justamente, la fase de relevamiento en el diseño o rediseño puede acotarse con una metodología sistemática, basada en el modelo de patrones de tráfico característicos de la red LAN, que simplifique el relevamiento, mejore la estimación, reduzca el sobredimensionamiento, y asegure la estabilidad de la red en costo, tiempo de respuesta y atraso, probabilidad de bloqueo, escalabilidad, administración, performance, confiabilidad y productividad.

2 Estado de Situación de la Ingeniería de Diseño de Redes

Se están introduciendo aceleradamente tecnologías y métodos avanzados para la optimización en los procesos de diseño y rediseño de redes LAN y WAN, que llevan al replanteo metodológico de la orientación actual de la Ingenieria de Redes de Datos. Los problemas relacionados al diseño de redes LAN de envergadura, y sus accesos WAN crecen en complejidad proporcionalmente al número de nodos, a su variedad y a los requerimientos de servicios que demanden [1][2][3].
Por ello, es un objetivo constante de investigación la búsqueda de nuevas metodologías en las etapas iniciales de diseño, para una efectiva aproximación para la puesta en marcha de redes de una manera estable, con la mayor calidad de servicios posible y una reducción de sus costos asociados [4][5][6][7].
Específicamente, una de las tareas mas complejas en la fase de diseño (o rediseño) de una gran red LAN (más de 200 puestos de trabajo) y WAN, es el relevamiento y el dimensionamiento del sistema, en función de todos los parámetros y las variables involucradas, entre redes. El relevamiento de la red no es una tarea sencilla, y requiere de una metodología que incluya todos los parámetros que deseamos medir para llegar a un buen diagnóstico de los requisitos necesarios para un buen funcionamiento de la misma.
Es por ello, que aunque las técnicas de diseño conocidas incluyen una secuencia de tareas perfectamente establecidas [8][9][10], la tarea de relevamiento es habitualmente omitida o simplificada.

3 Comportamiento de los Nodos de Red desde la Colección de Trazas

Las grandes Redes LAN (cableadas y/o wireless) están ampliamente instaladas, y la combinación de ambas propuestas ha ganado rápidamente popularidad. Además, una parte de lo nodos de red han conmutado a nodos wireless.
En este escenario, la importancia de entender la conducta de los nodos es importante. La colección de trazas de Redes LAN, en la comunidad de investigación, ha sido la técnica habitual para obtener conocimiento fundamental de sus nodos. La colección

de trazas, y la posterior modelación de la red, son importantes porque: 1) El análisis de la conducta del usuario y los patrones de uso habilitan el examen exacto de la utilización de la red, y ayuda en el desarrollo de mejores técnicas de entendimiento y mejor capacidad de decisiones planeadas, 2) El análisis de trazas es también un primer paso necesario para hacer modelos realistas que son cruciales para el diseño, simulación y evaluación de protocolos de redes, 3) Cada nueva tecnología evoluciona, entendiendo fundamentalmente la conducta de los nodos de red, y se vuelve esencial para el desarrollo exitoso de tales tecnologías emergentes.
Nuestra propuesta del estudio de las trazas, puntualmente las tramas de red, considera el análisis de patrones de comportamiento similares de los nodos (patrones característicos de tráfico), definiendo el menor número posible de subgrupos (perfiles de nodos de red). Los subgrupos (salvando la cantidad de miembros), se replicarán con bastante aproximación para la mayoría de ellos en diversas redes LAN, independientemente de la organización, y por lo tanto, disminuirán las horas de ingeniería de la etapa de relevamiento del diseño y/o rediseño de toda nueva red. Es decir, una organización distinta tendrá subgrupos bastante similares (ajustando la cardinalidad de sus miembros), tal que el modelo pueda ser utilizado para establecer, y simular el tráfico, y por lo tanto, para aproximar más apropiadamente, y en un menor tiempo, el dimensionamiento del equipamiento activo y pasivo de la nueva red.

