INTERNATIONAL JOURNAL OF NETWORK MANAGEMENT
Int. J. Network Mgmt 2013; **23**: 101–118
Published online 28 December 2012 in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/nem.1818

NetCube: a comprehensive network traffic analysis model based on multidimensional OLAP data cube

Daihee Park[1] , Jaehak Yu[2] , Jun-Sang Park[1] and Myung-Sup Kim[1,*†]

[1]*Department of Computer and Information Science, Korea University, Sejong 339-700, Korea*
[2]*Electronics and Telecommunications Research Institute, Daejeon 305-700, Korea*

SUMMARY

Network traffic monitoring and analysis are essential for effective network operation and resource management. In particular, multidimensional analysis for long-term traffic data is necessary for comprehensive understanding of the traffic trend and effective quality-of-service provision considering the extremely dynamic behavior of the current Internet, where various types of traffic occur from high-speed network links and greatly increasing number of applications. However, only limited analysis results are provided, as the existing network traffic analysis tools and systems are developed and deployed focusing on their own specialized analysis purposes. Consequently, it is difficult to understand the network comprehensively and deeply, which increases the necessity for multilateral analysis of long-term traffic data. In this paper, we propose a novel traffic analysis model for large volumes of Internet traffic accumulated over a long period of time. The *NetCube*, the proposed network traffic analysis model using online analytical processing (OLAP) on a multidimensional data cube, provides an easy and fast way to construct a multidimensional traffic analysis system for comprehensive and detailed analysis of long-term traffic data by utilizing simple OLAP operations and powerful data-mining techniques on various abstraction levels of traffic data to complete the analysis purpose. We validate the feasibility and applicability of the proposed *NetCube* traffic analysis model by implementing a traffic analysis system and applying it to our campus network. Copyright © 2012 John Wiley & Sons, Ltd.

Received 29 February 2012; Revised 18 October 2012; Accepted 19 October 2012

1. INTRODUCTION

Internet usage has increased exponentially owing to a number of advances such as the convergence of telecommunication and data networks, the integration of wired and wireless networks, the development of various services and application programs and diverse user demands. Along with this increase in usage, there has also been a dramatic growth in Internet traffic. However, this rapid increase of network traffic causes a number of problems for stable network services because of network overload and bottlenecks. Therefore, the importance of network traffic monitoring and analysis has increased to gain a more exact understanding of the overall network conditions for stable network operation and effective resource management [1–6]. More significantly, the multidimensional analysis of long-term traffic data is necessary for a comprehensive and detailed understanding of the traffic behavior to provide precise information for quality-of-service (QoS) policy establishment.

Network monitoring and analysis are largely studied in two ways: one active and the other passive. In the active method, inspection packets are injected into a target network. Their response packets are then utilized to measure network performance and traffic conditions. Ping and traceroute monitoring tools which use Internet Control Message Protocol (ICMP) packets are representative of this active method. However, this method can cause an overload on the network, as it generates additional Transmission Control Protocol (TCP), User Datagram Protocol (UDP) or ICMP packets in a real operational

*Correspondence to: Myung-Sup Kim, Department of Computer and Information Science, Korea University, Sejong 339-700, Korea.
†E-mail: tmskim@korea.ac.kr

network. In addition, this method is short in providing comprehensive performance information for QoS provision and resource allocation. Conversely, in the passive method, traffic data are collected in a specific section of a network directly and analyzed to measure the performance of a target network. Sometimes Simple Network Management Protocol (SNMP) management information base (MIB) [4] data are collected from SNMP agents. Sometimes raw packet or summarized flow data are collected from a target section of network and analyzed. While the former provides slightly limited performance information, the latter can provide a detailed and comprehensive analysis of long-term traffic data. This method can also provide exact performance information for a stable QoS provision and effective resource management. In addition, the passive monitoring method is widely selected as it does not produce any overload in an operational network. However, most current traffic analysis systems based on the passive method have not reached a level where they can offer comprehensive analysis services including real-time treatment, long-term storage or various analyses of large-volume traffic generated from high-speed networks.

In fact, most of the passive traffic-monitoring systems currently do not offer comprehensive analysis information by themselves [7–9]. For example, the Multi Router Traffic Grapher (MRTG) [10] and Paessler Router Traffic Grapher (PRTG) [11], which are widely used by network operators, effectively provide the traffic load change of a target link in the form of time series graph. However, it has difficulty in obtaining traffic characteristics such as the portion of certain application traffic in total traffic. Snort [12], a representative network intrusion detection system, provides effective detection for registered attacks, but it does not offer the entire traffic trend and traffic characteristics of an individual host. These traffic analysis systems were developed to be used for specific analysis purposes, whereas network operators need a multidimensional understanding of traffic data for the stable and effective management of their networks. Thus most of the current traffic analysis systems [10,12–16] do not satisfy such demands. They mainly focus on their own analysis purposes and give little attention to other analysis demands. Therefore, in order to achieve a comprehensive understanding of their network, the network operator should install a number of traffic analysis systems of multiple purposes and combine various analysis results from them. This requires a high level of expertise in this area. Moreover, it may also generate incorrect information about the network. Three basic reasons summarize why the current passive analysis systems do not perform multidimensional analyses. First, there is the lack of an effective storage management method for large-volume and long-term traffic data. Second, there is lack of an effective integration method for different analysis methodologies that have been built separately. Third, there is a lack of methods to effectively extract useful information from huge volumes of traffic data accumulated over long periods of time.

In this paper, we propose a new design methodology for an effective traffic analysis system that can overcome the limitations of the current passive analysis methods. The core of the methodology suggested in this paper is the construction of a data cube model for multidimensional analyses required by network operators using network traffic data accumulated long term in a data warehouse. Based on the construction of the data cube model, multidimensional traffic analysis is then performed using online analytical processing (OLAP) operation, varying the abstraction level according to the purpose of the traffic analysis. In addition, *NetCube*, the proposed network traffic analysis model for OLAP on a multidimensional data cube, provides comprehensive and detailed information of long-term traffic data through data-mining techniques for various analysis purposes. We validate the feasibility and applicability of the proposed *NetCube* traffic analysis model by implementing a traffic analysis system and then deploying it to our campus network.

The remainder of this paper is organized as follows. Section 2 classifies the existing traffic analysis systems according to the various analysis purposes and defines the requirements of traffic analysis systems that this paper attempts to resolve. We also describe some related works referenced in this paper. Section 3 presents the proposed multidimensional traffic data analysis model. Section 4 describes the experimental results and performance analysis. Finally, Section 5 discusses the conclusions and future research directions.

2. REQUIREMENT ANALYSIS AND RELATED WORK

As our dependence on networks has increased in recent years, various network traffic analysis systems to understand the state, event and characteristics of a network have been developed. The purpose of traffic

DOI: 10.1002/nem

analysis can be categorized according to the following four viewpoints: trend, point, layer and event analysis. Table 1 shows the analysis items included in the four traffic analysis viewpoints and their representative systems.

2.1. Categorization of traffic analysis systems

Trend analysis is aimed at analyzing network traffic with the focus on traffic changes over time. This method is used for analysis of the temporal behavior of traffic based on a variety of time slots, such as minute, hour, day, month or year. MRTG [10] and PRTG [11] are widely used analysis systems in this category. MRTG monitors the traffic load on network links over a long period of time. It shows the measured data in several time series graphs of different time units. However, it is insufficient in providing analysis information according to the hosts and subnets for point analysis, or analysis results according to the applications and protocols for layer analysis.

Point analysis addresses the host or location from where the traffic is generated. Ntop [13] is a representative system that offers various types of analysis information of traffic according to the hosts and subnet through flow-based analysis. It shows statistical traffic information for a regular period of time, such as the amounts and percentages of bytes, packets and flows generated from hosts or subnets in descending order of download and upload traffic. Ntop has an advantage in point analysis, as it provides information based on hosts or subnets. However, an operator faces difficulties in analyzing which application created the traffic and whether the traffic is produced normally or abnormally.

Layer analysis is a method used to analyze the characteristics of traffic according to the hierarchy of the network protocol structure. In this category, the Linux L7-filter [14] traffic classification system is a representative analysis system. It determines the application program or protocol of traffic flows by payload signature matching in real time. This type of analysis method is very helpful for a network operator to understand the traffic characteristics in terms of the application programs and protocols. However, it does not provide enough data to analyze the characteristics of the traffic produced in the hosts or sections over a long period of time, as it does not pay much attention to past accumulated traffic data over time. It also has a weakness in reflecting abnormally produced traffic, as it treats such traffic as being from normal application.

