HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Nguyễn Thị Tân

NGHIÊN CỨU VỀ ĐỐI SÁNH CHUỖI VÀ ỨNG DỤNG TRONG PHÂN TÍCH SÂU CÁC GÓI TIN

Chuyên ngành: Hệ thống thông tin

Mã số: 60.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI - 2013

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: TS. HOÀNG XUÂN DẬU

Phản biện 1: ……

Phản biện 2: ……

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: .. giờ ..... ngày ..... tháng .... năm .....

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

MỞ ĐẦU

Cùng với sự phát triển mạnh mẽ của công nghệ thông tin và hạ tầng mạng viễn thông, mạng Internet đã phát triển nhanh chóng và trở thành một phần không thể thiếu trong đời sống xã hội hiện đại. Rất nhiều ứng dụng trên nền Internet đã được phát triển và sử dụng rộng rãi như thư điện tử, diễn đàn, mạng xã hội, các mạng chia sẻ âm nhạc, phim, các ứng dụng lưu trữ và ứng dụng chuyên dùng trong kinh doanh.

Kèm theo các phần mềm hữu ích trên Internet, các phần mềm độc hại hoặc các phần mềm quảng cáo cũng phát triển và lan truyền nhanh chóng, như thư rác, các phần mềm độc hại với trẻ em, các phần mềm hoặc công cụ phục vụ tấn công, đột nhập trái phép. Việc phát hiện và ngăn chặn các ứng dụng độc hại và các hành vi tấn công, đột nhập trái phép, đảm bảo an toàn cho người dùng Internet là nhu cầu cấp thiết. Một trong các hướng giải quyết có hiệu quả là phân tích sâu nội dung các gói tin truyền trên mạng nhằm phát hiện sớm các nội dung độc hại cũng như các hành vi tấn công, đột nhập trái phép. Ưu điểm của phương pháp này là khả năng đảm bảo an toàn cho nhiều ứng

dụng, nhiều máy trạm trong mạng. Tuy nhiên, do lưu lượng thông tin truyền trên mạng thường rất lớn, nên việc phân tích nội dung một lượng rất lớn các gói tin là một thách thức thực sự, đặc biệt là phân tích trực tuyến.

Trong việc phân tích sâu nội dung các gói tin, công đoạn đối sánh chuỗi đóng vai trò quyết định. Ngoài việc đảm bảo tính chính xác trong đối sánh, vấn đề tốc độ xử lý cũng rất quan trọng do số lượng các gói tin cần xử lý thường rất lớn. Đề tài luận văn "*Nghiên cứu về đối sánh chuỗi và ứng dụng trong phân tích sâu các gói tin*" tập trung nghiên cứu, đánh giá các giải thuật đối sánh chuỗi. Trên cơ sở đó lựa chọn giải thuật phù hợp và ứng dụng trong mô hình phân tích sâu nội dung các gói tin. Cụ thể luận văn có cấu trúc như sau:

Chương 1- TỔNG QUAN VỀ ĐỐI SÁNH CHUỖI VÀ ỨNG DỤNG. Nghiên cứu tổng quan về đối sánh chuỗi và các ứng dụng của việc đối sánh chuỗi trên thực tế.

Chương 2 – CÁC THUẬT TOÁN ĐỐI SÁNH CHUỖI. Nghiên cứu các thuật toán đối sánh chuỗi chính xác thông dụng kèm theo phần đánh giá, so sánh giữa các thuật toán đối sánh.

Chương 3 - ỨNG DỤNG ĐỐI SÁNH CHUỖI TRONG PHÂN TÍCH SÂU GÓI TIN VÀ CÀI ĐẶT THỬ NGHIỆM. Giới thiệu tổng quan về việc phân tích sâu các gói tin, các ứng dụng của phân tích sâu gói tin và sử dụng các thuật toán đối sánh chuỗi vào việc phân tích sâu các gói tin. Từ đó cài đặt thuật toán để thử nghiệm và đánh giá kết quả.

Chương 1 – TỔNG QUAN VỀ ĐỐI SÁNH CHUỖI VÀ ỨNG DỤNG

Chương 1 trình bày tổng quan về đối sánh chuỗi và các ứng dụng của nó trong thực tế. Qua đó ta hiểu được một phần công việc của đối sánh chuỗi không những phục vụ những nhu cầu cơ bản của con người mà còn giúp con người tránh được những hành vi vi phạm trái phép.