4 Trabajos relacionados

Hay bastantes trabajos de investigación dirigidos al objetivo general de entender la conducta de los nodos de red, con focos diferentes y aplicaciones potenciales en mente. En estos estudios, por ejemplo, el objetivo ha sido la estadística total (agregada) de eventos de movilidad, cantidad total de handoff, o eventos de re-asociación, longitud de sesión promedio, tipos de protocolos y ancho de bandas demandados, latencias, pérdida de paquetes, etc., o identificando las propiedades de usuarios individuales separadamente, por ejemplo, las locaciones home de los usuarios (es decir, el AP con que el usuario está asociado la mayoría de su tiempo online), o estudio de los protocolos más demandados [11][12]. U otros dirigidos a evaluar los retardos de end-to-end en las redes wireless para ciertos servicios como telnet (sobre TCP) y NFS (sobre UDP), planteando comparaciones con las redes cableadas, y proponiendo optimizaciones [13].
Por contraste, en otros trabajos, el objetivo puede verse más general, como un problema multivariable, identificando los patrones en asociación de usuarios y grupos similares de usuarios. Unos pocos estudios tocan el tema de la predicción de las tendencias comunes en patrones de asociación de usuarios individuales [14]. Sin embargo, no son muchos los trabajos que focalizan el estudio de la conducta sobre la estructura de los patrones de asociación de los nodos de red basada sobre colección de tramas extensivas de redes reales, y que involucren aspectos como los protocolos.

5 Caso de Estudio Experimental de Tráfico Ethernet

5.1 Colección de trazas en red universitaria y análisis con WEKA

Se utilizó el programa Wireshark [15] (ex Ethereal). Es el sniffer más usado con el que se pueden analizar 480 protocolos distintos. Tiene una interfaz flexible con opciones muy ricas de filtrado.
Con Wireshark, se tomaron 7 muestras de 3 minutos cada una, cada media hora, sobre un switch 3COM, usando un puerto configurado como monitor, para capturar todo el tráfico que pasaba a través de todos los puertos, con el detalle dado en la Tabla nº 1.

Tabla 1 Muestras de Tráfico

Numero Muestra	Horario de toma	Numero de instancias
1	17:30 hs	18439
2	18:00 hs	42647
3	18:30 hs	20407
4	19:00 hs	30189
5	19:30 hs	36139
6	20:00 hs	22754
7	20:30 hs	26657

Los atributos originales disponibles para cada instancia son: tiempo desde el inicio de la muestra, hora real, desplazamiento de tiempo, número IP origen, número IP destino, número de puerto origen y destino, y longitud de trama ó paquete.
Para los análisis de agrupamientos se usó la herramienta WEKA (Figura 1), que es una extensa colección de algoritmos de máquinas de conocimiento desarrollados por la universidad de Waikato (Nueva Zelanda) implementados en Java. WEKA contiene las herramientas necesarias para realizar transformaciones sobre los datos, y tareas de clasificación, regresión, clustering, asociación y visualización [16].

Figura 1 WEKA

5.3 Atributos Objetivo y Análisis de clustering sobre los atributos Longitud y Distancia de Protocolo

Atento a la diversidad de información que puede obtenerse desde las muestras, se hizo foco sobre los siguientes atributos, generados desde los originales: Delta Time, Delta Time Acumulado, Número de Red Origen, Número de Nodo Origen, Número de Red Destino, Protocolo, Distancia de Protocolo y Longitud. Estos atributos objetivo, se analizaron preliminarmente sobre todas las instancias de las distintas muestras tal como fueron obtenidas, salvando una leve depuración previa que significó el descarte de menos de 0,5% del tráfico.
Para el nuevo atributo Distancia de Protocolos se usó la siguiente asignación cuantificada según el atributo Protocolo: TCP = 1, UDP = 9, ICMP = 5, HTTP = 2, SNMP = 8, y SMTP = 10. El protocolo SMTP no aparece en las muestras filtradas, dado que era uno de los protocolos que aparecía con muy poco tráfico. Se usó el algoritmo FarthestFirst para obtener las instancias agrupadas y los centroides de los clusters para cada muestra, usando 5 clusters. La Tabla 2 resumen los resultados.

