(12) **BẢN MÔ TẢ SÁNG CHẾ THUỘC BẰNG ĐỘC QUYỀN SÁNG CHẾ**

(19) **Cộng hòa xã hội chủ nghĩa Việt Nam (VN)**
CỤC SỞ HỮU TRÍ TUỆ

(11) **1-0020334**

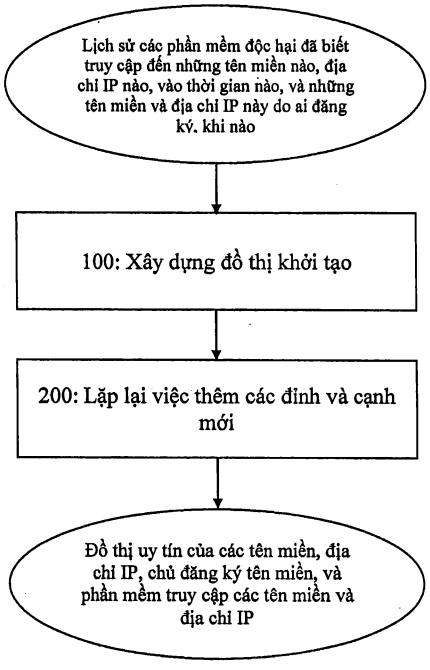
(51)[7] **H04L 29/06**, G06F 21/56

(13) **B**

(21) 1-2015-04690
(22) 09.12.2015
(45) 25.01.2019 370
(43) 25.10.2016 343
(73) VIỆN NGHIÊN CỨU CÔNG NGHỆ FPT - TRƯỜNG ĐẠI HỌC FPT (VN)
Số 8 Tôn Thất Thuyết, Mỹ Đình 2, Nam Từ Liêm, thành phố Hà Nội
(72) Nguyễn Minh Đức (VN)

(54) QUY TRÌNH XÂY DỰNG ĐỒ THỊ UY TÍN TRÊN KẾT NỐI INTERNET VÀ CẢNH BÁO PHẦN MỀM ĐỘC HẠI

(57) Sáng chế đề xuất hai quy trình có liên hệ mật thiết với nhau. Quy trình thứ nhất nhận đầu vào là lịch sử truy cập internet của các phần mềm độc hại đã biết, và đưa ra một đồ thị thể hiện mức độ uy tín của một lượng lớn các địa chỉ IP và tên miền, cũng như chủ đăng ký tên miền, gọi tắt là đồ thị uy tín. Quy trình thứ hai nhận đầu vào là đồ thị uy tín, cùng với hiện trạng truy cập internet của một máy tính, và cho đầu ra là cảnh báo rằng máy tính đang có phần mềm độc hại hay không, và nếu có thì phần mềm độc hại thuộc loại nào. Trong quy trình thứ nhất, việc xây dựng đồ thị uy tín được thực hiện bằng cách lặp lại việc thêm các đỉnh và các cạnh mới có quan hệ với các đỉnh và các cạnh đã có trên đồ thị thông qua một số mối quan hệ nhất định, đồng thời cũng dựa vào đặc điểm ngôn ngữ của các tên miền độc hại. Trong quy trình thứ hai, nếu máy tính có truy cập đến một trong các đỉnh trong đồ thị đã xây dựng trong quy trình thứ nhất, đưa ra cảnh báo máy tính chứa phần mềm độc hại. Ngoài ra, tuỳ theo loại của đỉnh trên đồ thị, có thể xác định loại độc hại của phần mềm độc hại được cảnh báo.



Lĩnh vực kỹ thuật được đề cập

Sáng chế đề cập đến những quy trình để từ các đầu vào là lịch sử truy cập internet của các phần mềm độc hại, cùng với hiện trạng truy cập internet của một máy tính để đưa ra đánh giá về việc máy tính đang có chứa phần mềm độc hại hay không.

Tình trạng kỹ thuật của sáng chế

Việc xác định liệu máy tính có chứa các phần mềm độc hại hay không là một trong các hoạt động quan trọng trong đảm bảo an ninh an toàn cho các hệ thống máy tính và hệ thống thông tin. Các phần mềm độc hại rất đa dạng và luôn biến đổi theo chiều hướng hoạt động ngày càng tinh vi để thoát khỏi sự phát hiện của các hệ thống an ninh an toàn thông tin, và một khi không bị phát hiện thì có thể gây ra nhiều loại thiệt hại khác nhau cho những người sở hữu hệ thống thông tin hay dữ liệu trên đó.

Các phương pháp truyền thống để nhận diện phần mềm độc hại thường là xác định các đặc điểm của các phần mềm này, sau khi các phần mềm này đã phát tác, rồi đưa các đặc điểm vào một bảng tra. Mỗi khi có một phần mềm mới cần kiểm tra, các hệ thống an ninh an toàn thông tin truyền thống so sánh các đặc điểm của phần mềm mới với các đặc điểm đã có trong bảng tra, để xác định xem phần mềm mới có độc hại hay không. Ví dụ, có thể đơn giản là đưa mã máy quan trọng của phần mềm độc hại đã biết qua một hàm băm, ví dụ hàm MD5, và lưu kết quả trong một cơ sở dữ liệu. Khi có phần mềm mới cần kiểm tra, có thể đưa mã máy của phần mềm mới qua cùng hàm băm, và so sánh kết quả với cơ sở dữ liệu, nếu có sự trùng lặp thì đưa ra cảnh báo phần mềm mới là độc hại. Hoặc một cách áp dụng phức tạp hơn là ghi nhận các đặc điểm trong hành vi, chứ không phải trong mã máy, của phần mềm độc hại, ví dụ như hành vi truy cập vào những khu vực lưu trữ trọng yếu trên máy tính, chụp lại màn hình, ghi lại nhật trình gõ

phím, vân vân, và lưu trữ các hành vi này vào trong cơ sở dữ liệu. Khi có phần mềm mới cần kiểm tra, xem xét lại các hành vi của phần mềm mới, và so sánh chúng với cơ sở dữ liệu hành vi độc hại đã biết, để đưa ra cảnh báo.