1.1 Tổng quan về đối sánh chuỗi

1.1.1 Khái niệm về đối sánh chuỗi

Đối sánh chuỗi là việc so sánh một hoặc một vài chuỗi (thường được gọi là mẫu hoặc pattern) với văn bản để tìm nơi và số lần xuất hiện của chuỗi đó trong văn bản.

1.1.2 Lịch sử phát triển

1.1.3 Phân loại đối sánh chuỗi

1.1.3.1 Theo thứ tự đối sánh

Đối sánh chuỗi có thể được thực hiện theo các thứ tự sau:

- Từ trái sang phải
- Từ phải sang trái
- Đối sánh tại vị trí cụ thể

- Không theo thứ tự nhất định

1.1.3.2 Theo số lượng pattern

- Đối sánh chuỗi đơn pattern.
- Đối sánh chuỗi đa pattern

1.1.3.3 Theo độ sai khác đối sánh

- Đối sánh chuỗi chính xác
- Đối sánh chuỗi gần đúng

1.1.3.4 Theo sự thay đổi của pattern và văn bản

- Pattern thay đổi, văn bản cố định
- Pattern cố định, văn bản thay đổi
- Pattern thay đổi, văn bản thay đổi.

1.2 Ứng dụng của đối sánh chuỗi

1.2.1 Ứng dụng trong soạn thảo văn bản, thư viện số và công cụ tìm kiếm

1.2.2 Ứng dụng trong phát hiện đột nhập mạng

1.2.3 Ứng dụng trong Tin sinh học và nghiên cứu cấu trúc hóa học

1.3 Kết chương

Chương 1 trình bày tổng quan về đối sánh chuỗi và một số ứng dụng điển hình của đối sánh chuỗi. Đối sánh chuỗi được ứng dụng trong nhiều lĩnh vực như xử lý văn bản, tin sinh học và trong phát hiện đột nhập mạng. Ứng

dụng đối sánh chuỗi trong phát hiện đột nhập mạng cho phép sớm nhận dạng các chuỗi mẫu, các chữ ký của các tấn công, đột nhập và các phần mềm độc hại trong nội dung các gói tin truyền trên mạng. Chương 2 của luận văn đi sâu nghiên cứu các thuật toán đối sánh chuỗi thông dụng từ đó đánh giá hiệu năng thực hiện của từng thuật toán.

Chương 2 – CÁC THUẬT TOÁN ĐỐI SÁNH CHUỖI THÔNG DỤNG

Chương 2 đi sâu nghiên cứu các thuật toán đối sánh chuỗi, từ đó đánh giá được hiệu năng của từng thuật toán đối sánh chuỗi. Việc nghiên cứu các thuật toán và đánh giá hiệu năng của từng thuật toán đối sánh chuỗi là công việc quan trọng, từ đó ta có thể đưa ra quyết định việc lựa chọn thuật toán đối sánh chuỗi phù hợp trong từng bài toán cụ thể.

2.1 Tiêu chí đánh giá các thuật toán đối sánh chuỗi

Để đánh giá hiệu năng của thuật toán đối sánh chuỗi, chúng ta có thể dựa trên những tiêu chí sau:

- Số lần tìm kiếm
- Nén văn bản
- Độ phức tạp thời gian
- Tiêu chuẩn đối sánh
- Số pattern
- Sự biểu diễn kỹ thuật pattern

2.2 Các thuật toán đối sánh chuỗi chính xác thông dụng

2.2.1 Thuật toán Brute-Force

- T[0 .. n-1] là văn bản gồm n ký tự.
- P[0 .. m-1] là pattern gồm m ký tự, với điều kiện $m \leq n$

Thuật toán sẽ duyệt tìm P trên T từ vị trí 0 đến vị trí n-m, mỗi lần dịch chuyển P trên T một ký tự, như vậy độ dịch chuyển s sẽ lần lượt tăng thêm 1 qua mỗi lần đối sánh.