Event analysis is used for the analysis of specific events on network traffic, such as abnormal behavior of traffic. Snort [12], a representative event analysis system, can detect various attacks based on predetermined detection rules. The detection results may be presented through a web-based interface such as ACID [16]. This system mainly focuses on the detection of abnormal events in real-time traffic data and it also provides good performance. However, it is difficult to acquire trend and special information about event occurrence from this system. For example, the number of hosts that are infected as zombie agents for a specific distributed denial-of-service (DDoS) attack or the amount of DDoS attack traffic changes during a certain period of time.

2.2. Requirements for comprehensive traffic analysis

Figure 1 shows the performance degree of the four representative analysis systems of each analysis viewpoints in the form of radar charts based on our investigation in Section 2.2. It shows that all of the four systems failed to cover the four analysis requirements, as they analyze the traffic data by focusing on only one specific view. For example, the MRTG system has good performance in trend analysis but not in other analysis viewpoints, while the Snort system has good performance in event analysis but not in others.

Table 1. Items based on the analysis viewpoint

Analysis view	Analysis items	Example systems
Trend analysis	Minute, hour, day, month, year	MRTG, PRTG
Point analysis	Host, subnet, department, building, network	Ntop
Layer analysis	Access type, network, transport, app. protocol, application	L-7 filter
Event analysis	Abnormal detection, misuse detection	Snort

DOI: 10.1002/nem

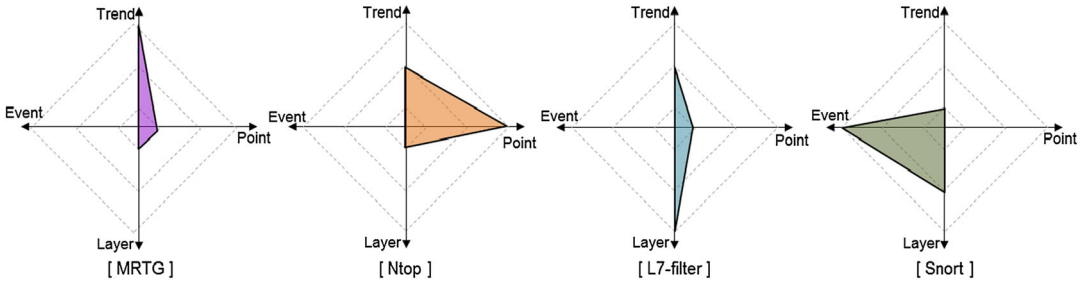


Figure 1. Performance evaluation based on traffic analysis viewpoints with radar charts

Following their specific purposes, many traffic-monitoring systems currently conduct their analysis from only a partial viewpoint due to the structural complexity of a system increasing when analyzing traffic data from multidimensional viewpoints.

One of the main concerns of network operators is how to provide a stable and efficient network environment to users on limited network resources and diverse user demands. For efficient network resource management a network operator requires a comprehensive and detailed understanding of the long-term traffic data from the target network which reflects all four analysis viewpoints. However, the existing analysis systems based on their own purposes independently collect, manage and analyze traffic data in order to meet their own reasons for analysis. No single traffic analysis system provides all the information that is required by the network operators. A comprehensive and detailed understanding of traffic data over a long period of time is very helpful for efficient QoS provision and resource management. To do this, a network operator should deploy a number of individual traffic analysis systems for the four viewpoints on a target network. He or she also needs to combine the analysis results from all these systems in order to infer what needs to be known about the network. As this work is so error prone and sophisticated, it requires much experience and expertise by the network operators.

To reduce any mistakes and extract the correct information from the analysis data that is generated by a number of individual analysis systems, we need a comprehensive analysis model which combines, stores and analyzes long-term and massive traffic data. A multidimensional analysis is needed to offer effective traffic information that fits various analytical purposes. For example, let us suppose that we want to carry out a traffic engineering task by analyzing a crowded section where a bottleneck often occurs or a section where the usage rate of a link is low. For this task, we should analyze the traffic generation trend for a specific section according to the time slots. In terms of the application layer, we have to obtain the exact characteristics of the traffic that is generated in a crowded section through the observation of application usage. Understanding the traffic from a multidimensional viewpoint is necessary to draw useful information that fits the purpose of our analysis for efficient QoS provision and resource management.

2.3. Related work for multidimensional long-term traffic analysis

Most of the current traffic analysis systems are interested in analyzing the traffic data on-the-fly from target network links [17,18]. They do not pay much attention to past traffic data. What they usually do on the past analyzed traffic data is to summarize their biased analysis results, store them in the form of log data and provide them as they are without any further analysis. The original traffic data are not stored, to reduce storage costs. Furthermore, although correlation analysis is necessary for a comprehensive and detailed understanding of traffic data, it is not performed on the analysis results obtained from the different analysis systems. There are several reasons why current passive traffic analysis systems do not perform comprehensive and detailed analyses for long-term massive traffic data. First is the lack of an effective storage management method for large-volume and long-term traffic data. Second is the lack of an effective integration method for different analysis methodologies that have been separately built. Third is the lack of an effective method to extract useful information from large volumes of traffic data that are accumulated over a long period of time.

DOI: 10.1002/nem

There are several related research studies [19–24] that may be referenced to achieve multidimensional traffic analysis for comprehensive and detailed understanding of long-term traffic data. Wang *et al.* [19] proposed a data-warehousing and data-mining approach to discover comprehensive and detailed information in BitTorrent P2P application. They collected BitTorrent data such as IP, Time To Live (TTL) and bandwidth from the Internet. They generalized the collected data to a certain level of abstraction, integrated the data into a data warehouse and mined association rules for the user. They extracted much valuable information such as the user's habitual behaviors, data access patterns, interested files and so on. In this paper, we took a similar technique such as correlation and data-mining technique to analyze long-term traffic data. Mansmann *et al.* [20] utilized the data cube and OLAP operation to handle massive long-term traffic data to develop a hierarchical network map which is a network traffic visualization method. Although they did not perform more comprehensive and detailed analysis such as data-mining analysis, the data cube and OLAP operation is a good reference for handling massive long-term traffic data. In this paper, we develop a data cube to store long-term traffic data and we also used the OLAP operation for retrieval of useful information from the cube. Zhang *et al.* [21] also utilized a data warehouse to store IPv6 traffic data and provided it to the next back-analysis node effectively in a cluster-based traffic analysis system. Many research studies and systems use a data warehouse to store massive data and OLAP operations to retrieve useful information from the data. We believe that these approaches can be selected to analyze long-term massive traffic data for QoS provision and resource management.

In this paper, we propose a new analysis model for long-term and massive traffic data that can overcome the limitations of the current passive analysis systems. The core of the proposed analysis model is the construction of a data cube in a data warehouse and an OLAP operation for multidimensional traffic analysis. The proposed model can conduct a number of data-mining techniques for the extraction of useful information for the network operators. We believe that a comprehensive and detailed analysis resulting from our proposed analysis model will be the basis for a QoS policy establishment that contributes to a more stable network operation.

3. *NETCUBE*: A COMPREHENSIVE TRAFFIC ANALYSIS MODEL

In this Section, we present the *NetCube* multidimensional traffic analysis model, which comprehensively reflects the viewpoints of trend, point and layer. *NetCube* can analyze long-term massive traffic data through various OLAP operations at multiple levels of abstraction.

3.1. Overall NetCube traffic analysis model

Figure 2 illustrates an overall concept of *NetCube* traffic analysis model which is proposed in this paper. As you can see in Figure 2, *NetCube* traffic analysis model consists of three layers from online traffic capture to the representation of analysis results: application layer traffic identification, data warehouse with OLAP and multidimensional analysis.

In the first layer, *NetCube* traffic analysis model performs application layer traffic identification process. Raw packets are captured from a target network link and aggregated into a group of flows. A flow is a set of packets belonging to the same session between two end points. The formation of flow can reduce the size of traffic data by removing unnecessary payload data and repeating header data from packets. Application name of each flow is determined in the first layer. The flow information is delivered into the next layer for storage. In the second layer, the flow information is stored in the data. The flow information contains IP address, port number, timing information such as start and end time of the flow, the number of packets and bytes in both directions, and application name. We used the data warehouse for the effective storage of long-term massive traffic data, where the flow data are stored in the form of a round robin database to solve the increasing storage size as time goes on [24,25]. We used the OLAP operation to retrieve information from the data warehouse. In the third layer, we performed multidimensional analysis using OLAP and data-mining technique. We also retrieved various useful items of information, based on which the network operator can understand the status of the target network correctly and establish QoS policy efficiently for network resource management.