Các phương pháp trên có hạn chế là chỉ sau khi phần mềm độc hại đã phát tác ở một số máy tính nào đó rồi, thì bảng tra mới có được thông tin về đặc điểm của phần mềm độc hại đó; nghĩa là phương pháp này không phát hiện và ngăn chặn được những phần mềm độc hại có đặc điểm mới, chưa biết.

Hầu hết các phần mềm độc hại hiện đại đều có liên lạc thông tin đến những máy tính ở trên mạng Internet. Cách hoạt động phổ biến của chúng là, bằng một cách nào đó, có mặt trên máy tính nạn nhân, nhưng chưa có bất cứ hoạt động gì, để chờ thời cơ trong một thời gian dài. Khi có thông tin điều khiển từ một máy tính nhất định trên mạng Internet gửi tới chúng, chúng mới phát tác. Chẳng hạn, phần mềm độc hại có thể ở dạng các '*bot net*' tấn công từ chối dịch vụ vào một dịch vụ Web nào đó. Chúng có thể nằm rải rác trong rất nhiều máy tính ở nhiều nơi khác nhau trên mạng Internet, và không phát tác trong một thời gian dài. Đến khi có lệnh điều khiển gửi tới chúng từ một máy tính nhất định của kẻ tấn công, chúng đồng loạt truy cập vào dịch vụ Web nạn nhân, khiến cho dịch vụ Web nạn nhân bị quá tải.

Sáng chế được đề xuất trong hồ sơ mã số 1-2015-04422 đã nộp ở Cục Sở hữu Trí tuệ Việt Nam đã đề cập đến quy trình giúp phát hiện phần mềm độc hại, ngay cả khi các đặc điểm của chúng chưa được biết đến, bằng cách lưu trữ sẵn trong một cơ sở dữ liệu tên miền hoặc địa chỉ IP của máy tính mà chúng kết nối đến có thể được. Việc lưu trữ những tên miền hoặc địa chỉ IP độc hại có độ phức tạp nhỏ hơn so với lưu trữ các đặc điểm của phần mềm độc hại. Do đó, với một cơ sở dữ liệu nhỏ chứa những tên miền hoặc địa chỉ IP độc hại, có thể phát hiện ra một lượng lớn phần mềm độc hại, ngay cả khi đặc điểm của chúng là chưa biết.

Tuy nhiên, sáng chế trong hồ sơ mã số 1-2015-04422 đã nộp ở Cục Sở hữu Trí tuệ Việt Nam mới chỉ đề cập đến cách xây dựng cơ sở dữ liệu tên miền

hoặc địa chỉ IP của máy tính độc hại hoàn toàn dựa trên kết nối trong lịch sử từ phần mềm độc hại đã biết. Nếu có tên miền độc hại mới xuất hiện mà không hề có bất cứ một mối quan hệ nào với các tên miền độc hại khác, tức là chưa có phần mềm độc hại nào kết nối đến, và cũng không cùng chung những chủ tên miền với các phần mềm độc hại đã biết, thì tên miền độc hại mới cũng không thể được cập nhật vào cơ sở dữ liệu, do đó có nguy cơ làm cho việc cảnh báo phần mềm độc hại giảm hiệu quả.

Bản chất kỹ thuật của sáng chế

Sáng chế đề xuất hai quy trình, cùng hướng đến mục tiêu giúp phát hiện ra các phần mềm độc hại, ngay cả khi các đặc điểm mã máy hoặc hành vi trên máy tính chưa được biết đến trong các cơ sở dữ liệu. Quy trình thứ nhất nhận đầu vào là lịch sử các phần mềm độc hại đã biết truy cập đến những tên miền nào, địa chỉ IP nào, vào thời gian nào, và những tên miền và địa chỉ IP này do ai đăng ký trong khoảng thời gian nào, và cho ra một đồ thị có các đỉnh là các phần mềm độc hại, các tên miền hoặc địa chỉ IP và các chủ đăng ký, cùng với các cạnh là các mối quan hệ giữa chúng, gọi là đồ thị uy tín. Quy trình thứ hai nhận đầu vào là đồ thị được xây dựng theo quy trình thứ nhất, và hiện trạng truy cập internet của một phần mềm trên hệ thống máy tính, để đưa ra cảnh báo xem phần mềm này có độc hại không.

Quy trình thứ nhất được cải tiến từ quy trình thứ nhất của sáng chế trong hồ sơ mã số 1-2015-04422 đã nộp ở Cục Sở hữu Trí tuệ Việt Nam, trong đó tên miền độc hại mới được thêm vào đồ thị uy tín không chỉ nhờ vào mối quan hệ với các tên miền độc hại khác đã biết, như trong sáng chế trong hồ sơ mã số 1-2015-04422 đã nộp ở Cục Sở hữu Trí tuệ Việt Nam, mà còn dựa vào đặc điểm ngôn ngữ trong cách đặt tên miền.