2.2.2 Thuật toán Rabin-Karp

- T[0 .. n-1] : là văn bản có n ký tự
- P[0 .. m -1]: là pattern có m ký tự với $m \leq n$
- t_s : là giá trị băm của chuỗi con tuần tự T[s .. s+m-1] trong T với độ dịch chuyển là s, trong đó $0 \leq s \leq n-m$
- p: là giá trị băm của P.

  Khi này thuật toán so sánh lần lượt giá trị t_s với p với s chạy từ 0 đến n-m, bước tiếp theo của thuật toán sẽ xảy ra với hai trường hợp như sau:

  - TH1: $t_s = p$, thực hiện phép đối sánh chuỗi giữa T[s .. s+m-1] và P[0.. m-1]
  - TH2: $t_s \neq p$, nếu $s \leq m$ tính gán s = s+1 và tính tiếp giá trị băm t_s.

2.2.3 Thuậtt toán Knuth-Morris-Pratt

- T[0 .. n-1] văn bản có n ký tự
- P[0 .. m-1] pattern có m ký tự với $m \leq n$

Thuật toán xác định vị trí dịch chuyển tiếp theo của P trong T được quyết định trên chính P mà vẫn làm cho phép đối sánh giữa P và T không thiếu xót bất kỳ xuất hiện nào của P trong T.

Đầu tiên, thuật toán tính được giá trị p[i] tương ứng với P[i] với $0 \leq i \leq m-1$ để xác định giá trị quyết định vị trí dịch chuyển tiếp theo của P trong T. Độ dịch chuyển tiếp theo của P trong T : s + (i – p[i]) với :

- s: độ dịch chuyển của P trong T ngay trước đó
- i: là ký tự thứ i đầu tiên trong P khi xảy ra $P[i] \neq T[s+i]$

Nếu $p[i] \neq -1$ thay bằng việc tiếp tục so sánh ký tự đầu tiên của P với T tại vị trí dịch chuyển s, thuật toán sẽ tiếp tục so sánh tại ký tự thứ p[i] của P với T.

2.2.4 Thuật toán Boyer-Moore

- T[0 .. n-1] văn bản có n ký tự
- P[0 .. m-1] pattern có m ký tự với $m \leq n$

Thuật toán duyệt các ký tự trong P từ phải qua trái, trong trường hợp không khớp (hoặc tìm thấy P trong T) nó sử dụng hai hàm tính lại giá trị để dịch chuyển P. Hai hàm dịch chuyển được dùng trong thuật toán gọi là phép dịch chuyển hậu tố tốt (good-suffix shift) hay còn gọi là phép dịch chuyển trùng khớp và dịch chuyển ký tự tồi (bad-character shift).

2.3 So sánh các thuật toán đối sánh chuỗi

Mỗi thuật toán đều đưa ra các phương pháp khác nhau để tìm kiếm pattern trong văn bản. Bảng 2.1 là tổng hợp sự khác biệt giữa các thuật toán đã trình bày trong mục 2.2.

Bảng 2.1 Sự khác biệt giữa các thuật toán

Tên thuật toán	Thứ tự đối sánh	Độ phức tạp tiền xử lý	Độ phức tạp đối sánh	Đặc điểm chính
Brute-Force	Không theo thứ tự nhất	Không thực hiện	O(mn)	Dịch chuyển từng kí tự một. Đây

	định.	việc tiền xử lý		không phải là một thuật toán tối ưu.
Rabin-Karp	Từ trái qua phải	O(m)	O(mn)	Sử dụng hàm băm, rất hiệu quả trong các thuật toán đối sánh đa pattern
Knuth-Morris-Pratt	Từ trái qua phải	O(m)	O(m+n)	Dựa vào chính pattern để quyết định bước dịch chuyển tiếp theo. Tăng khả năng thực thi, giảm độ trễ và thời gian đối sánh
Boyer	Từ phải	O(m)	O(mn)	Sử dụng hai

Moore	sang trái			hàm dịch chuyển là hậu tố tốt (good suffix) và ký tự tồi (bad character). Thuật toán cho kết quả tìm kiếm nhanh và được áp dụng nhiều trong thực tế.

2.4 Kết chương

Trong chương 2, luận văn đi sâu nghiên cứu các thuật toán đối sánh chuỗi chính xác thông dụng. Mỗi thuật toán có sự khác nhau về tiền xử lý dữ liệu từ đó quyết định đến bước dịch chuyển tiếp theo của pattern trong văn bản.