Tabla 2 Clustering Atributo Longitud y Protocolo

Numero Muestra		Muestra 1 17:30	Muestra 2 18:00	Muestra 3 18:30	Muestra 4 19:00	Muestra 5 19:30	Muestra 6 20:00	Muestra 7 20:30
Numero de Clusters		5	5	5	5	5	5	5
Cluster 0	Porcentaje	20%	10%	18%	1%	15%	21%	17%
	Centroide Longitud Protocolo	301 SNMP UDP	60 SNMP UDP	432 SNMP UDP	666 SNMP UDP	430 SNMP UDP	60 SNMP UDP	74 SNMP UDP
Cluster 1	Porcentaje	11%	7%	6%	7%	6%	20%	8%
	Centroide Longitud Protocolo	1514 HTTP TCP	1514 HTTP TCP	1514 HTTP TCP	1514 HTTP TCP	1514 HTTP TCP	1514 HTTP TCP	1514 HTTP TCP
Cluster 2	Porcentaje	57%	57%	63%	71%	68%	47%	64%
	Centroide Longitud Protocolo	66 HTTP TCP	66 HTTP TCP ICMP	66 HTTP TCP	66 HTTP TCP ICMP	64 HTTP TCP	60 HTTP TCP	60 HTTP TCP
Cluster 3	Porcentaje	7%	26%	7%	6%	8%	5%	7%
	Centroide Longitud Protocolo	780 HTTP TC	792 HTTP TCP	780 HTTP TCP	801 HTTP TCP	789 HTTP TCP	747 HTTP TCP	775 HTTP TCP
Cluster 4	Porcentaje	5%	0%	7%	16%	4%	5%	5%
	Centroide Longitud Protocolo	86 ICMP	1034 UDP	152 ICMP	85 UDP SNMP ICMP	98 ICMP	102 ICMP	98 ICMP

Se observa que la clusterización resultante entre la Longitud y los protocolos se mantiene bastante estable independientemente de la muestra. Además, que:

- El cluster 2 reúne, para las distintas muestras un 62% promedio de instancias, con un Centroide de Longitud ubicado entre 60 y 66, y de los protocolos HTTP y TCP (salvando para las muestras de las 18:00 y 19:00 hs que incluyen ICMP).
- El cluster 1 reúne, para las distintas muestras un 9% promedio de instancias, con un Centroide de Longitud de 1514, y de los protocolos HTTP y TCP.
- El cluster 3 reúne, para las distintas muestras un 9% promedio de instancias, con un Centroide de Longitud ubicado entre 747 y 801, y de los protocolos HTTP y TCP.
- El cluster 4 reúne, para las distintas muestras un 6% promedio de instancias, con un Centroide de Longitud ubicado entre 86 y 152, y del protocolo ICMP (salvando para las muestras de las 18:00 y 19:00 hs donde ICMP es parte de otros clusters).
- El cluster 0 se muestra más diverso en porcentaje de instancias para las distintas muestras, con un Centroide de protocolos de SNMP y UDP, y de Longitud ubicado entre 60 y 666.

En resumen: 1) Los protocolos HTTP y TCP barren todas las longitudes permitidas de paquetes, 2) El protocolo ICMP se concentra en los paquetes de longitud mínima, y 3) Los protocolos UDP y SNMP en la parte baja de las longitudes permitidas.
La figuras 2 y 3 presentan como ejemplo la clusterización y visualización del atributo Longitud y Protocol para la muestra 1 de las 17:30 hs.

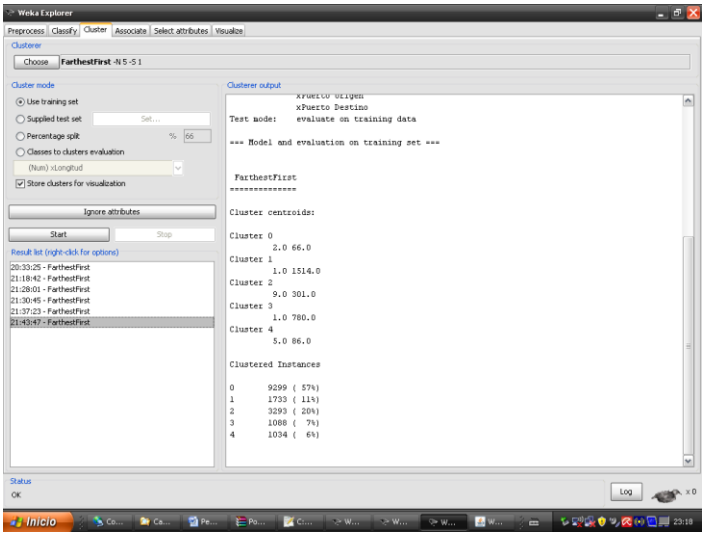

Figura 2 Clusterización usando FarthestFirst sobre el Atributo Longitud y Protocol muestra 17:30 hs