Cụ thể, quy trình thứ nhất nhận đầu vào là lịch sử các phần mềm độc hại đã biết truy cập đến những tên miền nào, địa chỉ IP nào, vào thời gian nào, và những tên miền và địa chỉ IP này do ai đăng ký trong khoảng thời gian nào, và cho ra một đồ thị có các đỉnh là các phần mềm độc hại, các tên miền hoặc địa chỉ

IP và các chủ đăng ký, cùng với các cạnh là các mối quan hệ giữa chúng, gồm các bước:

đặt các đỉnh ứng với các phần mềm độc hại đã biết;

với mỗi đỉnh là phần mềm độc hại đã biết, thêm các đỉnh vào đồ thị, là các tên miền hoặc địa chỉ IP mà các phần mềm độc hại này có liên lạc đến, và thêm các cạnh là mối quan hệ liên lạc từ các phần mềm độc hại đến các tên miền hoặc địa chỉ IP này;

với mỗi tên miền mới xuất hiện trong một khoảng thời gian nhất định *T*, tính đến thời điểm hiện tại, mà chưa có trong đồ thị, xác định xem nó có liên quan đến phần mềm độc hại không, theo các bước:

bỏ đuôi bên phải ngoài cùng sau ký tự '.' và bỏ đi các ký tự '.' khỏi tên miền, để thu được tên miền rút gọn;

xác định bốn con số:

số thứ nhất là độ dài của tên miền rút gọn, bằng với số ký tự có trong tên miền rút gọn;

số thứ hai là entropy của tên miền rút gọn, bằng với:

$$-\sum_{i} P(x_i) \ln P(x_i)$$

trong đó x_i là ký tự thứ i trong tên miền rút gọn, $i = 1, 2, \ldots$, $P(x_i)$ là thương số của số lần xuất hiện của ký tự bằng với ký tự x_i trong tên miền rút gọn chia cho độ dài của tên miền rút gọn;

số thứ ba là tích vô hướng của véc tơ *n*-gram của tên miền rút gọn với véc tơ *n*-gram tham chiếu thứ nhất, và số thứ tư là tích vô hướng của véc tơ *n*-gram của tên miền rút gọn với véc tơ *n*-gram tham chiếu thứ hai, trong đó *n* là một số nguyên dương chọn trước và:

véc tơ *n*-gram của tên miền rút gọn là véc tơ gồm có các thành phần được tính bằng số lần xuất hiện của một chuỗi độ dài *n* ký tự nhất định - chuỗi *n* ký tự này tương ứng với thành phần đang xét - xuất hiện trong tên miền rút gọn;

véc tơ *n*-gram tham chiếu thứ nhất là một véc tơ có độ dài bằng với véc tơ *n*-gram của tên miền rút gọn, và có các thành phần được đặt bằng những giá trị cho trước;

véc tơ *n*-gram tham chiếu thứ hai là một véc tơ có độ dài bằng với véc tơ *n*-gram của tên miền rút gọn, và có các thành phần được đặt bằng những giá trị cho trước;

đưa bốn con số trên vào phân loại bằng kỹ thuật học máy thống kê có giám sát như phương pháp phân loại Rừng Ngẫu nhiên, để phân loại nó thuộc về loại bình thường hay là có liên hệ với phần mềm độc hại, sử dụng một mô hình Rừng Ngẫu nhiên được xây dựng sẵn, từ tập hợp bốn con số đặc trưng tính từ những tên miền rút gọn đã biết chính xác là bình thường hay là có liên hệ với phần mềm độc hại; nếu tên miền thuộc loại có liên hệ với phần mềm độc hại thì thêm vào thành đỉnh mới của đồ thị;

với mỗi tên miền hoặc địa chỉ IP, thêm các đỉnh là các chủ đăng ký đã từng đăng ký sử dụng trong quá khứ, kèm theo thuộc tính của các đỉnh này là thời gian đăng ký sử dụng, và các cạnh là mối quan hệ đăng ký từ các tên miền hoặc địa chỉ IP đến các chủ đăng ký.

lặp đi lặp lại các hoạt động sau, cho đến khi không thêm được đỉnh và cạnh mới vào đồ thị được nữa:

với mỗi đỉnh ứng với chủ đăng ký có trong đồ thị, thêm các đỉnh là các tên miền hoặc địa chỉ IP chưa có trong đồ thị nhưng cũng

đã từng được đăng ký bởi chủ đăng ký đang xét, và các cạnh là mối quan hệ đăng ký từ các tên miền hoặc địa chỉ IP mới thêm đến các chủ đăng ký đang xét;

với mỗi tên miền hoặc địa chỉ IP mới thêm ở trên, thêm các đỉnh là chủ đăng ký đã từng đăng ký sử dụng trong quá khứ, mà chưa có trong đồ thị, kèm theo thuộc tính của các đỉnh này là thời gian đăng ký sử dụng, và các cạnh là mối quan hệ đăng ký từ các tên miền hoặc địa chỉ IP đang xét đến các chủ đăng ký mới thêm.

Còn quy trình thứ hai nhận đầu vào là đồ thị được xây dựng theo quy trình thứ nhất, và hiện trạng truy cập internet của một phần mềm trên hệ thống máy tính, để đưa ra cảnh báo xem phần mềm này có độc hại không gồm các bước:

kiểm tra xem phần mềm có kết nối tới một trong các tên miền hay địa chỉ IP có trong đồ thị không; nếu có thì cảnh báo phần mềm độc hại.

Mô tả vắn tắt các hình vẽ

Hình 1 thể hiện một ví dụ của đồ thị uy tín, là đồ thị có các đỉnh là các phần mềm độc hại, các tên miền hoặc địa chỉ IP và các chủ đăng ký, cùng với các cạnh là các mối quan hệ giữa chúng;

Hình 2 là sơ đồ khối của các bước thực hiện quy trình thứ nhất, xây dựng đồ thị có các đỉnh là các phần mềm độc hại, các tên miền hoặc địa chỉ IP và các chủ đăng ký, cùng với các cạnh là các mối quan hệ giữa chúng;

Hình 3 là sơ đồ khối của các bước thực hiện quy trình thứ hai, cảnh báo xem một phần mềm có độc hại hay không.

Mô tả chi tiết sáng chế

Sáng chế đề xuất hai quy trình, có mối liên hệ mật thiết với nhau, trong việc giúp phát hiện phần mềm độc hại, trong đó quy trình thứ nhất được cải tiến từ quy trình thứ nhất được nêu trong sáng chế trong hồ sơ mã số 1-2015-04422 đã nộp ở Cục Sở hữu Trí tuệ Việt Nam.