Về cơ bản, thuật toán Brute-Fore và Rabin-Karp đưa ra độ dịch chuyển của pattern trong văn bản là giống

nhau. Tuy nhiên, thuật toán Rabin-Karp tránh được việc đối sánh những chuỗi dài mà thay vào đó là sử dụng hàm hash để chuyển thuật toán về đối sánh mảng số nguyên, việc đối sánh chuỗi chỉ thực sự xảy ra khi xuất hiện các giá trị hash bằng nhau.

Thuật toán KMP và Boyer-Moore đều dựa trên pattern để quyết định bước dịch chuyển tiếp theo của pattern trên văn bản. Tuy nhiên, thứ tự đối sánh là khác nhau và biểu diễn thuật toán Boyer-Moore phức tạp hơn vì nó dựa trên hai quy tắc để dịch chuyển pattern trên văn bản. Thuật toán Boyer-Moore được ứng dụng nhiều trong thực tế như được cài đặt sẵn trong các bộ soạn thảo văn bản.

Chương 3 - ỨNG DỤNG ĐỐI SÁNH CHUỖI TRONG PHÂN TÍCH SÂU GÓI TIN VÀ CÀI ĐẶT THỬ NGHIỆM

Phân tích sâu các gói tin truyền trên mạng là một trong các biện pháp được sử dụng nhằm phát hiện sớm các dấu hiệu hoặc các hành vi tấn công, đột nhập hoặc sự lây lan các phần mềm độc hại. Chương này đi sâu nghiên cứu vấn đề phân tích sâu gói tin và ứng dụng của việc đối sánh chuỗi trong phân tích sâu các gói tin.

3.1 Tổng quan về phân tích sâu gói tin

3.1.1 Khái niệm phân tích sâu gói tin

Phân tích sâu gói tin (DPI - Deep Packet Inspection) là một giải pháp về phần mềm và phần cứng nhằm theo dõi luồng dữ liệu trên mạng và xác định các giao thức và ứng dụng, những địa chỉ web (URL) không thích hợp, phát hiện đột nhập và các phần mềm độc hại bằng việc phân tích kỹ các thành phần của các gói tin dữ liệu. Việc phân tích sâu gói tin giúp nhận dạng các dấu hiệu, các chuỗi đặc trưng, các chữ ký của các tấn công, đột nhập hoặc mã độc hại nhúng trong các gói tin gửi đến các dịch vụ và ứng dụng. Từ đó có thể giúp hệ thống bảo mật

gửi cảnh báo sớm, hoặc kịp thời ngăn chặn các tấn công, đột nhập hoặc sự lan truyền của các phần mềm độc hại.

3.1.2 Các ứng dụng của phân tích sâu gói tin

3.1.2.1 Ngăn chặn virus và các phần mềm độc hại

3.1.2.2 Phát hiện và ngăn chặn tấn công, đột nhập

3.1.2.3 Lọc URL

3.1.3 Thách thức trong việc phân tích sâu gói tin

Những yếu tố ảnh hưởng đến việc phân tích sâu gói tin trên mạng như:

- Độ phức tạp của thuật toán tìm kiếm.
- Số lượng chữ ký ngày càng tăng.
- Dữ liệu được mã hóa.
- Các vấn đề về phần cứng và phần mềm.

