Clasificación de Tráfico de Redes para la Agrupación de Usuarios

Jorge E. RODRÍGUEZ
Universidad Distrital Francisco José de Caldas
Bogotá, Colombia

Diana C. MACHADO
Universidad Distrital Francisco José de Caldas
Bogotá, Colombia

Gina A. ALZATE
Universidad Distrital Francisco José de Caldas
Bogotá, Colombia

RESUMEN

En este artículo se presentan algunos algoritmos de agrupación de tráfico para un conjunto de datos de monitoreo de red. Primero, se realiza la descripción del conjunto de datos a evaluar y los trabajos relacionados en este aspecto. Segundo, se describe el algoritmo de gravitación seleccionado y finalmente se muestran las conclusiones con respecto a la selección, y trabajos futuros.

Palabras Claves: agrupación, tráfico en la red, agrupación gravitacional, algoritmo de agrupación.

1. INTRODUCCION

La clasificación del tráfico, es un elemento que establece importantes tareas en la administración de la red, tales como: priorización del flujo, que según [13] consiste en determinar la importancia de la aplicación según la obtención o utilización de los recursos; control del tráfico, (permite controlar la máxima tasa de transferencia dentro o fuera de la red), categorización de las aplicaciones, categorización del volumen de tráfico, planeación de la capacidad y aprovisionamiento del enrutamiento.

El enfoque tradicional para clasificar el tráfico en la red consiste en categorizar el número de puerto ó protocolo, con base en estos parámetros el operador de red puede asignar el ancho de banda adecuado para cada usuario. Este método es posible ya que inicialmente las aplicaciones P2P usaban números de puerto estáticos y protocolos tales como, HTTP y FTP de esta manera se diferenciaba el tipo de archivo transferido o la aplicación consumida. Sin embargo, luego aparecieron versiones de aplicaciones P2P como bitTorrent que ofrecían descargas rápidas de archivos de gran tamaño, por medio de un protocolo y puerto estático; dado que este tipo de aplicaciones fueron restringidas surgió en el mercado el enmascaramiento de puertos dificultando bloquear de manera directa aquellos puertos conocidos como fuera del enfoque del negocio.

Debido al enmascaramiento de puertos y protocolos esta técnica de detección de tráfico en la red comenzó a ser ineficaz, además de generar alto costo computacional y de almacenamiento.

Las limitaciones expuestas con anterioridad han motivado el uso de las características de la capa de transporte para la clasificación de tráfico. Por ejemplo, en [4] y [13], se muestra que el análisis de grupos tiene la habilidad de clasificar el tráfico de internet usando dicha información.

Teniendo en cuenta la agrupación de características necesaria para realizar la clasificación del tráfico se recurre a técnicas

de agrupamieno ofrecidas por algoritmos de aprendizaje computacional no supervisado.

La aplicación de dichas técnicas permite a cada usuario comprar planes según el ancho de banda consumido. Este es el algoritmo de Agrupación Gravitacional Aleatoria, el cual usa la misma teoría propuesta por [7], afianzando en este caso la utilización de una heurística para mejorar los resultados, tal como se describe en la sección V. Según los resultados obtenidos, a partir de la aplicación del algoritmo en el conjunto de datos ejemplificado, se determina que el algoritmo es apropiado para la agrupación de usuarios.

En este artículo se describen varios trabajos relacionados con la clasificación de tráfico de la red y los resultados obtenidos en cada caso, con el fin de tener un referente para la selección del algoritmo.

Por último, en la sección V se presenta el algoritmo seleccionado para posteriormente construir una herramienta de agrupación de usuarios de acuerdo a las características de la capa de transporte, dado que esta técnica es efectiva para clasificación del tráfico en la red.

2. DESCRIPCIÓN DEL PROBLEMA

Debido al crecimiento de internet y a la construcción de aplicaciones más robustas en términos de consumo de ancho de banda, las compañías donde existen uno o varios segmentos de LAN, con una red troncal formada normalmente por redes Ethernet o fibra, conexiones remotas para usuarios que trabajan desde la casa, conexiones con internet, entre otras, ha sido necesario administrar apropiadamente sus circuitos WAN en busca de encontrar una distribución más eficiente, en lugar de incrementar el tamaño de los circuitos.