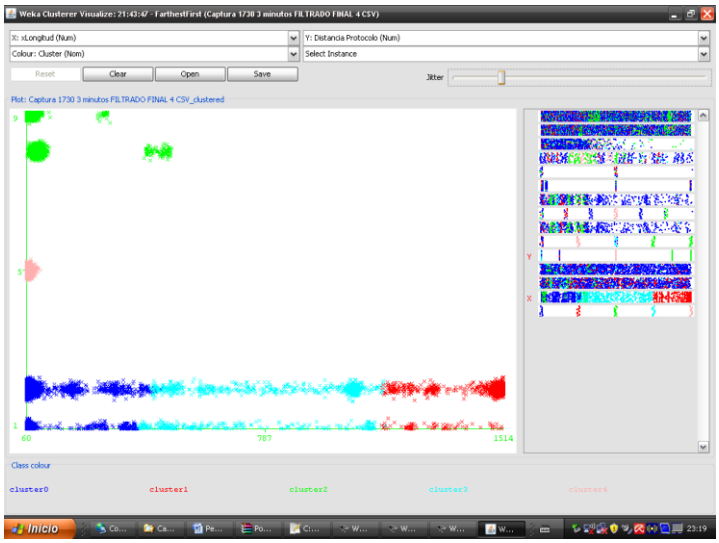

Figura 3 Visualización usando FarthestFirst sobre el Atributo Longitud y Protocol muestra 17:30 hs

5.3 Análisis de clustering sobre los atributos Delta Time Acumulado, Número de Nodo Origen, Distancia de Protocolo y Longitud para la red origen 192.168.23

La muestra de las 17:30 hs de la red origen 192.168.23 tenía 6910 instancias distribuidas entre los 18 nodos activos, como indica la Tabla 3 de la siguiente manera:

Tabla 3 Instancias por Nodo Origen

Número De Orden	Numero de Nodo IP	Número De Instancias	Máximo Delta Time Acumulado	Protocolos
1	30	83	150.837618	TCP
2	31	15	144.977294	HTTP,TCP
3	33	534	178.229240	HTTP,TCP
4	165	551	181.012942	UDP,ICMP
5	169	1220	181.438292	HTTP,TCP,ICMP
6	170	188	181.009561	UDP,ICMP
7	181	4	127.507899	ICMP
8	182	954	181.205839	HTTP,TCP
9	184	104	135.010842	HTTP,TCP
10	187	25	136.724155	HTTP,TCP,ICMP
11	199	381	150.529362	HTTP,TCP,UDP
12	200	12	75,422951	HTTP,TCP
13	209	240	181.011104	HTTP,TCP
14	212	911	179.919663	HTTP,TCP
15	229	551	181.018642	UDP,ICMP
16	243	37	160.489615	HTTP,TCP,UDP
17	244	550	181.011829	UDP,ICMP
18	248	550	181.021322	UDP,ICMP

Se usó el algoritmo EM para obtener un número de instancias agrupadas y los centroides de los cluster para la muestra. La configuración default dio una segmentación sobre 5 clusters. La Tabla 4 resumen los resultados. Se observa que:
- El cluster 3 reúne un 52% de instancias, con un Centroide de Longitud ubicado en 61, de los protocolos HTTP y TCP, de Nodo Origen 217 y de Delta Time Acumulado de 77.9854. Se observa que el cluster recupera las características más frecuentes, como son: la Longitud mínima, los protocolos más usados HTTP, TCP e ICMP, sobre un nodo intermedio de tráfico de paquetes e intermedio Delta Time Acumulado.
- El cluster 1 reúne un 29% de instancias, con un Centroide de Longitud ubicado en 67, de los protocolos UDP e ICMP, de Nodo Origen 189 y de Delta Time Acumulado de 89.4017. Se observa que el cluster recupera características similares para una longitud mínima un poco mayor, para los protocolos UDP e ICMP, sobre un nodo intermedio de tráfico de paquetes e intermedio Delta Time Acumulado.