3.2 Ứng dụng đối sánh chuỗi trong phân tích sâu gói tin

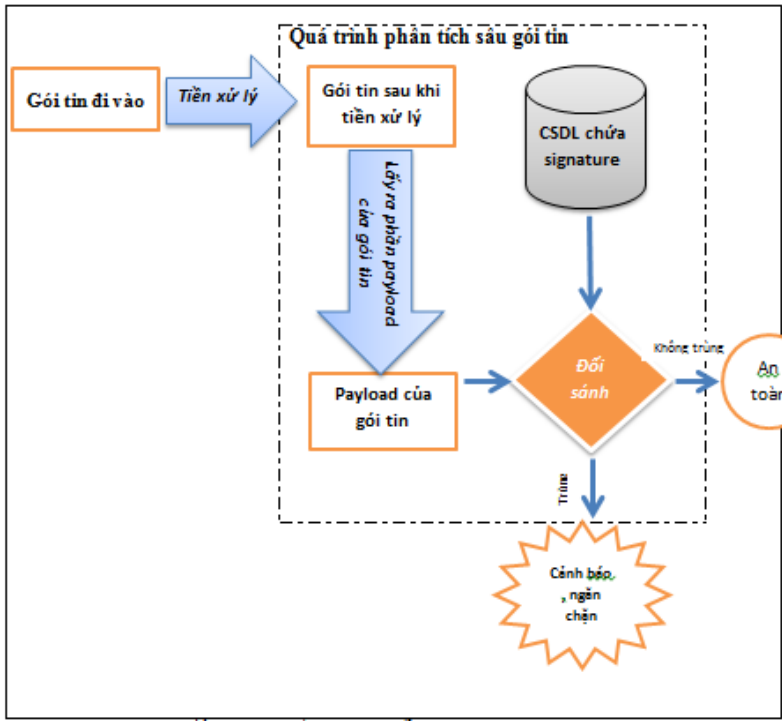
3.2.1 Phương pháp tiếp cận đối sánh chuỗi trong phân tích sâu gói tin

3.2.1.1 Phương pháp tiếp cận dựa trên máy trạng thái

3.2.1.2 Phương pháp tiếp cận dựa trên Heuristic

3.2.1.3 Phương pháp tiếp cận dựa trên lọc

3.2.2 Mô hình áp dụng thuật toán đối sánh chuỗi trong phân tích sâu gói tin



Hình 3.1 – Mô hình đối sánh chuỗi trong việc phân tích sâu gói tin

3.3 Cài đặt thuật toán, thử nghiệm và đánh giá kết quả

3.3.1 Tập CSDL sử dụng

Hai tập cơ sở dữ liệu các gói tin được sử dụng:

- Tập cơ sở dữ liệu chứa các signature. Trong tập này, các gói tin đã được gán nhãn chứa các loại đột nhập.
- Tập cơ sở dữ liệu kiểm thử: chứa những gói tin đã được chuẩn hóa với các signature trong CSDL dùng

để đối sánh với CSDL signature để đưa ra kết luận của việc phát hiện đột nhập.

Tập CSDL được sử dụng được trích từ tập CSDL KDD CUP 99.

3.3.1.1 Tổng quan về tập CSDL KDD CUP 99

3.3.1.2 Các thuộc tính của tập CSDL

3.3.1.3 Phân loại tấn công đột nhập trong tập CSDL

- Tấn công từ chối dịch vụ (DoS)
- Tấn công từ người dùng đến root (U2R)
- Tấn công truy cập từ xa đến nội bộ (R2L)
- Tấn công thăm dò (Probe)

Bảng 3.1 Danh sách các kiểu tấn công

DoS	U2R	R2L	Probe
back	buffer_overflow	guess_passwd	ipsweep
land	loadmodule	multihop	nmap
neptune	perl	phf	portsweep
pod	rootkit	spy	satan
smurf		warezclient	
teardrop		warezmaster	

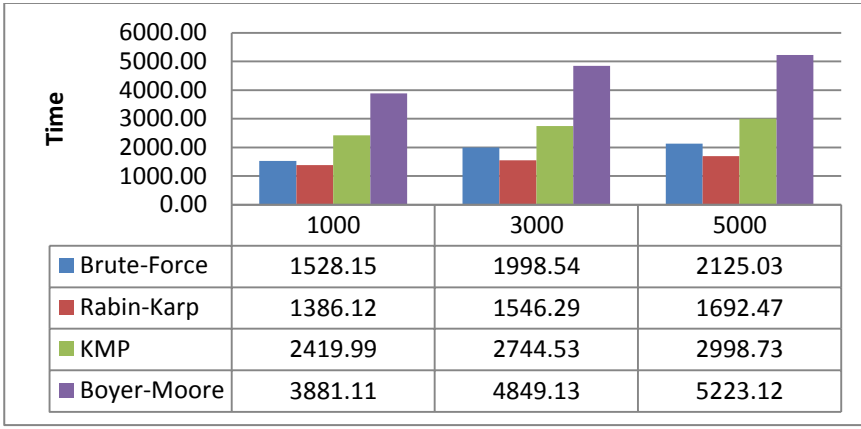
3.3.2 Cài đặt thuật toán và thử nghiệm

Chương trình thử nghiệm được phát triển bằng ngôn ngữ C++ trên nền Dev-C++ 4.9.9.2. Hệ thống máy

tính thử nghiệm có cấu hình: Bộ vi xử lý: Intel(R) Core(TM) i5-3210M CPU @ 2.50GHz 2.50GHz, RAM : 4.00 GB, Windows 7 Professional 32-bit .