Por esta razón, en el mercado existen herramientas con este propósito como el algoritmo de balanceo de carga de la red, el cual por su simplicidad y velocidad permite alcanzar una alta efectividad incluyendo una baja utilización del canal y un rápido tiempo de respuesta en un amplio número de aplicaciones cliente/servidor. Dicha efectividad puede ser medida en: sobrecarga del procesador, tiempo de respuesta, ocupación del canal, ocupación del switch; de manera que estas herramientas proveen una solución costo beneficio ideal para aumentar la escalabilidad y alta disponibilidad de las aplicaciones.

No obstante, las compañías ofrecen a los usuarios el mismo ancho de banda en planes con características idénticas, aún cuando el tráfico no sea equivalente entre usuarios, ya que tal como se describe en el algoritmo de balanceo de carga, una clasificación de los usuarios es necesaria para una distribución eficiente del tráfico en la red; si bien este proceso no es transparente para el usuario, es realizado por las compañías; es por esta razón que el objetivo de este proyecto es construir un prototipo de software que permita la creación de perfiles de consumo basados en la agrupación de usuarios según el tipo de aplicaciones y ancho de banda consumido.

3. TRABAJOS RELACIONADOS

Variedad de técnicas utilizan la información de la capa de transporte para resolver problemas asociados a la clasificación del tráfico en la red, en esta sección se hará un recorrido por las herramientas relacionadas.

En [16], se proponen técnicas de aprendizaje computacional para agrupar el flujo de datos presentado en un conjunto de datos por el algoritmo EM (Expectation - Maximization), este divide el flujo en grupos: el paso inicial es definir la constante del valor esperado, primero se asigna una densidad probabilística al conjunto de datos, luego en el proceso de maximización, los valores de densidad se estiman para los parámetros de cada modelo. Estos pasos son repetidos hasta conseguir un

máximo local que sea igual al máximo global definido.

Las pruebas con este algoritmo muestran que es apropiado para clasificar información con las mismas características generales, se desagrega la cabecera del paquete e identifica un tipo de tráfico particular.

En [5] se emplearon dos algoritmos de agrupamiento, llamados K-Means y DBSCAN para la clasificación de tráfico en estos algoritmos los grupos son determinados por una similitud definida, según la distancia Euclidiana, mientras que una gran distancia implica estar en un grupo diferente. K-Means inicialmente elige aleatoriamente los objetos dentro del conjunto de datos, los objetos son divididos en los grupos más cercanos, K-Means iterativamente calcula los nuevos centros, y luego los particiona nuevamente, basado en los nuevos centros hasta converger, con los conjuntos de datos probados la precisión superaban el 79%. DBSCAN, está basado en la densidad así como el algoritmo EM propuesto [12].

El concepto depende de dos parámetros épsilon (eps) y minPts, donde eps es la distancia alrededor del objeto que define el vecindario, y minPts es el valor mínimo para determinar el núcleo del objeto. El grupo es formado por cada densidad de objeto conectada basada en eps y minPts hasta asignar cada objeto en el conjunto de datos.

La precisión de este algoritmo es de 72%. Estos algoritmos son probados con aplicaciones basadas en el protocolo TCP, considerando las siguientes características de flujo estadístico: número total de paquetes, tamaño medio de los paquetes, es decir el tamaño de la carga útil excluyendo las cabeceras, bytes transferidos y tiempo medio entre llegadas de los paquetes.

Los métodos tradicionales de clasificación de tráfico de la red están basados en la identificación de puertos, sin embargo existen otras alternativas, en [4] se propone la clasificación de acuerdo al flujo de estadísticas, este enfoque es evaluado en las trazas de un conjunto de datos de un variado tipo de aplicaciones con el cual se obtiene una precisión cercana al 94%.

La clasificación consiste en dos pasos: agrupación, en la que se divide el conjunto de datos en grupos disjuntos, de acuerdo al flujo similar o diferente; dicho procedimiento es realizado con K-Means, luego estos grupos utilizan las etiquetas disponibles de acuerdo al flujo y obtiene un mapeo de los grupos. Este procedimiento se enfoca exclusivamente en la clasificación de tráfico TCP.