Tương tự như trong sáng chế trong hồ sơ mã số 1-2015-04422 đã nộp ở Cục Sở hữu Trí tuệ Việt Nam, quy trình thứ nhất, dựa vào thông tin về danh sách các phần mềm độc hại đã biết, đã được công bố bởi các trang web hoặc dịch vụ phục vụ an toàn thông tin, xây dựng lại một đồ thị thể hiện quan hệ giữa các phần mềm độc hại, địa chỉ IP hoặc tên miền mà chúng kết nối đến, và với những cá nhân hoặc tổ chức đăng ký các tên miền này. Các cá nhân hay tổ chức đăng ký còn được gọi là các chủ đăng ký. Đồ thị này gọi là đồ thị uy tín.

Hình 1 thể hiện một ví dụ của đồ thị uy tín. Đồ thị có một số loại đỉnh sau:

phần mềm độc hại, thể hiện ở các đỉnh nằm dọc theo một cột ngoài cùng bên trái trên Hình 1;

các địa chỉ IP, thể hiện ở các đỉnh nằm dọc theo một cột ở ngay bên phải của cột ứng với các phần mềm độc hại, trên Hình 1;

các tên miền, thể hiện ở các đỉnh nằm dọc theo một cột ở ngay bên phải của cột ứng với các địa chỉ IP, trên Hình 1;

các tổ chức hoặc cá nhân đăng ký sử dụng các tên miền và địa chỉ IP, thể hiện ở các đỉnh nằm dọc theo một cột ngoài cùng bên phải trên Hình 1.

Các đỉnh được kết nối bằng các mối quan hệ giữa chúng. Ví dụ, trên Hình 1, "Phần mềm độc hại thứ nhất" có liên lạc tới "Tên miền thứ nhất", tên miền này có "Địa chỉ IP thứ nhất" và được đăng ký sử dụng bởi "Chủ đăng ký thứ nhất". "Chủ đăng ký thứ nhất" còn đăng ký sử dụng cả "Tên miền thứ hai" và "Địa chỉ IP thứ hai".

Nếu có một phần mềm mới, chưa biết có độc hại hay không, kết nối với "Địa chỉ IP thứ hai" là một địa chỉ IP cũng chưa từng được kết nối bởi bất cứ một phần mềm độc hại nào trước đây, thì vẫn có thể đưa ra cảnh báo phần mềm mới là độc hại, do "Địa chỉ IP thứ hai" được đăng ký sử dụng bởi cùng "Chủ đăng ký thứ nhất".

Như vậy quy trình thứ hai, theo dõi phần mềm kết nối đến tên miền hoặc địa chỉ IP nào, nếu tên miền hoặc địa chỉ IP đó thuộc vào đồ thị uy tín đã được xây dựng trước, thì cảnh báo phần mềm độc hại.

Tuy nhiên có điểm khác biệt quan trọng giữa quy trình thứ nhất được đề xuất, so với quy trình thứ nhất được nêu trong sáng chế trong hồ sơ mã số 1-2015-04422 đã nộp ở Cục Sở hữu Trí tuệ Việt Nam, đó là ở cách xây dựng đồ thị uy tín.

Hình 2 là sơ đồ khối của các bước thực hiện quy trình thứ nhất, xây dựng đồ thị uy tín. Sơ đồ khối này giống với sơ đồ thể hiện trên Hình 2 trong hồ sơ mã số 1-2015-04422 đã nộp ở Cục Sở hữu Trí tuệ Việt Nam. Khác biệt giữa sáng chế này với sáng chế trong hồ sơ mã số 1-2015-04422 đã nộp ở Cục Sở hữu Trí tuệ Việt Nam nằm ở bước 100, xây dựng đồ thị khởi tạo.

Trong bước 100, thực hiện:

đặt các đỉnh ứng với các phần mềm độc hại đã biết;

với mỗi đỉnh là phần mềm độc hại đã biết, thêm các đỉnh là các tên miền hoặc địa chỉ IP mà các phần mềm độc hại này có liên lạc đến, và thêm các cạnh là mối quan hệ liên lạc từ các phần mềm độc hại đến các tên miền hoặc địa chỉ IP này;

thêm các tên miền mới xuất hiện trong một khoảng thời gian nhất định *T*, ví dụ *T* là trong khoảng một năm tính đến thời điểm hiện tại, mà chưa có trong các tên miền đã được thêm vào đồ thị, và được xác định là tên miền độc hại theo quy trình nêu trong hồ sơ mã số 1-2015-04495 đã nộp ở Cục Sở hữu Trí tuệ Việt Nam;

với mỗi tên miền hoặc địa chỉ IP, thêm các đỉnh là các tổ chức hoặc cá nhân đã từng đăng ký sử dụng trong quá khứ, kèm theo thuộc tính của các đỉnh này là thời gian đăng ký sử dụng, và các cạnh là mối quan hệ đăng ký từ các tên miền hoặc địa chỉ IP đến các tổ chức hay cá nhân đăng ký.