DOI: 10.1002/nem

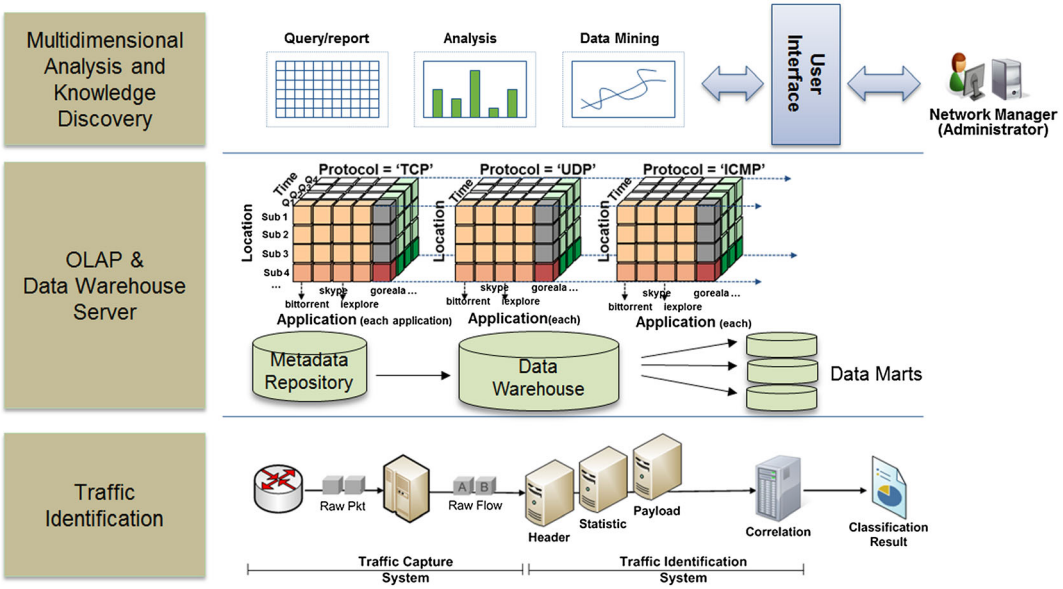


Figure 2. Overall *NetCube* traffic analysis model

3.2. *Application traffic identification*

In this section, we describe the first layer of Figure 3, the application traffic identification method, which we developed in the *NetCube* traffic analysis model. Flow-based identification is the basis of our *NetCube* analysis model to realize the trend, point and layer analysis concerning long-term and massive traffic data. Figure 3 describes the overall architecture of our traffic identification system, which consists of four consecutive levels.

In the first level, the flow generator (FG) captures all the raw packets from a target network and aggregates them into a group of flows. A flow is a summarization of packets generated in a connection between two end hosts. The flow data are delivered to the second level, where several individual identification modules work in order to identify the application name of each flow separately. We developed three signature-based identification methods (HSC, SSC and PSC) and one behavior-based identification

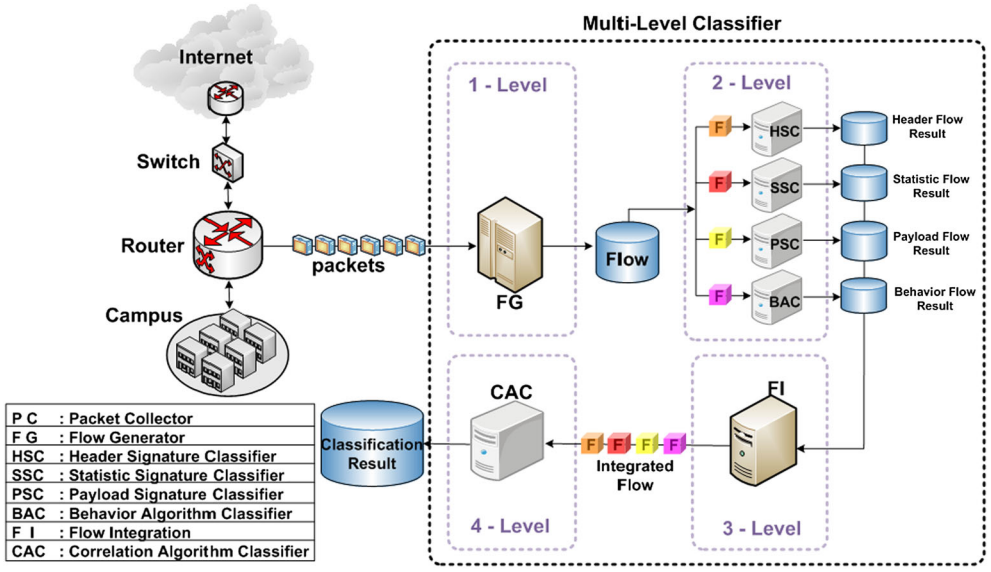


Figure 3. Multi-level application traffic identification system

DOI: 10.1002/nem

method (BAC) [13]. Each identification module does its best to determine the application name of each flow based on its own algorithm. For a single flow each identification module may or may not determine the application name.

The identification results of each flow are integrated in the third level: flow integration (FI). We used a priority-based selection algorithm for the selection of the best result. In the final level, the correlation algorithm classifier (CAC) additionally determines the application name of unknown flows which are not identified in the second level. We utilized the relationship among the already identified flows and unknown flows to determine the application name of the unidentified flows [6].

The identified flows by the application are stored in the data warehouse for comprehensive and detailed analysis. The IP information in a flow can be utilized for point analysis. Moreover, the port number, protocol name and application name are utilized for layer analysis. Timing information such as flow starting and ending time is the basis for trend analysis. We can also utilize the flow data, which was accumulated for a long period of time for the trend analysis. Furthermore, the *NetCube* analysis model can extend to the event analysis by adding any event analysis method such as snort [12] in the first layer of Figure 2. This is will form one of our future studies.

3.3. Multidimensional data cube

Data warehouses and OLAP tools are based on a multidimensional data model. This model views data in the form of a data cube [23,24,26–29]. A data cube allows data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts. In general, dimensions are the perspectives or entities with respect to which a user wants to keep records. The fact table contains the names of the facts, or measures, as well as keys to each of the related dimension tables. In this paper, we defined four dimensions: time for a trend analysis, location for a point analysis, and application and protocol for a layer analysis. We also used the bandwidth and flow count of the traffic as measurements in a fact table. The data cube model in this paper consists of tables for the four dimensions and one fact table. The dimension in the data cube can be easily added if a new analysis viewpoint is required.

A data warehouse requires a concise, subject-oriented schema that facilitates online data analysis. The most popular data model for a data warehouse is the multidimensional model [26–31]. In this paper, we used a star schema for the multidimensional analysis model. The data warehouse in it contains a large central table (fact table). This table contains a bulk of data and a set of smaller attendant tables (dimension tables), one for each dimension. Figure 4 shows the star schema for *NetCube*.

The storage space necessary to store traffic data is continuously increasing as time passes. In this paper we propose a time-based incremental compression method inspired from the round robin database. Usually the information of the past traffic data is less important than that of the recent traffic data. The MRTG and PRTG reduces the necessary storage space by summarizing past traffic data in the time

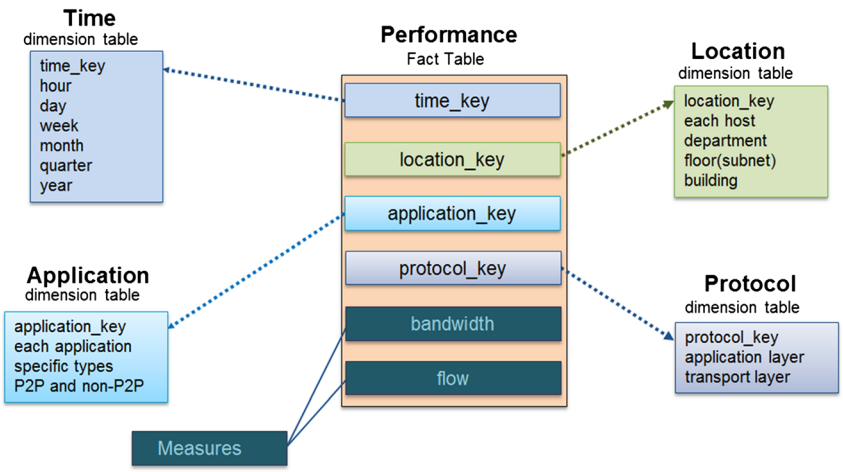


Figure 4. *NetCube* star schema

DOI: 10.1002/nem

dimension. Basically the *NetCube* storage model stores the traffic data in the form of flow, not packet. The *NetCube* storage model increases the compression level on each dimension more and more for old traffic data. It reduces the storage space by filtering out unimportant information in the application dimension, by aggregating traffic data with regard to a common metric in the location dimension and by summarizing measures in the time dimension. Therefore, the *NetCube* data warehouse provides more summarized and aggregated information about past traffic data as time goes on, and more detailed and specific information regarding recent traffic data, respectively. The time-based incremental compression method makes the *NetCube* able to retrieve comprehensive and detailed information for long-term and massive traffic data within a limited storage space. The compression level can be determined dynamically in accordance with storage size, analysis preference on each dimension, etc.