En este artículo, se evalúan algoritmos de agrupación aplicados en la clasificación del tráfico de red tales como EM, K-Means y DBSCAN, otros trabajos relacionados son mencionados en [1], [3], [6], [11] y [17].

4. ALGORITMO DE AGRUPACIÓN GRAVITACIONAL

La clasificación del tráfico de la red se puede abordar desde múltiples técnicas, por lo que se revisaron principalmente aquellas con relación al campo del problema planteado. Algunas de estas técnicas realizan el análisis de tráfico con datos tomados de los dispositivos de red, siendo esta la mayor diferencia con muchas de las propuestas revisadas.

La descripción presentada a continuación, se enfoca en aquellas propuestas más relevantes para el caso de estudio.

El algoritmo de gravitación universal propuesto en [7] halla automáticamente el número de grupos en un conjunto de datos. Cada objeto en el conjunto de datos es considerado un objeto en el espacio futuro y es movido usando la ley de la fuerza de gravitación universal y la segunda ley de newton.

Basado en las técnicas de agrupación propuestas en la teoría de minería de datos que según [15] es un proceso que permite extraer información útil de grandes conjuntos de datos o bases de datos, o según [2], [9] y [10] quienes exponen que la minería de datos, es un proceso con el cuál, partiendo de una serie de datos (generalmente grandes volúmenes) y haciendo uso de métodos estadísticos, de aprendizaje computacional, e inteligencia artificial entre otros, se encarga de descubrir conocimiento implícito en los datos.

Cada registro de datos en la fuente es considerado como un objeto en el espacio futuro y es movido por la utilización de la fuerza gravitacional y la segunda ley de newton. Esta propuesta se basa en el algoritmo de gravitación propuesto por [14], la ventaja principal sobre las técnicas expuestas son: velocidad, robustez y no supervisión.

Un punto de datos ejerce una fuerza superior en un punto de datos que no está en el mismo grupo, a continuación, los puntos son movidos en la dirección del centro del grupo, con esta técnica se determinaran los grupos en el conjunto de datos. Si algún punto no pertenece a algún grupo no será asignado a ningún grupo, de tal manera que se elimina el ruido del conjunto de datos agrupados.

Cada punto en el conjunto de datos es movido de acuerdo a una versión simplificada de la ecuación (1).

$$x(t+1) = x(t) + \frac{\vec{d}G}{\|\vec{d}\|^3} \qquad (1)$$

Donde, $\vec{d} = \vec{y} - \vec{x}$, y G la constante de gravitación considera la velocidad en cualquier espacio de tiempo, v(t), tal como el vector zero y Δ(t)=1. La distancia entre puntos y la constante de gravitación G es reducida, mientras los puntos de información son movidos a los grupos correspondientes, dependiendo el valor definido por la constante G.

GCA crea un conjunto de grupos usando una óptima estructura de separación del conjunto de datos y unión-correlación y la distancia entre objetos. RGC usa un parámetro extra (α) para determinar el mínimo de puntos de datos que un grupo debe incluir con el fin de ser considerado como un grupo válido. Examinando el tiempo de complejidad del algoritmo obtenido es O(N), y esto define que el tiempo de complejidad de la función GetClusters es generalmente delimitada por la ecuación O(#clusters) ≤ (N).

5. AGRUPACIÓN GRAVITACIONAL ALEATORIO

Agrupación gravitacional aleatorio (RGC), [8] es una generalización del algoritmo de agrupación gravitacional, para permitir las diferentes funciones de movimiento y asignar automáticamente la constante de gravitación.

En RAIN [8], se definen tres elementos importantes:

- Distancia máxima entre dos puntos cercanos, el cual se determina en la ecuación 2 de acuerdo a las tres posibles trayectorias para un punto de información.

$$\hat{d} = \frac{2 * \sqrt{n}}{\sqrt{3} * N^{1/n}} \qquad (2)$$

- Funciones de movimiento, define la posición final del punto de datos x que esta interactuando con otro punto de datos y.

$$x(t+1) = x(t) + G * \vec{d} * f\left(\frac{\|\vec{d}\|}{\hat{d}}\right) \qquad (3)$$

- Asignación de la constante de gravitación inicial G, en orden de definir los grupos necesarios, dicha constante es reducida en una proporción constante (ΔG).