Tabla 4 Clustering Atributo Longitud, Protocol, Delta Time Acumulado y Nodo Origen muestra 17:30 hs

Numero Muestra		Muestra 1 17:30
Numero de Clusters		5
Cluster 0	Porcentaje	11%
	Centroide Longitud Protocolo Nodo Origen Delta Time Acumulado	814 HTTP,TCP 147 76.6132
Cluster 1	Porcentaje	29%
	Centroide Longitud Protocolo Nodo Origen Delta Time Acumulado	67 UDP,ICMP 217 89.4017
Cluster 2	Porcentaje	6%
	Centroide Longitud Protocolo Nodo Origen Delta Time Acumulado	61 HTTP,TCP 32 90.3359
Cluster 3	Porcentaje	52%
	Centroide Longitud Protocolo Nodo Origen Delta Time Acumulado	61 HTTP, TCP 189 77.9854
Cluster 4	Porcentaje	1%
	Centroide Longitud Protocolo Nodo Origen Delta Time Acumulado	97 HTTP,TCP,ICMP 190 140.3976

La Figura 4 presenta como ejemplo la clusterización y visualización del atributo Longitud, Protocol, Delta Time Acumulado y Nodo Origen para la muestra.
Se observa que la segmentación de las instancias tiene una preponderancia organizativa desde las longitudes de los paquetes.
También se destaca que el tráfico del tipo TCP-HTTP, que es el mayor en la muestra, observa una distribución menos continua durante el periodo de muestreo de los Delta Time Acumulados (ver cluster 3 – Figura nº 4). Es decir, el tráfico UDP sería entonces menos intenso pero más uniformemente distribuido en el tiempo. Podría concluirse que el tráfico de ráfagas estaría más influido por TCP que por UDP. Tal afirmación requeriría nuevas muestras por un periodo más extenso

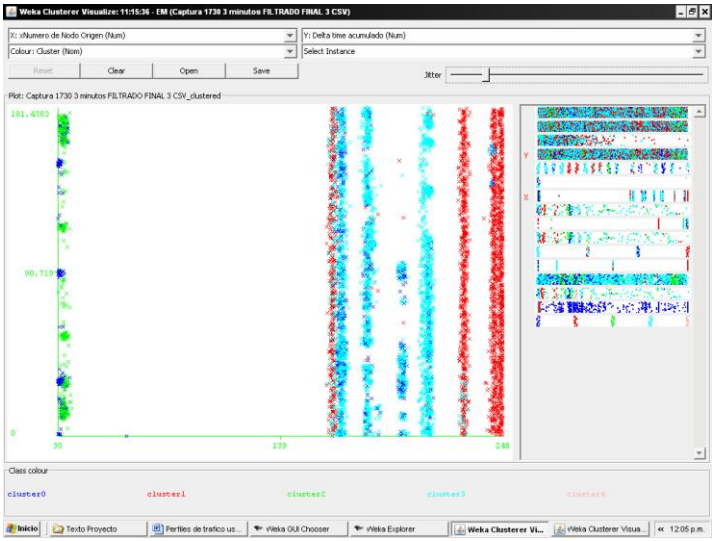

Figura 4 Visualización usando EM sobre el Atributo Longitud, Protocol. Delta Time Acumulado y Nodo Origen muestra 17:30 hs

6 Conclusiones

En este documento, se presentó una metodología de análisis del comportamiento de las redes desde la perspectiva de clustering, poniendo en evidencia la utilidad para la validación de aspectos conocidos por otras técnicas, y el descubrimiento de nuevos. A partir del método de colección de trazas (tramas) con el uso de sniffers, se generó una base de datos, sobre la cual se han aplicado técnicas de preprocesamiento.
A lo largo del trabajo se ha verificado que el tráfico de los diversos nodos de la red tienen perfiles característicos de comportamiento, con un alto grado de similitud. Entre las características más sencillas se destacan: a) Una proporción muy elevada de paquetes tiene una longitud casi mínima de 60 bytes; b) En el momento de muestreo la mayoría del tráfico (95%) se concentraba en 5 protocolos (HTTP; UDP; TCP, ICMP y SNMP), y c) El protocolo HTTP era el predominante (76,5%).