Quy trình nêu trong hồ sơ mã số 1-2015-04495 đã nộp ở Cục Sở hữu Trí tuệ Việt Nam, giúp xác định xem một tên miền có tiềm năng có mối liên hệ với các phần mềm độc hại hay không, hoàn toàn chỉ dựa vào đặc điểm ngôn ngữ của tên miền đó. Tóm tắt quy trình nêu trong hồ sơ mã số 1-2015-04495 đã nộp ở Cục Sở hữu Trí tuệ Việt Nam là nhận đầu vào là tên miền và đưa ra đánh giá xem tên miền đó bình thường hay là có liên quan đến các phần mềm độc hại, gồm các bước:

bỏ đuôi bên phải ngoài cùng sau ký tự '.' và bỏ đi các ký tự '.' khỏi tên miền, để thu được tên miền rút gọn;

xác định bốn con số:

số thứ nhất là độ dài của tên miền rút gọn, bằng với số ký tự có trong tên miền rút gọn;

số thứ hai là entropy của tên miền rút gọn, bằng với:

$$-\sum_i P(x_i) \ln P(x_i)$$

trong đó x_i là ký tự thứ i trong tên miền rút gọn, $i = 1, 2, \ldots$, $P(x_i)$ là thương số của số lần xuất hiện của ký tự bằng với ký tự x_i trong tên miền rút gọn chia cho độ dài của tên miền rút gọn;

số thứ ba là tích vô hướng của véc tơ n-gram của tên miền rút gọn với véc tơ n-gram tham chiếu thứ nhất, và số thứ tư là tích vô hướng của véc tơ n-gram của tên miền rút gọn với véc tơ n-gram tham chiếu thứ hai, trong đó n là một số nguyên dương chọn trước và:

véc tơ n-gram của tên miền rút gọn là véc tơ gồm có các thành phần được tính bằng số lần xuất hiện của một chuỗi độ dài n ký tự nhất định - chuỗi n ký tự này tương ứng với thành phần đang xét - xuất hiện trong tên miền rút gọn;

véc tơ *n*-gram tham chiếu thứ nhất là một véc tơ có độ dài bằng với véc tơ *n*-gram của tên miền rút gọn, và có các thành phần được đặt bằng những giá trị cho trước;

véc tơ *n*-gram tham chiếu thứ hai là một véc tơ có độ dài bằng với véc tơ *n*-gram của tên miền rút gọn, và có các thành phần được đặt bằng những giá trị cho trước;

đưa bốn con số trên vào phân loại bằng kỹ thuật học máy thống kê có giám sát như phương pháp phân loại Rừng Ngẫu nhiên, để phân loại nó thuộc về loại bình thường hay là có liên hệ với phần mềm độc hại, sử dụng một mô hình Rừng Ngẫu nhiên được xây dựng sẵn, từ tập hợp bốn con số đặc trưng tính từ những tên miền rút gọn đã biết chính xác là bình thường hay là có liên hệ với phần mềm độc hại.

Đồ thị thu được sau bước 100 được gọi là đồ thị khởi tạo. Các bước tiếp theo được thực hiện như đã nêu trong quy trình thứ nhất trong hồ sơ mã số 1-2015-04422 đã nộp ở Cục Sở hữu Trí tuệ Việt Nam, để thu được đồ thị uy tín.

Hình 3 là sơ đồ khối của các bước thực hiện quy trình thứ hai, cảnh báo phần mềm mới có độc hại không và cập nhật đồ thị uy tín nếu cần. Từ đồ thị uy tín đã được xây dựng trong quy trình thứ nhất, quy trình thứ hai được thực hiện tương tự như trong quy trình thứ hai trong hồ sơ mã số 1-2015-04422 đã nộp ở Cục Sở hữu Trí tuệ Việt Nam.

Trong một phương án mở rộng để thực thi quy trình thứ hai, sau bước cập nhật đồ thị uy tín, cụ thể là sau khi:

thêm các đỉnh là các tên miền hoặc địa chỉ IP mà phần mềm độc hại mới được cảnh báo có liên lạc đến, và chưa có trong đồ thị uy tín, và thêm các cạnh là mối quan hệ liên lạc từ phần mềm độc hại mới được cảnh báo đến các tên miền hoặc địa chỉ IP này, vào đồ thị uy tín;

thì bổ sung tên miền, mà phần mềm độc hại mới được cảnh báo này có liên lạc đến, vào tập hợp tên miền có liên hệ với các phần mềm độc hại, dùng để xây dựng mô hình Rừng Ngẫu nhiên và dùng để tính ra véc tơ *n*-gram tham chiếu

thứ hai, dùng trong bước xác định tên miền có độc hại không dựa vào đặc điểm ngôn ngữ của tên miền đó.

Yêu cầu bảo hộ

1. Quy trình nhận đầu vào là lịch sử các phần mềm độc hại đã biết truy cập đến những tên miền nào, địa chỉ IP nào, vào thời gian nào, và những tên miền và địa chỉ IP này do ai đăng ký trong khoảng thời gian nào, và cho ra một đồ thị có các đỉnh là các phần mềm độc hại, các tên miền hoặc địa chỉ IP và các chủ đăng ký, cùng với các cạnh là các mối quan hệ giữa chúng, gồm các bước:

đặt các đỉnh ứng với các phần mềm độc hại đã biết;

với mỗi đỉnh là phần mềm độc hại đã biết, thêm các đỉnh vào đồ thị, là các tên miền hoặc địa chỉ IP mà các phần mềm độc hại này có liên lạc đến, và thêm các cạnh là mối quan hệ liên lạc từ các phần mềm độc hại đến các tên miền hoặc địa chỉ IP này;

với mỗi tên miền mới xuất hiện trong một khoảng thời gian nhất định T, tính đến thời điểm hiện tại, mà chưa có trong đồ thị, xác định xem nó có liên quan đến phần mềm độc hại không, theo các bước:

bỏ đuôi bên phải ngoài cùng sau ký tự '.' và bỏ đi các ký tự '.' khỏi tên miền, để thu được tên miền rút gọn;

xác định bốn con số:

số thứ nhất là độ dài của tên miền rút gọn, bằng với số ký tự có trong tên miền rút gọn;

số thứ hai là entropy của tên miền rút gọn, bằng với:

$$-\sum_{i} P(x_i) \ln P(x_i)$$

trong đó x_i là ký tự thứ i trong tên miền rút gọn, $i = 1, 2, \ldots$, $P(x_i)$ là thương số của số lần xuất hiện của ký tự bằng với ký tự x_i trong tên miền rút gọn chia cho độ dài của tên miền rút gọn;