6. CONCLUSIONES

En este artículo fueron revisados los algoritmos de agrupamiento: EM, K-Means, DBSCAN, RGC, GCA, para el problema de clasificación de tráfico en la red. El análisis se basa en elegir un algoritmo para producir los grupos necesarios con gran precisión.

El algoritmo seleccionado es RGC (Agrupación Gravitacional Aleatoria) debido a los resultados obtenidos en los conjuntos de datos probados expuestos, que según los autores son cercanos al 99%, lo cual nos permite precisar que dicho algoritmo es apropiado para la clasificación de consumo de ancho de banda, en un conjunto de datos determinado por el puerto y protocolo consumido por determinada aplicación.

7. TRABAJOS FUTUROS

Nuestro trabajo futuro consiste en la implementación del algoritmo RGC [7], en un conjunto de datos de información recolectada a partir de la información de los protocolos y puertos usados por una aplicación de internet.

Esta implementación es enfocada en generar los grupos correspondientes al consumo de ancho de banda, de manera que se permita a las compañías ofrecer un servicio o plan basado en las necesidades del usuario.

8. REFERENCIAS

[1] Bakarat, Thiran, Iannacone, Diot, Owezarski. *Modeling Internet Backbone Traffic at the Flow Level*. IEEE Transactions on Signal processing. (2003).

[2] Charkrabarti, S., *Data mining: Know it all* Know it all. 2009, Burlintong: Morgan Kufmann. p. 477.

[3] Corner T., Leiserson C, and Rivest, R Introduction to Algorithms. Mc Graw Hill, 1990.

[4] Erman J, Mahanti A, Arlitt M, Cohen I, Williamson C. Offline/Realtime *Traffic Classification Using Semi-Supervised Learning* Performance Evaluation, Volume 64, Issues 9-12, October 2007, Pages 1194-1213.

[5] Erman, Arlitt, Mahanti. *Traffic Classification Using Clustering Algorithms*. Proceedings of the 2006 SIGCOMN workshop on Mining network data. 2006.

[6] Erman, Arlitt, Mahanti. *Internet Traffic Identification using Machine Learning*. Global Telecommunications Conference, (2006).

[7] Gomez J., Dasgupta D., and Nasraoui O., *"Clustering Gravitational Algorithm"* in Proceedings of the Third SIAM International Conference on Data Mining 2003, 2003.

[8] Gomez J., Dasgupta D., and x Nasraoui D., *"RAIN, Data Clustering using Randomized Interactions between Data Points"* in Proceedings of the 2004 International Conference on Machine Learning and Applications, 2004.

[9] Hand D., Mannila H., and Smyth P., *Principles of data mining*. 2001, Cambridge: A Bradford Book The MIT Press. p. 546.

[10] Han, J and Kamber, M, *Data mining:Concepts and techniques* 2 ed. ed. Data management systems, ed. J. Gray. 2006, San Francisco, CA: Morgan Kufmann Publishers. p. 770.

[11] Karagiannis, Papagiannaki, Faloutsos. *BLINC: Multilevel Traffic Classification in the Dark*. Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications. *(2005).*

[12] McGregor, Mark, Lorier, Brunskil. *Flow Clustering Using Machine Learning* Techniques. Lecture Notes in Computer Science (2004).

[13] McCabe, J. (2007). *Network Analysis, Architecture, and Design*. (3th ed.). Morgan Kaufmann.

[14] Right W. E., Gravitational Clustering. Pattern Recognition, 9:151-166, Pergamon Press, 1977.

[15] Sivanandam, S N and Sumathi, S, *Introduction to data mining and its applications*. 2006, Berlin Heidelberg: Springer-Verlag Berlin Heidelberg. p. 835.

[16] Williams N, Zander S, Armitage G. *A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification*. Computer Communication Review. (2006).

[17] Zander S, Nguyen T, Armitage G. *Automated Traffic Classification and Application Identification using Machine Learning.* The IEEE Conference on Local Computer Networks 30th Anniversary. (2005).