WEB TRAFFIC CLASSIFICATION

SAMRUAY KAOPRAKHON

A RESEARCH PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE (COMPUTER SCIENCE) FACULTY OF GRADUATE STUDIES MAHIDOL UNIVERSITY 2009

WEB TRAFFIC CLASSIFICATION

SAMRUAY KAOPRAKHON 4836589 ITCS/M

M.Sc. (COMPUTER SCIENCE)

RESEARCH PROJECT ADVISORY COMMITTEE: VASAKA VISOOTTIVISETH, Ph.D., SUDSANGUAN NGAMSURIYAROJ, Ph.D., SIWARUK SIWAMOGSATHAM, Ph.D.

ABSTRACT

A wide use of the Internet introduces various online services such as online games, radio, music, TV and video clips, which communicate over Hyper Text Transfer Protocol (HTTP). These services maybe consume a large bandwidth. Therefore, if there are some employees using these services, the misuse could directly impact network bandwidth consumption. In this work, we aim to classify web traffic into three types: audio, video and normal web traffic. We propose a classification method based on the information flow. Our classification use a combination of keyword matching techniques and statistical behavior profiles. Keywords are pre-defined by observing from both audio and video traffic. Behavior profiles consist of three attributes, which are the average received packet size, a ratio of the number of server-client packets, and the flow duration. Each attribute has an independent threshold of mean (μ) and standard deviation (σ). The experimental results show that our method can classify audio, video and normal web traffic with high precision and recall.

KEY WORDS: WEB TRAFFIC CLASSIFICATION /VIDEO TRAFFIC/AUDIO TRAFFIC/KEYWORD MATCHING/BEHAVIOR PROFILE MATCHING/FLOW BASED CLASSIFICATION

128 pages.

การจำแนกประเภทข้อมูลเว็ป

WEB TRAFFIC CLASSIFICATION

สำรวย เกาประโคน 4836589 ITCS/M

วท.ม.(วิทยาการคอมพิวเตอร์)

คณะกรรมการที่ปรึกษาโครงการวิจัย : วัสกา วิสุทธิวิเศษ, Ph.D., สุดสงวน งามสุริยโรจน์, Ph.D., ศิวรักษ์ ศิวโมกษธรรม, Ph.D.

บทคัดย่อ

เนื่องจากปัจจุบันอินเทอร์เน็ตได้รับความนิยมอย่างแพร่หลาย ทำให้มีการพัฒนาการให้บริการต่าง ๆ บนเครือข่ายอินเทอร์เน็ตนั้นหลากหลายมากยิ่งขึ้น เช่น การให้บริการเล่นเกมส์ออนไลน์ การให้บริการวิทยุออนไลน์ การให้บริการฟังเพลงออนไลน์ การให้บริการดูโทรทัศน์ออนไลน์ และการให้บริการวีดีโอคลิปออนไลน์ โดยการให้บริการดังกล่าวสามารถให้บริการโดยใช้มาตรฐานของ Hyper Text Transfer Protocol (HTTP) ซึ่งบริการดังกล่าวอาจมีการบริโภค Bandwidth ในประมาณที่สูง ดังนั้นถ้ามีบางคนในองค์กรใช้บริการดังกล่าวโดยไม่ถูกกาละก็อาจจะทำให้เกิดผลกระทบต่อการทำงานของเพื่อนร่วมงานในองค์กรได้

ดังนั้นในโครงการนี้ผู้วิจัยมีวัตถุประสงค์เพื่อจำแนกข้อมูลบนเครือข่ายอินเทอร์เน็ตเป็นออกเป็นสามประเภทคือ ประเภทเว็ปไซต์ทั่วไป ประเภทเสียง และประเภทวีดีโอโดยนำเสนอวิธีการแบ่งแยกในระดับ flow และใช้เทคนิค keyword matching รวมกับ statistic behavior profile ซึ่งได้จากการเรียนรู้จากกลุ่มข้อมูลตัวอย่างของข้อมูลประเภทเสียงและวีดีโอ โดยในส่วนของ statistic behavior profile ผู้วิจัยได้เก็บสถิติ ค่าเฉลี่ย และค่าส่วนเบี่ยงเบนมาตรฐานของแต่ละคุณลักษณะจำนวนสามประการคือ ค่าเฉลี่ยของข้อมูลที่ได้รับ สัดส่วนของจำนวน packet ที่ได้รับกับจำนวนของ packet ที่ส่งและระยะเวลา จากผลการทดลองพบว่าการใช้เทคนิคดังกล่าวทำให้สามารถจำแนกประเภทของข้อมูลได้อย่างแม่นยำ

128 หน้า