số thứ ba là tích vô hướng của véc tơ n-gram của tên miền rút gọn với véc tơ n-gram tham chiếu thứ nhất, và số thứ

tư là tích vô hướng của véc tơ *n*-gram của tên miền rút gọn với véc tơ *n*-gram tham chiếu thứ hai, trong đó *n* là một số nguyên dương chọn trước và:

véc tơ *n*-gram của tên miền rút gọn là véc tơ gồm có các thành phần được tính bằng số lần xuất hiện của một chuỗi độ dài *n* ký tự nhất định - chuỗi *n* ký tự này tương ứng với thành phần đang xét - xuất hiện trong tên miền rút gọn;

véc tơ *n*-gram tham chiếu thứ nhất là một véc tơ có độ dài bằng với véc tơ *n*-gram của tên miền rút gọn, và có các thành phần được đặt bằng những giá trị cho trước;

véc tơ *n*-gram tham chiếu thứ hai là một véc tơ có độ dài bằng với véc tơ *n*-gram của tên miền rút gọn, và có các thành phần được đặt bằng những giá trị cho trước;

đưa bốn con số trên vào phân loại bằng kỹ thuật học máy thống kê có giám sát như phương pháp phân loại Rừng Ngẫu nhiên, để phân loại nó thuộc về loại bình thường hay là có liên hệ với phần mềm độc hại, sử dụng một mô hình Rừng Ngẫu nhiên được xây dựng sẵn, từ tập hợp bốn con số đặc trưng tính từ những tên miền rút gọn đã biết chính xác là bình thường hay là có liên hệ với phần mềm độc hại; nếu tên miền thuộc loại có liên hệ với phần mềm độc hại thì thêm vào thành đỉnh mới của đồ thị;

với mỗi tên miền hoặc địa chỉ IP, thêm các đỉnh là các chủ đăng ký đã từng đăng ký sử dụng trong quá khứ, kèm theo thuộc tính của các đỉnh này là thời gian đăng ký sử dụng, và các cạnh là mối quan hệ đăng ký từ các tên miền hoặc địa chỉ IP đến các chủ đăng ký;

lặp đi lặp lại các hoạt động sau, cho đến khi không thêm được đỉnh và cạnh mới vào đồ thị được nữa:

với mỗi đỉnh ứng với chủ đăng ký có trong đồ thị, thêm các đỉnh là các tên miền hoặc địa chỉ IP chưa có trong đồ thị nhưng cũng đã từng được đăng ký bởi chủ đăng ký đang xét, và các cạnh là mối quan hệ đăng ký từ các tên miền hoặc địa chỉ IP mới thêm đến các chủ đăng ký đang xét;

với mỗi tên miền hoặc địa chỉ IP mới thêm ở trên, thêm các đỉnh là chủ đăng ký đã từng đăng ký sử dụng trong quá khứ, mà chưa có trong đồ thị, kèm theo thuộc tính của các đỉnh này là thời gian đăng ký sử dụng, và các cạnh là mối quan hệ đăng ký từ các tên miền hoặc địa chỉ IP đang xét đến các chủ đăng ký mới thêm.

2.) Quy trình nhận đầu vào là đồ thị được xây dựng theo quy trình ở điểm 1, và hiện trạng truy cập internet của một phần mềm trên hệ thống máy tính, để đưa ra cảnh báo xem phần mềm này có độc hại không gồm các bước:

kiểm tra xem phần mềm có kết nối tới một trong các tên miền hay địa chỉ IP có trong đồ thị không; nếu có thì cảnh báo phần mềm độc hại.

3. Quy trình theo điểm 2, khác biệt ở chỗ, nếu phần mềm được cảnh báo là độc hại, có thêm bước:

bổ sung đỉnh ứng với phần mềm độc hại mới được cảnh báo vào đồ thị;

thêm các đỉnh là các tên miền hoặc địa chỉ IP mà phần mềm độc hại mới được cảnh báo có liên lạc đến, và chưa có trong đồ thị, và thêm các cạnh là mối quan hệ liên lạc từ phần mềm độc hại mới được cảnh báo đến các tên miền hoặc địa chỉ IP này;

thêm các đỉnh là các chủ đăng ký chưa có trong đồ thị, mà đã đăng ký sử dụng trong quá khứ các tên miền hoặc địa chỉ IP mới thêm vào đồ thị ở bước trên, và thêm các cạnh là mối quan hệ đăng ký từ các tên miền