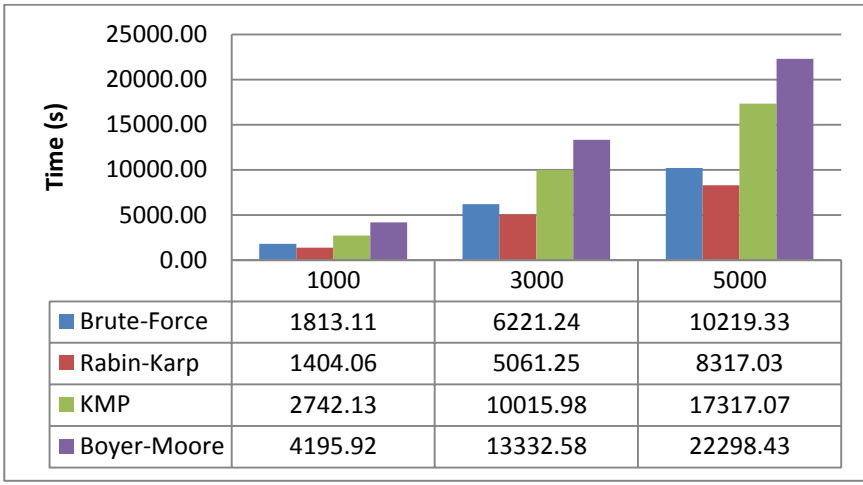
Thuật toán đối sánh sẽ lấy từng gói tin trong tập kiểm thử (tập số 1 và tập số 2) để đối sánh với tập chữ ký và đưa ra kết quả:

- Độ phức tạp thời gian chạy của từng thuật toán.
- Tổng số gói tin đột nhập và kiểu đột nhập.



	1000	3000	5000
Brute-Force	1528.15	1998.54	2125.03
Rabin-Karp	1386.12	1546.29	1692.47
KMP	2419.99	2744.53	2998.73
Boyer-Moore	3881.11	4849.13	5223.12

Hình 3.1 Biểu đồ so sánh hiệu năng các thuật toán đối sánh chuỗi với tập số 1



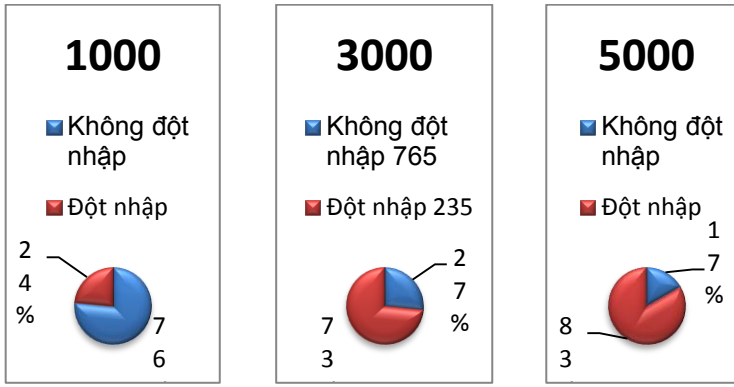
	1000	3000	5000
Brute-Force	1813.11	6221.24	10219.33
Rabin-Karp	1404.06	5061.25	8317.03
KMP	2742.13	10015.98	17317.07
Boyer-Moore	4195.92	13332.58	22298.43