3.4. Concept hierarchy and OLAP operation

A concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level and more general concepts [26,27]. Concept hierarchies allow data to be handled at varying levels of abstraction. The attributes of a time dimension are organized in a partial order, forming a lattice. The partial order for the time dimension based on attributes such as day, week, month, quarter, and year is 'day < {month quarter; week} < year' (Figure 5a). Figure 5(b) shows a concept hierarchy for the location dimension. Here, a host indicates each individual IP in a target network. The concept hierarchy for the location dimension is based on the structure of a hierarchical tree [26,27], from the lowest host to the subnet of each floor and that of each building. An application dimension table is defined as a low-level concept, i.e. each application, to the highest concept, i.e. P2P and non-P2P, as classified in an Internet application traffic classification system [32] (Figure 5c). Table 2 shows the most popular 30 application types that are currently used at home and abroad. These take up over 96% of the entire traffic from around 200 applications that are currently generated in a campus network [25,32]. Finally, Figure 5(d) shows a concept hierarchy for the protocol dimension that has a structure mapping from an application layer to a transport layer.

The incremental compression of traffic data is performed according to the concept hierarchy at each dimension. For recent traffic data all of the leaf node information in the hierarchy is stored at the *NetCube* data warehouse. As time passes, the leaf node data in the time hierarchy are summarized into upper nodes up to the month or week nodes. The leaf nodes in the location hierarchy are aggregated up to the building

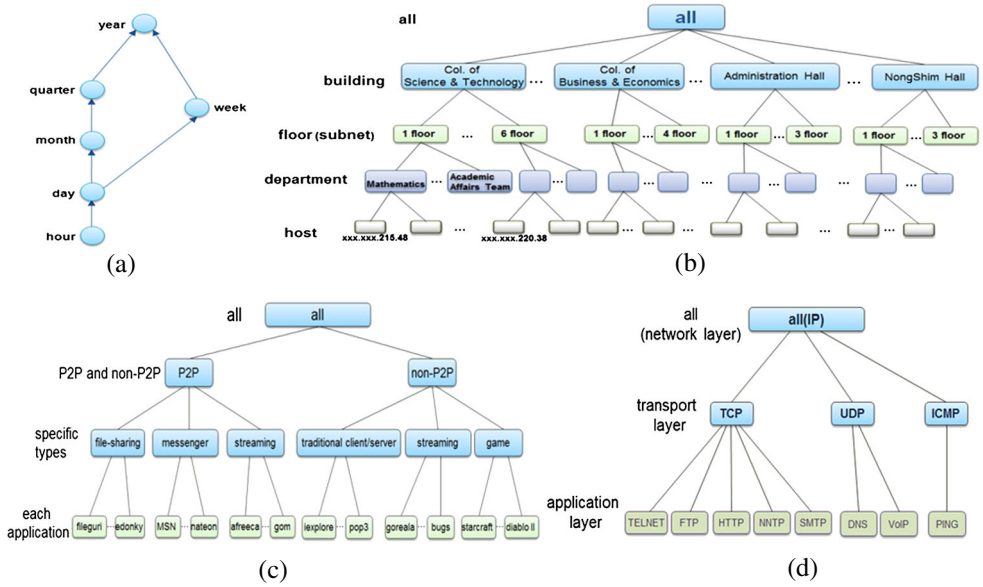


Figure 5. Concept hierarchy for (a) the dimension time; (b) the dimension location; (c) the dimension application; (d) the dimension protocol

DOI: 10.1002/nem

Table 2. P2P and non-P2P applications used for the experiment

P2P	File sharing	Fileguri, Donkey, BitTorrent, jjangfile.net, ToToBrowser, Soribada, ...
	Messenger	NateOn, MSN (Windows Live), Skype, Tachy, ...
	Streaming (TV)	Afreeca, Gretech (GOM), PotPlayer, TVUPlayer, ...
Non-P2P	Traditional client/server	Internet Explorer, V3 (AhnLab), FileMon, OnDisk, CoDisk, QFile, ...
	Streaming	Gorealra, Bugs, ALSong, Rainbow, cyMini, ...
	Game	Netmarble, StarCraft, Diablo II, Hangame, Pmang, ...

level. In the application dimension, the application nodes with a small amount of traffic are filtered out and combined into one common application type to reduce storage space. The removing order of nodes in the incremental compression in the hierarchies can be dynamically determined according to the user's analysis preference.

In the multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple levels of abstraction defined by concept hierarchies. This organization provides users with the flexibility to view data from different perspectives. A number of OLAP data cube operations exist to materialize these different views, allowing interactive querying and analysis of the data at hand. Hence OLAP provides a user-friendly environment for interactive data analysis [26–31], which makes it possible to construct a flexible and extensible analysis system.

Several OLAP operations such as *roll-up*, *drill-down*, *slice* and *dice* are used to retrieve useful information from the data warehouse of traffic data. The *roll-up* operation performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction. When a *roll-up* is performed by dimension reduction, one or more dimensions are removed from the given cube. The *drill-down* operation is the reverse of the *roll-up* operation. It navigates from less detailed data to more detailed data. The *drill-down* operation can be realized by either stepping down a concept hierarchy for a dimension or by introducing additional dimensions. As a *drill-down* operation adds more detail to the given data, it can also be performed by adding new dimensions to a cube. The *slice* operation selects one dimension of the given cube and this results in a subcube. The *dice* operation defines a subcube by performing a selection on two or more dimensions.

3.5. Traffic analysis using data-mining techniques

In this section, we describe a couple of data-mining techniques to retrieve useful information for a comprehensive and in-depth understanding of traffic data by using the multidimensional analysis results: application growth analysis and similarity analysis.

Understanding the traffic usage of users or subnets for long periods of time and also classifying users and subnets according to the characteristics of traffic are very important for a network operator in order to establish QoS policy for efficient network resource management. In this paper, we utilized the period-based trend analysis method proposed by Lee *et al*. [33] to measure the changes in traffic usage by applications, and we also utilized the cosine similarity analysis method to group subnets that possess similar traffic characteristics. Analysis of application traffic growth during several continuous periods is very important considering the current network environment where the application usage changes frequently and rapidly. Information about the bandwidth changes of a certain application during the past six months or one year will be very helpful for a network operator to understand the current popularity of the application. Moreover, a network operator can expect future usage of the application correctly for QoS policy establishment. The purpose of the application traffic growth analysis is to retrieve exact data about the traffic changes of each application over a long period of time.

In this paper we defined an application growth index (AGI), as indicated in equation (1), to analyze the application traffic growth analysis. The AGI is a relative indicator used to represent the growth rate of an application in the second period compared to that of the first period in traffic bandwidth:

DOI: 10.1002/nem

$$\text{AGI} = \frac{\text{Number of second period applications} - \text{Number of first period applications}}{\text{Number of second period applications} + \text{Number of first period applications}} \quad (1)$$

However, the AGI cannot represent an absolute growth value of each application, as the AGI is a relative rate between two specific periods of time. Therefore, we can derive WAGI (weighted AGI) from the AGI to indicate an absolute application growth value as shown in equation (2):

$$\text{WAGI} = |\text{Number of second period applications} - \text{Number of first period applications}| \times \text{AGI} \quad (2)$$

The WAGI is an indicator that reflects on the absolute application growth while the AGI is an indicator that reflects on the relative growth of an application. That is, the WAGI increases proportionally with the increased value of the applications. The WAGI is more helpful in understanding the traffic growth of an application for a long period of time than the AGI. In this paper we utilize the WAGI rather than the AGI for the analysis of traffic changes of each application for a target period of time.

Enterprise network is usually subdivided into a number of subnets to increase manageability. The traffic characteristics of each subnet may be different from each other as the user's interest and work contents are different. Understanding the traffic characteristics of each subnet can be helpful for a network operator in establishing different QoS policies that are suitable for each subnet. The subnets can also be categorized into groups that have similar traffic characteristics:

$$\cos(d_x, d_y) = \frac{\sum_{i=1}^{n} \left(w(a_i, d_x) \times w(a_i, d_y)\right)}{\sqrt{\sum_{i=1}^{n} w(a_i, d_x)^2 \times \sum_{i=1}^{n} w(a_i, d_y)^2}} \quad (3)$$

In this paper we propose a similarity measurement method among the subnets which is used as a basis for the establishment of QoS policy in the enterprise network. In order to measure the similarity of traffic characteristics among subnets we used the cosine similarity measurement, which is broadly utilized in text mining with good performance. The function to measure the similarity between two subnets d_x and d_y is represented in equation (3), where a_i and w indicate an application and weight factor, respectively. We used the application bandwidth as the input data for the weight matrix in the cosine similarity function. The function gives the similarity measures between every two subnets based on which network operators can easily set up a QoS policy suitable to each subnet.