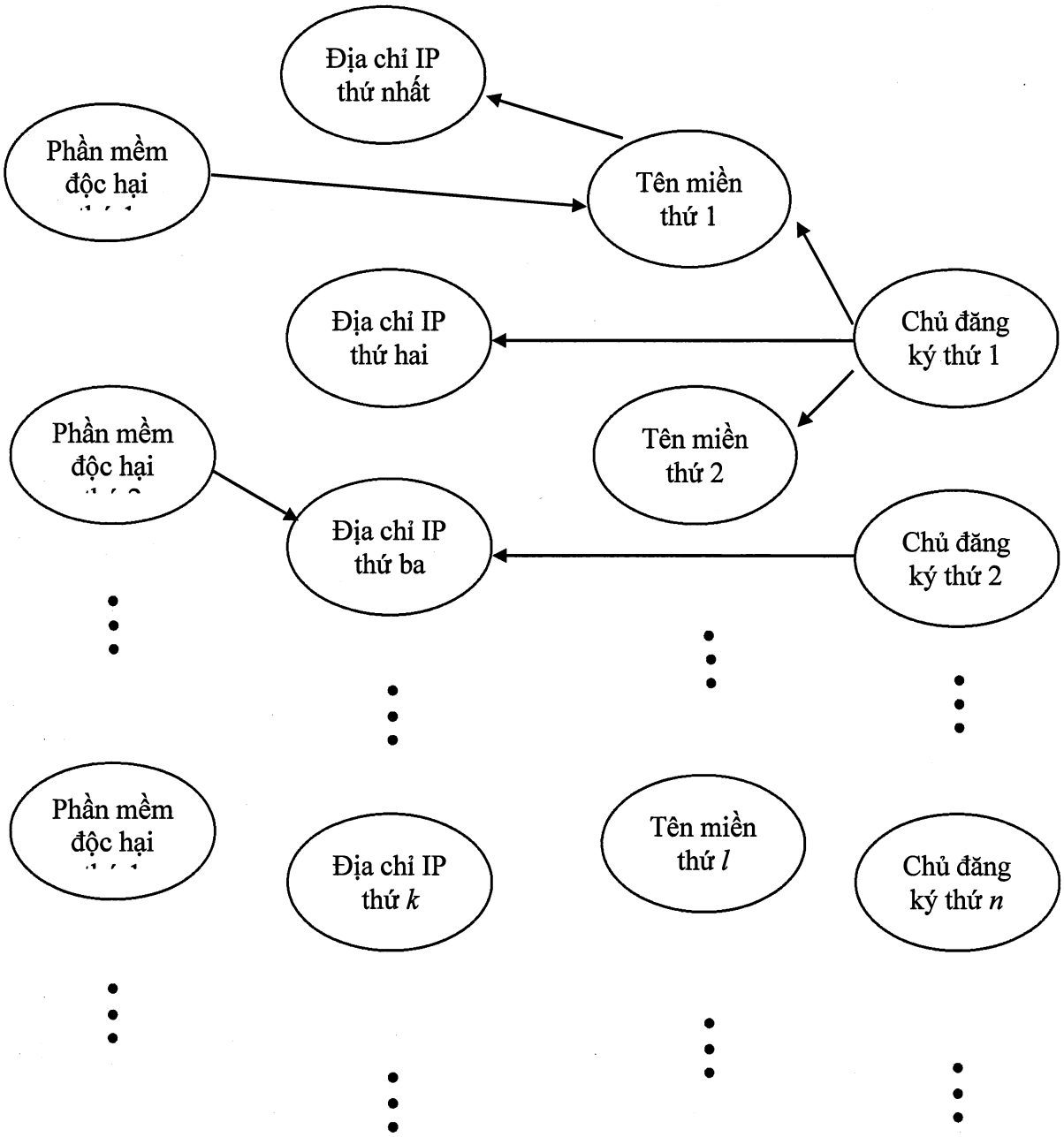
hoặc địa chỉ IP mới thêm vào đồ thị ở bước trên đến các các chủ đăng ký mới thêm;

lặp đi lặp lại các hoạt động sau, cho đến khi không thêm được đỉnh và cạnh mới vào đồ thị được nữa:

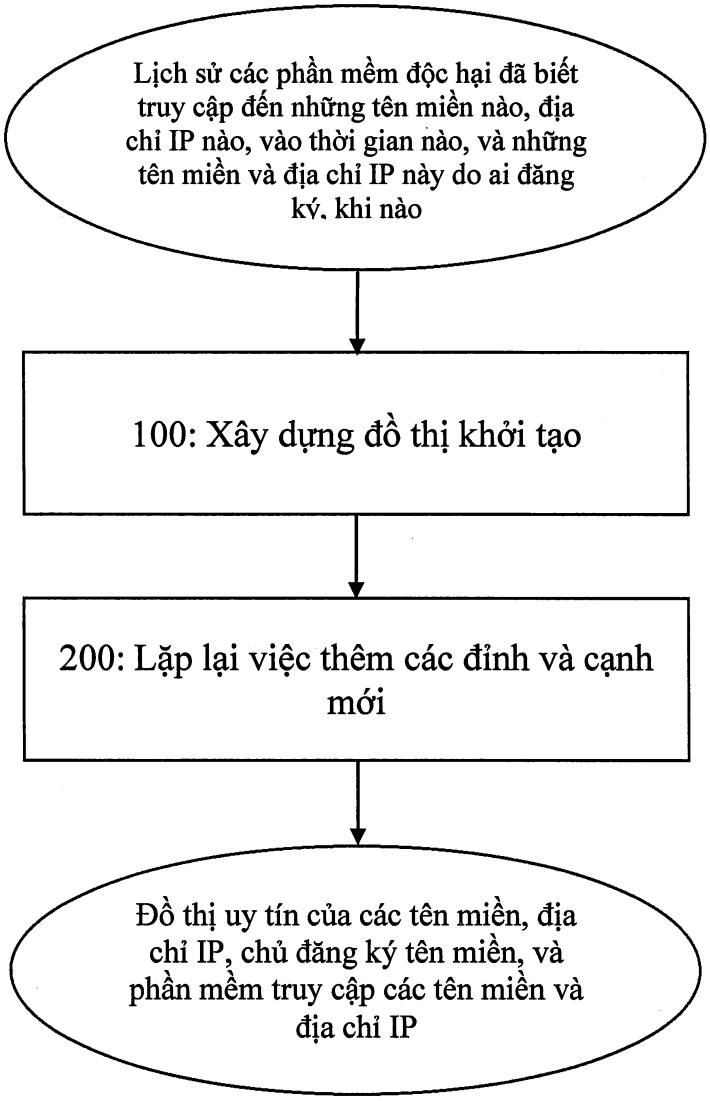
với mỗi đỉnh ứng với chủ đăng ký có trong đồ thị, thêm các đỉnh là các tên miền hoặc địa chỉ IP chưa có trong đồ thị nhưng cũng đã từng được đăng ký bởi chủ đăng ký đang xét, và các cạnh là mối quan hệ đăng ký từ các tên miền hoặc địa chỉ IP mới thêm đến các chủ đăng ký đang xét;

với mỗi tên miền hoặc địa chỉ IP mới thêm ở trên, thêm các đỉnh là chủ đăng ký đã từng đăng ký sử dụng trong quá khứ, mà chưa có trong đồ thị, kèm theo thuộc tính của các đỉnh này là thời gian đăng ký sử dụng, và các cạnh là mối quan hệ đăng ký từ các tên miền hoặc địa chỉ IP đang xét đến các chủ đăng ký mới thêm.