Hình 3.2 Biểu đồ so sánh hiệu năng các thuật toán đối sánh chuỗi với tập số 2

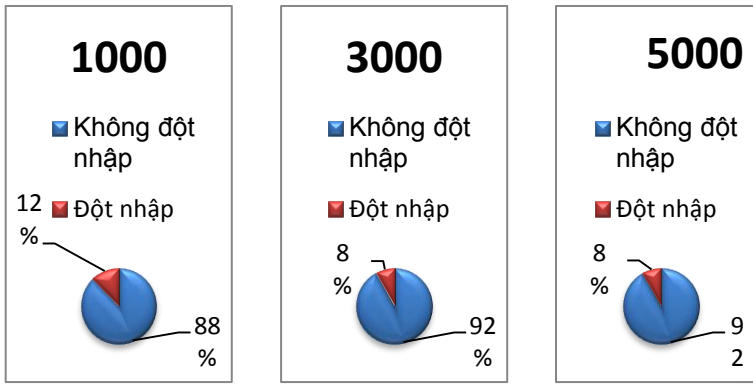
3.3.3 Đánh giá kết quả

Theo hình 3.5 - Biểu đồ so sánh hiệu năng thực hiện của thuật toán trên cùng một tập CSDL signature, ta có thể thấy độ phức tạp thời gian trong các trường hợp này phục thuộc chủ yếu vào các yếu tố sau:

- Số gói tin được phát hiện đột nhập.
- Thuật toán đối sánh chuỗi.



Hình 3.3 Biểu đồ so sánh hiệu năng các thuật toán đối sánh chuỗi với tập số 1



Hình 3.4 Tỷ lệ các gói tin được phát hiện đột nhập và không đột nhập trong tập số 2

3.4 Kết chương

Việc phân tích sâu các gói tin trên mạng đóng một vai trò quan trọng trong việc quản lý lưu lượng cũng như đảm bảo an ninh trên mạng. Do tốc độ mạng lên đến hàng Gb đã ảnh hưởng không nhỏ đến hiệu năng của việc phân tích sâu các gói tin. Khi lưu lượng gói tinlớn, việc phân tích gói tin sẽ làm tắc nghẽn mạng. Chính vì vậy việc đưa ra giải pháp về việc phân tích gói tin hiệu quả trong đó có việc nâng cấp các thuật toán đối sánh với tốc độ cao chúng ta cũng nâng cấp tốc độ phần cứng cũng như khả năng mở rộng của bộ nhớ.

Từ các thuật toán đối sánh chuỗi chính xác điển hình đã được nêu cụ thể ở chương 2 và áp dụng tập CSDL KDD'99 CUP, chương 3 đã đưa ra kết quả việc phát hiện các gói tin đột nhập cũng như hiệu năng thực hiện của các thuật toán.

KẾT LUẬN

Luận văn đi sâu nghiên cứu về đối sánh chuỗi và ứng dụng trong phân tích sâu nội dung các gói tin. Cụ thể, luận văn đã thực hiện được các nội dung sau:

- Nghiên cứu khái quát về đối sánh chuỗi, phân loại đối sánh chuỗi, và các ứng dụng của đối sánh chuỗi trong thực tế.
- Đi sâu nghiên cứu về các thuật toán đối sánh chuỗi thông dụng, từ đó đánh giá được hiệu năng của từng thuật toán.
- Nghiên cứu về việc phân tích sâu các gói tin, qua đó chúng ta có thể thấy rõ phân tích sâu gói tin không thể thiếu công đoạn đối sánh chuỗi. Nhờ việc đối sánh payload của gói tin với tập CSDL chứa các dấu hiệu được cho là gây hại đến hệ thống mạng, máy tính. Nếu tìm được sự xuất hiện của các signature trong CSDL trong payload, hệ thống có thể đưa ra cảnh báo hoặc ngăn chặn gói tin đó.
- Trên cơ sở lú thuyết về đối sánh chuỗi và phất tích sâu các gói tin, Luận văn đã cài đặt mô hình ứng dụng để đánh giá hiệu năng thực hiện của một số

thuật toán đối sánh chuỗi và phát hiện các gói tin đột nhập.

Trong tương lai, luận văn có thể được phát triển theo các hướng sau:

- Thực hiện việc bắt các gói tin trong thời gian thực trên mạng, tiền xử lý gói tin để trích chọn ra các payload phục vụ cho việc đối sánh tìm ra signature.
- Tìm hiểu sâu về các thuật toán đối sánh chuỗi đa pattern. Dạng thuật toán này cho phép ta có thể so sánh nhiều pattern trong cùng một lúc. Nó phù hợp hơn với tốc độ truyền gói tin trên mạng hiện nay.