4. EXPERIMENTS

In this section, we describe a multidimensional traffic analysis system based on the *NetCube*. We also prove the feasibility and applicability of the *NetCube* by describing the experimental results that were applied to the traffic data collected in a real campus network.

4.1. NetCube: a comprehensive traffic analysis system

We developed a proof-of-concept traffic analysis system of the *NetCube* traffic analysis model based on the KU-MON traffic analysis system [9]. The overall traffic analysis system is illustrated in Figure 6. KU-MON enables real-time data collection and analysis of the Internet traffic in an enterprise network. This is very flexible and can be easily combined with other traffic analysis systems. KU-MON performs the application traffic identification in real time. The *NetCube* traffic analysis system takes the identified flow data from the KU-MON system and performs a comprehensive and detailed analysis using OLAP operation and data mining.

The overall *NetCube* traffic analysis system is composed of five modules: (1) the flow generator module, which captures all the packets from a target network link and creates a corresponding flow information in real time; (2) the flow store module, which stores the created flow data for a period of time for online cascading analysis and offline analysis; (3) the traffic identifier module, which performs application traffic identification and determines the application name for each flow; (4) the

DOI: 10.1002/nem

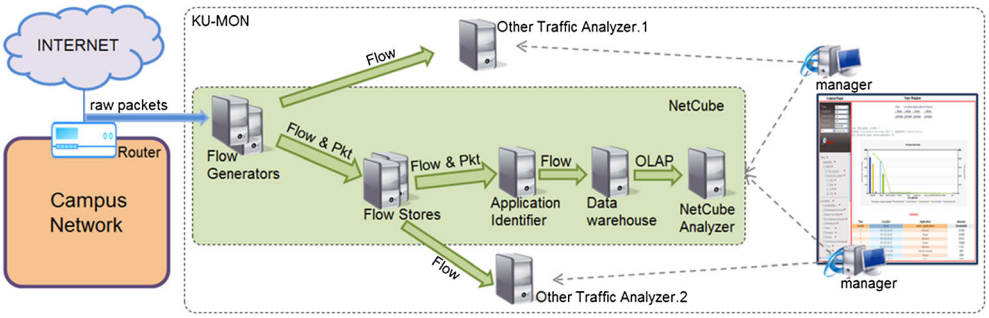


Figure 6. KU-MON: real-time traffic collection for *NetCube* traffic analysis

data warehouse module, which stores the identified flow in a data cube for OLAP operation; 5) the *NetCube* analyzer, which performs comprehensive analyses using various data-mining methods.

We deployed the *NetCube* traffic analysis system on our campus network. The Internet access point of our campus network handles up to 300 Mbps of bandwidth. This is the total inbound and outbound traffic. A campus network has various types of traffic, generated by about 3000 hosts. We implemented most of the system using C and C++ languages. We especially implemented a prototype of data warehouse and *NetCube* analyzer using Oracle 10 g and JSP, respectively. The user interface of the *NetCube* system was designed to allow a network operator to easily draw up questions on the Web, not through traditional complex query languages regarding OLAP operations, but through a number of convenient buttons which allow an easy and fast retrieval of analysis results. It is illustrated in Figure 6.

The experiments were conducted with the traffic data collected through KU-MON. We collected the entire flows and packets of a campus network by using KU-MON from March to July 2011 to monitor and analyze the campus traffic. Table 3 shows the statistics for the traffic data used in this experiment. The amount of traffic collected was around 168 Tbytes for a five-month period. When it was transformed into a flow format, it amounted to around 507 Gbytes.

We can dramatically reduce the storage space necessary to store traffic data by storing the flow data instead of packet data in the proposed *NetCube* data warehouse. Furthermore, we can save storage space efficiently by applying the proposed time-based incremental compression method. The compression level might be changed according to the user's interest and storage space. Figure 7(a) shows the compression level we applied in accordance with the age of traffic data in this experiment. It shows that compression in the location dimension begins from traffic data 2 years old. Traffic data over 2, 4 and 5 years old are aggregated into department, subnet and building level, respectively. In the time dimension the compression occurs at hour level for traffic data more than 1 year old. They are summarized into the day or week level as the age reaches 3 or 5 years, respectively. In the application dimension, all the flow data less than the corresponding threshold of flow size are aggregated into application type. The threshold of flow size increases as time goes on from 0.2 kbytes to 10 kbytes.

Figure 7(b) shows the amount of storage required over years when the proposed compression level is applied to each of the 5 months' traffic data in Table 3. Figure 7(b) shows that only 2 Mbytes of storage space is sufficient to store the flow data of July 2011 when it is 5 years old. It is only 2% of

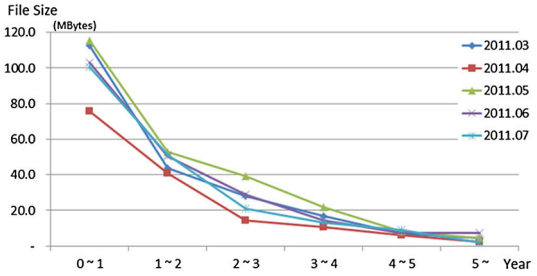
Table 3. Traffic trace statistics

Period	Flow Count(×1000)	Packet Count (×1 000 000)	Byte Size (Gbytes)	Flow file size (Mbytes)
2011. 03	469,618	37,309	28,124	112,708
2011. 04	314,592	27,345	21,661	75,816
2011. 05	476,555	48,470	38,141	115,326
2011. 06	428,189	47,729	36,776	103,193
2011. 07	416,401	52,814	43,818	100,769
Total	2,105,355	213,667	168,520	507,813

DOI: 10.1002/nem

Age of Traffic Data	Compression Level		
	Location Dimension	Time Dimension	Application Dimension
0 ~ 1 years	Host	Minute	App. Flow All
1 ~ 2 years	Host	Hour	App. Flow >0.2KB
2 ~ 3 years	Department	Hour	App. Flow >0.5KB
3 ~ 4 years	Department	Day	App. Flow >1KB
4 ~ 5 years	Subnet	Day	App. Flow >10KB
5 ~ years	Building	Week	App. Type All

(a)



(b)

Figure 7. (a) Compression level in time-based incremental compression. (b) Storage size required to store each of the 5 months of traffic data over time

the original size of flow data (100.8 Mbytes from Table 3). We can estimate that the total amount of storage space required to store all of the flow data during a 6-year period is 2.3 Tbytes on our campusThe following formula is a query statement with a symbolic network, supposing that the same amount of traffic data to July 2011 are generated during the whole 5-year period. This result shows that the proposed time-based incremental compression method is efficient for storing large-volume and long-term traffic data in the *NetCube* data warehouse.

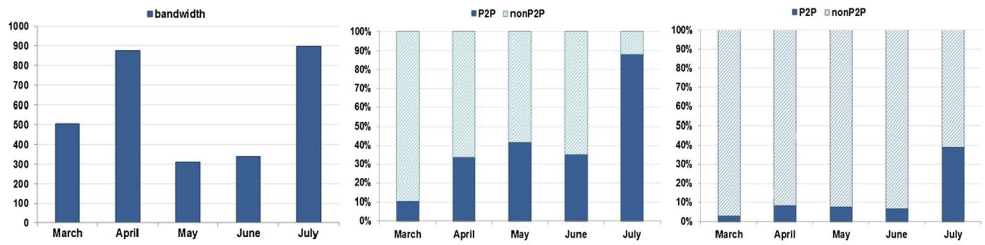
4.2. Traffic analysis results using OLAP

This section introduces a multidimensional analysis that can be obtained by choosing a dimension within the viewpoint that a network manager is attempting to analyze, and also by adjusting the level of abstraction using the *NetCube* model described in Section 3. First, we show that the system can be used to analyze network user trends and usage patterns.

The following formula is a query statement with a symbolic meaning that is used to analyze the changes in application usage according to the time slots:

$$\begin{aligned} &\textbf{Dice} \text{ for Time } = \text{''March 2011 ? July 2011''} \text{ AND} \\ &\text{Application } = \text{''all''} \text{ AND measures } = \text{''bandwidth and flow''} \end{aligned} \tag{4}$$

Figure 8 shows a group of graphs that result from formula (4). The graphs show that the bandwidth proportion of P2P application programs is relatively higher than the flow proportion in all months (Figure 8b, c). This indicates that P2P application consumes more bandwidth than non-P2P applications in each connection. In particular, P2P bandwidth occupies more than 80% of the entire bandwidth in a campus network during the July vacation period (Figure 8a, b). Thus QoS management policy for related P2P application programs is needed for the summer vacation, as excessive usage of P2P interferes with the stable services for daily business traffic.