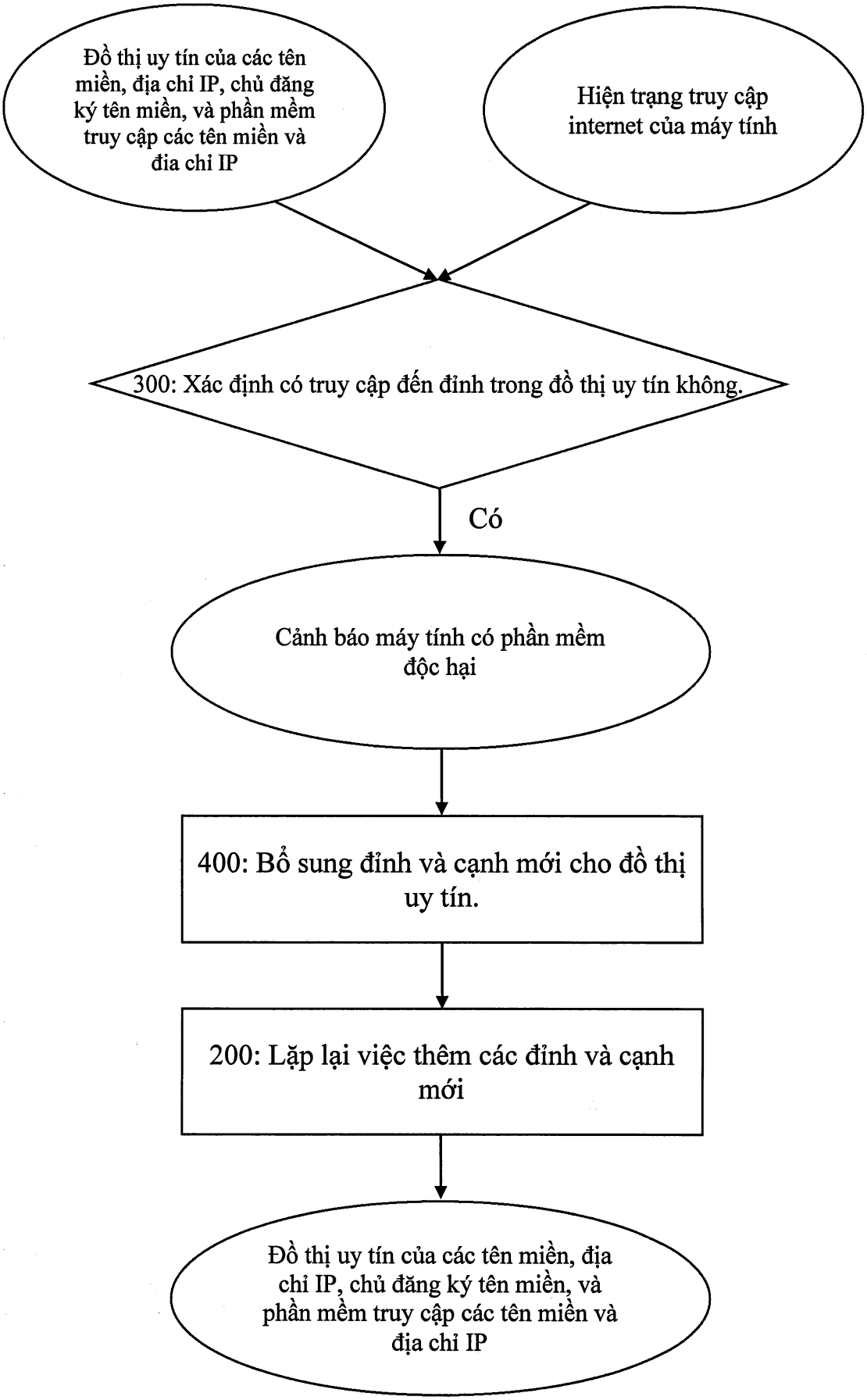
Hình 1


Địa chỉ IP thứ nhất
Phần mềm độc hại
Tên miền thứ 1
Địa chỉ IP thứ hai
Chủ đăng ký thứ 1
Phần mềm độc hại
Tên miền thứ 2
Địa chỉ IP thứ ba
Chủ đăng ký thứ 2
Phần mềm độc hại
Địa chỉ IP thứ *k*
Tên miền thứ *l*
Chủ đăng ký thứ *n*

Hình 2


Lịch sử các phần mềm độc hại đã biết truy cập đến những tên miền nào, địa chỉ IP nào, vào thời gian nào, và những tên miền và địa chỉ IP này do ai đăng ký, khi nào
100: Xây dựng đồ thị khởi tạo
200: Lặp lại việc thêm các đỉnh và cạnh mới
Đồ thị uy tín của các tên miền, địa chỉ IP, chủ đăng ký tên miền, và phần mềm truy cập các tên miền và địa chỉ IP

Hình 3


Đồ thị uy tín của các tên miền, địa chỉ IP, chủ đăng ký tên miền, và phần mềm truy cập các tên miền và đia chỉ IP
Hiện trạng truy cập internet của máy tính
300: Xác định có truy cập đến đỉnh trong đồ thị uy tín không.
Có
Cảnh báo máy tính có phần mềm độc hại
400: Bổ sung đỉnh và cạnh mới cho đồ thị uy tín.
200: Lặp lại việc thêm các đỉnh và cạnh mới
Đồ thị uy tín của các tên miền, địa chỉ IP, chủ đăng ký tên miền, và phần mềm truy cập các tên miền và địa chỉ IP