Usando la herramienta WEKA para el agrupamiento de la longitud y protocolos, pudo observarse que: 1) Los protocolos HTTP y TCP barren todas las longitudes permitidas de paquetes, 2) El protocolo ICMP se concentra en los paquetes de longitud mínima, y 3) Los protocolos UDP y SNMP en la parte baja de las longitudes permitidas.
El agrupamiento de los atributos Longitud, Protocolos, Nodo Origen y Delta Time Acumulado, presenta agrupamientos más diversos, coincidiendo en 5 perfiles o comportamientos de tráfico similares. Por ello, en el desarrollo del proyecto se reiteró el ensayo, para su comparación, con los algoritmos EM, FarthestFirst y SimpleKMeans. Se observa que los segmentos son diferentes, coincidiendo en general en los siguientes aspectos: 1) El tráfico UDP y de longitudes mínimas sería menos intenso, pero estaría más uniformemente distribuido en el tiempo, 2) El tráfico de ráfagas estaría mas influido por TCP que por UDP. Estos últimos comentarios, sugieren la necesidad de extender y profundizar los procedimientos seguidos, sobre nuevas muestras por un periodo más extenso, y en otras redes para contraste. Finalmente, las conclusiones podrían usarse en diversas aplicaciones, aunque el objetivo inmediato es proponer una nueva metodología sistemática de relevamiento, para reducir las horas de ingeniería de la etapa de diseño o rediseño de redes LAN. El modelo deberá contemplar la conducta destacadas a lo largo del trabajo.

Referencias

1. Schwartz, M: Report of the Internet Perspective Working Group, IITA Digital Libraries Workshop, 2005.
2. Clark, D.: Rethinking the design of the Internet: The end-to-end arguments vs the brave new world, MIT, agosto 2001.
3. Pérez, J.: Evolución y Tendencias del sector de las Telecomunicaciones. Universidad Politécnica de Madrid, 2005.
4. Hauger, S.: A scalable architecture for flexible high-speed packet classification, Universidad de Stuttgart, diciembre 2006.
5. Hu, G.: Analysis of the CSMA/CA Protocol en a new optical MAN network architecture, Universidad de Stuttgart, febrero 2004.
6. Balazinska, M.: Characterizing Mobility and network usage in a corporate wireless LAN, MIT Laboratory for Computer Science, Paul Castro, IBM, 2003.
7. Battiti, R.: Wireless LANs: From WarChalking to Open Access networks, Departamento de Informatica y telecomunicaciones, Universidad de Trento, febrero 2005.
8. http://www.cisco.com/univercd/cc/td/doc/cisintwk/idg4/index.htm CISCO PRESS
9. Bruno, A.: CCDA Certification Guide, Cap. 2, CISCO PRESS
10. Lammle, T., Baril, A.: Guia de Estudio CCDA, Certified Design Asóciate, 2º Edicion, San Francisco, London, Ed. Sybex, www.sybex.com
11. Lakhina, A., Papagiannaki, K., Crovella, M., Diot, C., Kolaczyk, D., Taft, N.:Structural Analysis of Network Trafficc Flows, ACM SIGMETRICS, New York, June 2004.
12. Chen, G., Huang, H., Kim, M.: Mining Frequent and Periodic Association Patterns, Dartmouth College Computer Science Technical Report TR2005-550, July 2005
13. Lai, K., Roussoponlos, M., Tang. D., Zhao, X., Baker, M.: Experiences with a Mobile Testbed, Stanford University, Proceedings of The Second International Conference on Worldwide Computing and its Applications (WWCA'98)
14. Merugu, S., Ammar, S.:Space-Time Routing in Wireless Networks with Predictable Mobility, Georgia Institute of Technology, Technical Report GIT-CC-04-07, Mar/04
15. http://www.wireshark.org
16. http://www.cs.waikato.ac.nz/ml/weka/