(a) (b) (c)

Figure 8. (a) Monthly bandwidth. (b) Monthly application bandwidth. (c) Monthly application flow

DOI: 10.1002/nem

Using formula (4), we discovered that P2P traffic occupied an excessively large amount of the campus network in July. However, detailed data regarding the applications used are required to exactly understand the traffic features more and to establish an effective QoS management policy. Thus we conducted a *drill-down* operation for the application dimension, as shown in formula (5), in order to understand the results of the detailed analysis. Figure 9 shows the results of the operation. Figure 9 (a) reveals that the type of P2P application which used a large amount of bandwidth in July was a file-sharing application:

$$\begin{aligned}&\textbf{Drill-down} \text{ on Application (from all to specific types),}\\&\textbf{Dice} \text{ for Time } = \text{″ March 2011 ? July 2011″ AND}\\&\text{Application } = \text{″ specific types″ AND measures } = \text{″ bandwidth″}\end{aligned} \tag{5}$$

For the application dimension, a network manager can conduct a *drill-down* operation, as shown in Formula (6), down to the lowest level to confirm in detail what file-sharing application programs are being used. The results of the P2P application analysis show that the file-sharing programs used were Fileguri, which is a domestic Korean program, and BitTorrent, which has many users both at home and abroad (Figure 9b):

$$\begin{aligned}&\textbf{Drill-down} \text{ on Application (from specific types to each application),}\\&\textbf{Dice} \text{ for Time } = \text{″ March 2011 ? July 2011″ AND}\\&\text{Application } = \text{″ each application″ AND measures } = \text{″ bandwidth″}\end{aligned} \tag{6}$$

The system was able to analyze trends and application layers simultaneously. It can also obtain the kinds of application programs that generated large amounts of traffic out of the entire traffic in the campus network.

Next, we chose the time and location dimensions and we analyzed changes in the bandwidth usage of each department according to the time slots:

$$\begin{aligned}&\textbf{Dice} \text{ for Time } = \text{″ March 2011 ? July 2011″ AND}\\&\text{Location } = \text{″ department″ AND measures } = \text{″ bandwidth″}\end{aligned} \tag{7}$$

From formula (7), Figure 10 presents the changes in bandwidth usage according to university departments from March to July of 2011. In particular, the bandwidth used in the Department of Computer Science in July covered more than 90% of the total bandwidth of the campus.

In order to carry out a detailed analysis for the Department of Computer Science, we drilled down from month to hour for the time dimension, as illustrated in formula (8), from the department to each host for the location dimension, and from P2P and non-P2P to each application for the application dimension. This department used a large amount of traffic in July. We also chose a transport layer for the protocol dimension to conduct an additional analysis on the protocols used:

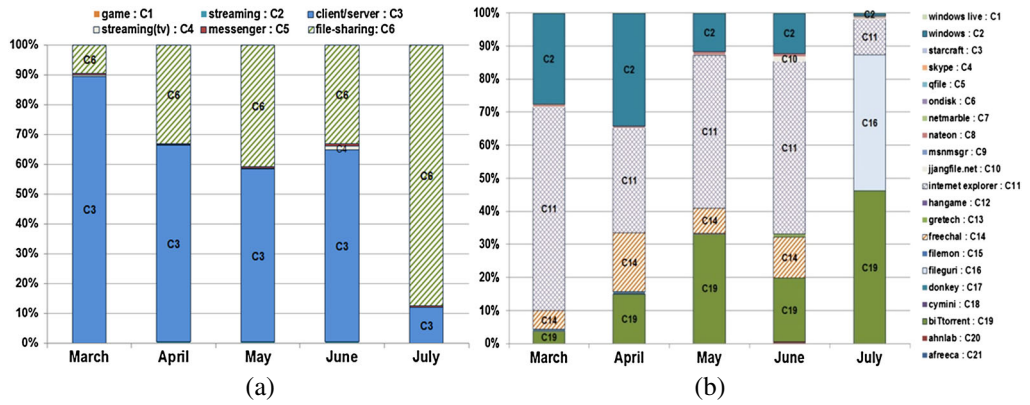


Figure 9. (a) Monthly application (specific types) bandwidth. (b) Monthly application (each application) bandwidth

DOI: 10.1002/nem

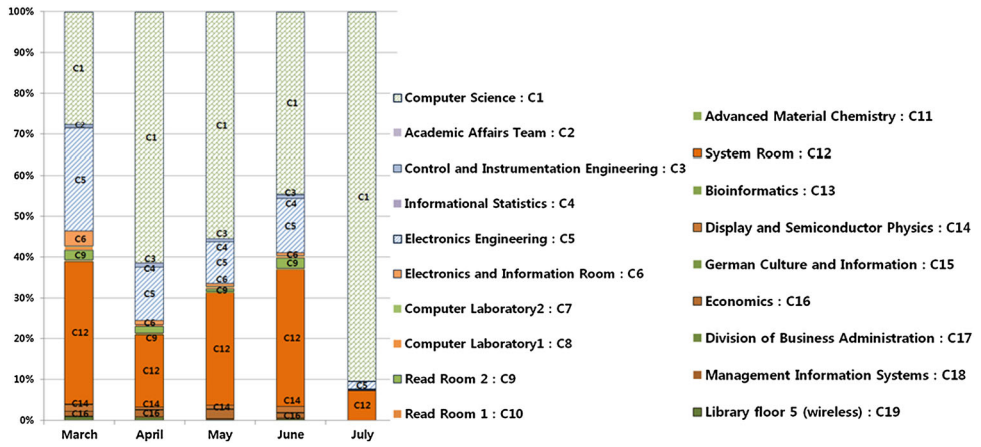


Figure 10. Usage rate of traffic for each department by time slot

Drill-down on Time (from month to hour) and Location (from department to each host) and Application (from P2P and non-P2P to each application) and Protocol (from all to transport layer),
Dice for Time = ″ hour and July 2011″ AND Location = ″ each host and Computer Science″ AND Application = ″ each application″ AND Protocol = ″ transport layer″ AND measures = ″ bandwidth (8)

The analysis shows that unusual large amount of traffic usages appear six times over several days in July. These unusual high traffic usages are caused by high increase of UDP flows, most of which were data transfer flows by P2P applications. Actually, over 50% of the entire flow used the UDP protocol (Figure 11a). In particular, hosts xxx.xxx.230.3 and xxx.xxx.230.39 used a large amount of bandwidth for BitTorrent, and hosts xxx.xxx.230.3 and xxx.xxx.230.63 did the same for Fileguri (Figure 11b).

In this section, through a multidimensional analysis, we described a process based on a certain scenario that elicits information that includes the type and name of an application which generates a vast amount of traffic within a specific period, as well as the protocol, building, department, host and time used. Results show that *NetCube* is a system that offers an environment whereby a network manager can extract various results for a given purpose. This is done by choosing a dimension within the viewpoint that he or she is trying to analyze, by adjusting the level of abstraction for the chosen dimension, and by conducting a multidimensional analysis.

4.3. Traffic analysis results using data mining

In this section, based on the traffic analysis results from the *NetCube* analysis model, we describe an example process for the retrieval of useful information by applying the data-mining analysis method for the establishment of QoS management policy.

Figure 12 describes the WAGI (weighted application growth index) calculated between two periods: the first period being June, the end of spring semester; and the second period July, the summer vacation. Figure 12(a, b) shows the WAGI of P2P and non-P2P applications from the perspective of bandwidth and flow count, respectively. As seen in Figure 12(a, b), the WAGI of P2P has a positive value in bandwidth and flow count, while the WAGI of non-P2P has a negative value. This indicates that the traffic of P2P has increased in July compared to that in June. In order to examine this in more detail, we performed a *drill-down* operation into the application dimension. We found out that the file-sharing application has significantly increased in the second period over other applications, as illustrated in Figure 12(c, d).

As the summer vacation began, the use of file-sharing P2P applications increased compared with other types of applications. The use of applications and the resultant traffic amount generated in the

DOI: 10.1002/nem

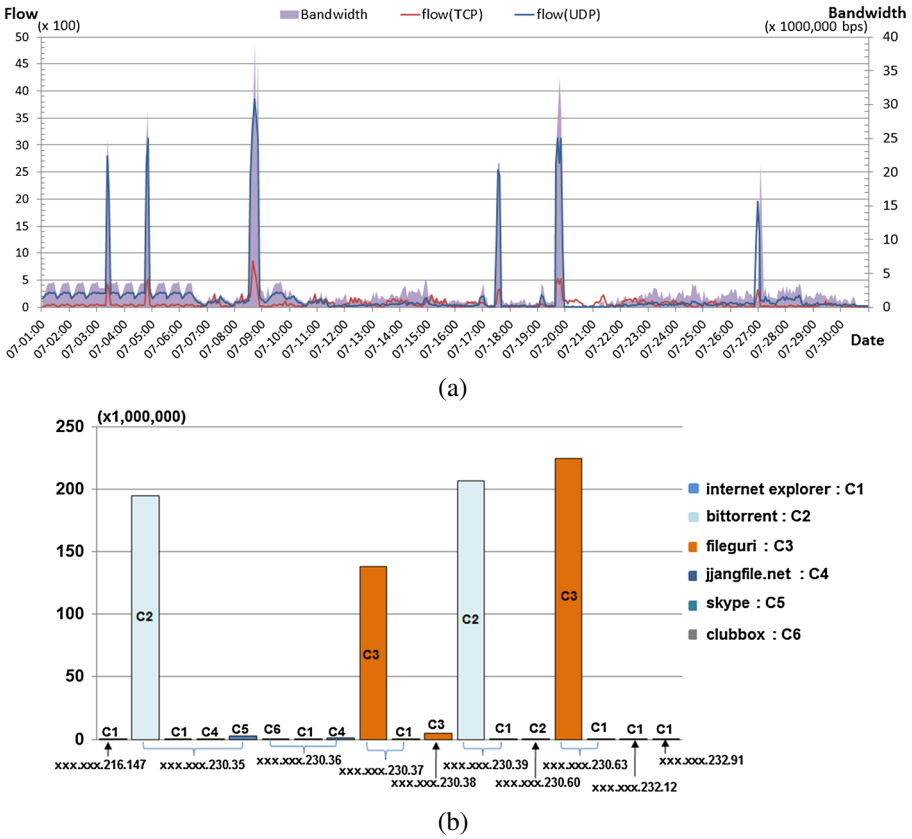


Figure 11. (a) Bandwidth and flow usage of each protocol in the Department of Computer Science in July. (b) An analysis of the application bandwidth used by the host

network changed dynamically over time. Therefore, a network operator needs to establish an adaptive and flexible QoS policy to control network resources based on application growth.

Figure 13(a) presents an application traffic distribution for each subnet analyzed using formula (9). Figure 13(b) shows a subnet similarity map that was drawn by applying the cosine similarity function. The size of the circle in Figure 13(b) indicates the amount of traffic generated by each subnet. The thickness of the links among the nodes represents the degree of relationship between the nodes:

$$\begin{aligned}&\text{Dice for Location} = \text{''department''} \text{ AND}\\&\text{Application} = \text{''each application''} \text{ AND measures} = \text{''bandwidth''}\end{aligned} \tag{9}$$

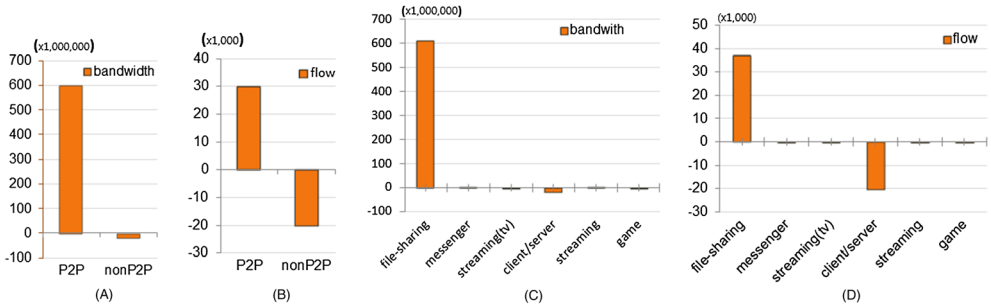


Figure 12. The WAGI of P2P and non-P2P applications in (a) bandwidth and (b) flow. Lower application level WAGI in (c) bandwidth and (d) flow

DOI: 10.1002/nem

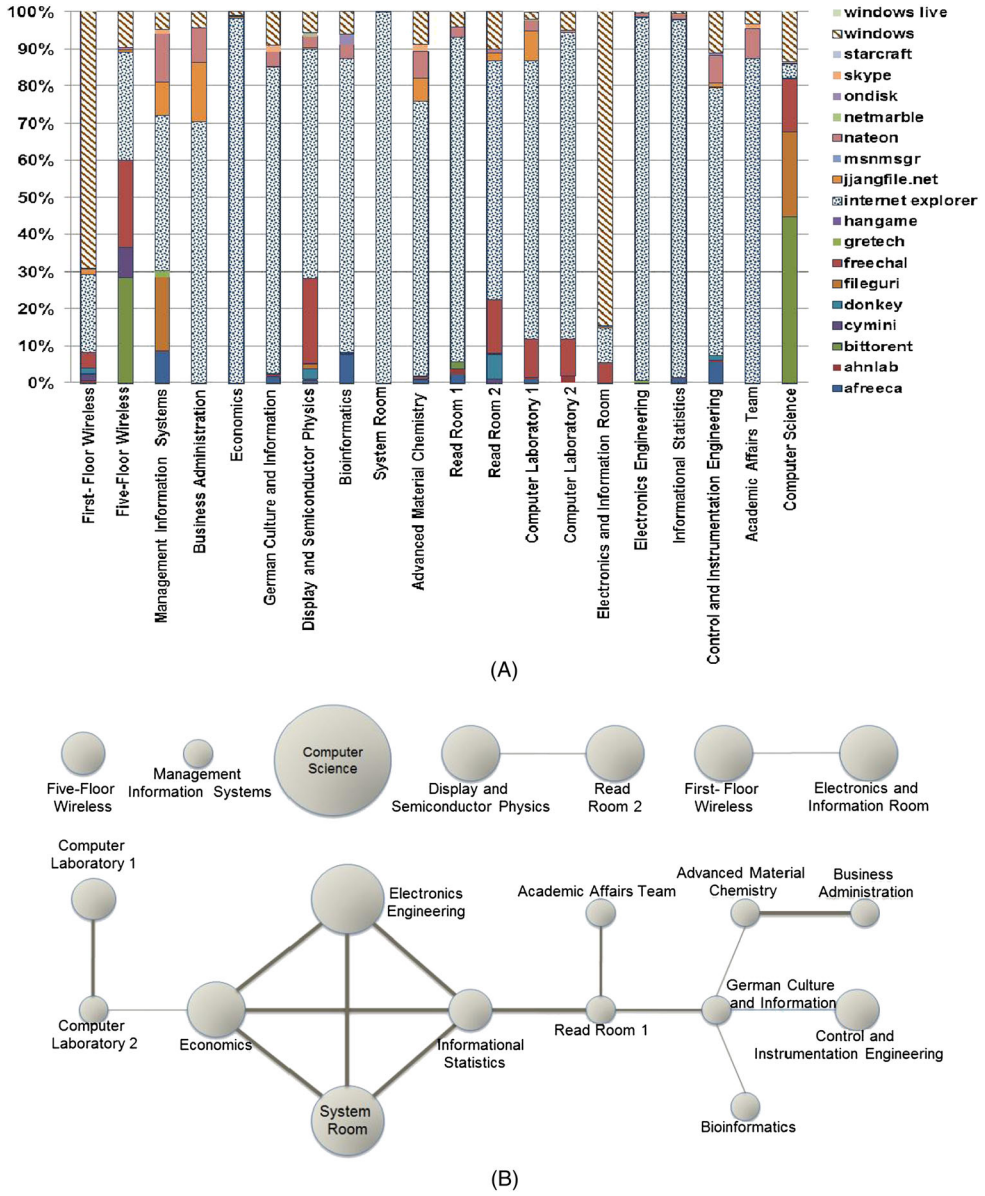


Figure 13. (a) Application traffic distribution by subnets. (b) Traffic pattern similarity map by subnets

First, the subnet of Economics, the subnet of Electronics Engineering, the system room, and the subnet of Informational Statistics have tightly connected links among them. This indicates that these subnets have very similar traffic patterns during the experimental period. Indeed, the major application consuming most of the network bandwidth in these subnets was the web browser from Microsoft: Internet Explorer.

The two subnets, the subnet of Advanced Material Chemistry and the subnet of Business Administration, are very similar to each other in terms of traffic patterns, even though users from these departments have significant differences in research area. The major applications in these departments were Microsoft IE, JJangfile.net (a P2P file-sharing application) and NateOn (a popular P2P messenger program in Korea). Conversely, the subnet of Computer Science is isolated from other subnets. It is the biggest bandwidth consumer, having a unique traffic characteristic among all the subnets in our campus.

The application distribution result and the relationship map generated by similarity analysis can be very helpful for a network manager to understand the traffic characteristics of each subnet and for the establishment of QoS policy differently and effectively, according to the similarity result. This analysis result can be utilized to plan the future network extent as well.

DOI: 10.1002/nem

5. CONCLUSION AND FUTURE WORK

As dependence on networks has increased in recent years, many different network traffic analysis systems have been developed to effectively operate the present state of a network. This paper outlined the problems of existing traffic analysis methodologies. We suggested a multidimensional traffic analysis model using data cube to elicit comprehensive analysis results that fits various purposes. As the traffic analysis system using data cube can easily extract comprehensive analysis results within the viewpoints of trend, point and layer analyses for long-term traffic data, the system can be used beyond traffic monitoring in an existing limited area. It can also be expanded to monitor and analyze traffic enterprise wide. The data cube system, named *NetCube*, consists of four dimensions—time, location, application type and protocol—and uses an incremental compression storage model. The system can be used to conduct a multidimensional analysis through an OLAP operation according to the level of abstraction for each dimension. The application growth analysis over a long period and the relationship analysis among subnets in an enterprise network were proposed for the retrieval of detailed and comprehensive traffic characteristics. The proposed *NetCube* multidimensional traffic analysis model can be effectively utilized to establish QoS policy and make a correct decision right on time to the network by providing detailed information on a dynamically and rapidly changing current network environment. We proved its practicality by applying it to a campus network.

For future research, we will utilize the analysis result of the proposed *NetCube* analysis model to establish a QoS policy on our campus network. We believe that this will contribute to network managers' understanding the traffic more comprehensively and accurately, and to realizing a more stable network environment. In addition, we plan to extend the *NetCube* model to include abnormal traffic and analyze traffic multidimensionally in terms of trend, point, layer and event. We also plan to make the *NetCube* more powerful by embedding various data-mining analysis techniques.

ACKNOWLEDGEMENT

This research was supported by a Korea University grant.

REFERENCES

1. Eittenberger PM, Krieger UR. A workbench for Internet traffic analysis. In *Proceedings of the MMB and DFT* 2012: *LNCS7201*, 2012; 240–243.
2. Chung JY, Choi Y, Park B, Hong JW-K. Measurement analysis of mobile traffic in enterprise networks. In *Proceedings of the 13th APNOMS*, 2011; 1–4.
3. Gebert S, Pries R, Schlosser D, Heck K. Internet access traffic measurement and analysis. In *Proceedings of the TMA 2012: LNCS7189*, 2012; 29–42.
4. Mauro D, Schmidt K. Essential SNMP (2nd edn). O'Reilly: Sebastopol, CA, 2005.
5. Chang K-D, Chen J-L, Chen C-Y, Chao H-C. IoT operations management and traffic analysis for Future Internet. In *Proceedings of the Computing, Communications and Applications Conference (ComComAp)*, 2012; 138–142.
6. Callado A, Kelner J, Sadok D, Alberto C, Fernandes S. Better network traffic identification through the independent combination of techniques. *Network and Computer Applications* 2012; **33**(4): 433–446.
7. Liao Q, Striegel A, Chawla N. Visualizing graph dynamics and similarity for enterprise network security and management. In *Proceedings of the Seventh International Symposium on Visualization for Cyber Security*, 2010; 34–45.
8. Liao M-Y, Luo M-Y, Yang C-S, Chen C-H, Wu P-C, Chen Y-C. Design and evaluation of deep packet inspection system: a case study. *IET Networks* 2012; **1**(1): 2–9.
9. Yoon S, Park J, Oh Y, Park J, Kim M. Internet application traffic classification using fixed IP-port. In *Proceedings of 12th Asia-Pacific Network Operations and Management Symposium: LNCS5787*, 2009; 21–30.
10. MRTG – The Multi Router Traffic Grapher, Tobi Oetiker 2012. Available: http://oss.oetiker.ch/mrtg/
11. PRTG - Paessler Router Traffic Grapher, Peassler AG 2012. Available: http://www.paessler.com/prtg
12. Snort 2012. Available: http://www.snort.org
13. Ntop 2012. Available: http://www.ntop.org
14. L7-filter 2012. Available: http://l7-filter.sourceforge.net/
15. RRDtool: Round Robin Database Tool, Tobi Oetiker 2012. Available: http://oss.oetiker.ch/rrdtool/
16. ACID 2012. Available: http://www.andrew.cmu.edu/user/rdanyliw/snort/snortacid.html
17. Xie G, Iliofotou M, Keralapura R, Faloutsos M, Nucci A. SubFlow: towards practical flow-level traffic classification. In *Proceedings of the IEEE INFOCOM*, 2012; 2541–2545.

DOI: 10.1002/nem

18. Asrodia P, Patel H. Network traffic analysis using Packet Sniffer. *International Journal of Engineering Research and Applications* 2012; **2**(3): 854–856.
19. Wang M, Feng J, Tsai M. A data warehousing approach to discover knowledge in peer-to-peer application. In *Proceedings of 22nd International Conference on Advanced Information Networking and Applications Workshop*, 2008; 1181–1186.
20. Mansmann F, Vinnik S. Interactive exploration of data traffic with hierarchical network maps. *IEEE Transactions on Visualization and Computer Graphics* 2006, **12**(6): 1440–1449.
21. Zhang B, Wu J, Liu J, Li C, Hu Y. Cluster based analysis of IPv6 network traffic. In *Proceedings of the 2nd International Conference on Networking and Distributed Computing*, 2011; 83–87.
22. Qi B, Dong YF. A new model of intrusion detection based on data warehouse and data mining. *Advanced Materials Research* 2012; **383–390**: 303–307.
23. Tanaksaranond G, Cheng T, Chow A, Santacreu A. Conceptual design of a star-schema OLAP to support multi-dimensional traffic analysis. In *Proceedings of the International Symposium on Spatial-Temporal Analysis and Data Mining*, July 2011.
24. Vokorokos L, Pekar A, Adam N. Preparing databases for network traffic monitoring. In *Proceedings of the IEEE 10th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 2012; 13–18.
25. Paloalto networks 2012. Available: http://www.paloaltonetworks.com
26. Han J, Kamber M. Data Mining: Concept and Techniques (3rd edn). Morgan Kaufmann: Burlington, MA, 2012; 105–145.
27. Schymik G, Corral K, Schuff D, St Louis R. Architecting a dimensional document warehouse. In *Proceedings of the International Conference on System Sciences*, 2007; 216–223.
28. Lee K, Chung Y, Kim M. An efficient method for maintaining data cubes incrementally. *Information Sciences* 2010; **180**(6): 928–948.
29. Zhang D, Zhai C, Han J. Topic cube: topic modeling for OLAP on multidimensional text databases. In *Proceedings of the International Conference on Data Mining*, Sparks, NV, April 2009.
30. Eavis T, Dimitrov G, Dimitrov I, Cueva D, Lopez A, Taleb A. Parallel OLAP with the Sidera server. *Future Generation Computer Systems* 2010; **26**(2): 259–266.
31. Romero O, Abello A. Automatic validation of requirements to support multidimensional design. *Data and Knowledge Engineering* 2010; **69**(9): 917–942.
32. Yu J, Lee H, Im Y, Kim M, Park D. Real-time classification of Internet application traffic using a hierarchical multi-class SVM. *KSII Transactions on Internet and Information Systems* 2010; **4**(5): 859–876.
33. Lee J, Moon J, Kim H. Examining the intellectual structure of records management and archival science in Korea with text mining. *Journal of Korean Society for Library and Information Science* 2007; **41**(1): 345–369 (in Korean).

AUTHORS' BIOGRAPHIES

Daihee Park received his B.S. degree in mathematics from Korea University, Korea, in 1982, and his Ph.D. degree in computer science from the Florida State University, USA, in 1992. He joined Korea University in 1993, where he is currently a Professor in the Dept. of Computer and Information Science. His research interests include data mining and intelligent database.

Jaehak Yu received his B.S. degree in Computer Science from Konkuk University in 2001. He received his M.S. and Ph.D. degrees in Computer Science from Korea University, Korea, in 2003 and 2010, respectively. He is currently a senior member of Electronics and Telecommunications Research Institute, Korea. His recent research interests include data and information analysis, intelligent network management, USN/IoT service platform, and network mining.

Jun-Sang Park received the B.S. and M.S. degree in computer science from Korea University, Korea, in 2008 and 2010, respectively. He is currently a Ph.D. candidate student of Korea University, Korea. His research interests include Internet traffic classification and network management.

Myung-Sup Kim received his B.S., M.S., and Ph.D. degree in Computer Science and Engineering from POSTECH, Korea, in 1998, 2000, and 2004, respectively. From September 2004 to August 2006 he was a postdoctoral fellow in the Department of Electrical and Computer Engineering, University of Toronto, Canada. He joined Korea University, Korea, in 2006, where he is working currently as an associate professor in the Department of Computer and Information Science. His research interests include Internet traffic monitoring and analysis, service and network management, and Internet security.

DOI: 10.1002/nem