(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) **1-0047507**

(51)[2020.01] **H04W 8/18**

(13) **B**

(21) 1-2021-08364

(22) 24/12/2021

(45) 25/06/2025 447

(43) 25/02/2022 407A

(73) TẬP ĐOÀN CÔNG NGHIỆP - VIỄN THÔNG QUÂN ĐỘI (VN)
Lô D26 Khu đô thị mới Cầu Giấy, phường Yên Hoà, quận Cầu Giấy, thành phố Hà Nội

(72) ĐỖ ANH DŨNG (VN); NGUYỄN TRUNG KIÊN (VN); ĐỖ TIẾN ĐẠT (VN).

(74) Công ty TNHH NACILAW (NACILAW)

(54) PHƯƠNG PHÁP PHÁT HIỆN THUÊ BAO GIAN LẬN LẬU CƯỚC TRONG VIỄN THÔNG

(21) 1-2021-08364

(57) Sáng chế đề xuất phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông bao gồm: bước 1: thu thập, phân tích dữ liệu hồ sơ người dùng, bản tin mã hóa người dùng gian lận lậu cước; bước 2: trích chọn các đặc trưng của người dùng; bước 3: xây dựng mô hình phân loại thuê bao nghi ngờ gian lận lậu cước; bước 4: xây dựng bộ luật phát hiện thuê bao gian lận lậu cước trên bản tin mã hóa người dùng; bước 5: đề xuất các phương án chặn lọc thuê bao gian lận lậu cước thời gian thực.

Hình 1

Lĩnh vực kỹ thuật được đề cập

Sáng chế đề cập đến phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông. Cụ thể, phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông được ứng dụng để phát hiện và chặn lọc thuê bao điện thoại di động có hành vi gian lận, cụ thể là lậu cước.

Tình trạng kỹ thuật của sáng chế

Một vấn đề mà các nhà mạng trên toàn cầu đang phải đối mặt hàng năm là đối tượng gian lận viễn thông, hành vi này gây ảnh hưởng đến mảng doanh thu đăng ký đầu số thương hiệu. Trong đó, bài toán gian lận lậu cước là một ngách trong kiểu gian lận giả mạo tin nhắn, với tập đối tượng lậu cước lớn đang tác động trực tiếp tới các nhà mạng. Hệ thống phát hiện gian lận lậu cước viễn thông không chỉ giúp dịch vụ an toàn, tăng trải nghiệm cho khách hàng với các tin nhắn dịch vụ mà còn loại bỏ được các thất thu tài chính cho nhà mạng.

Trên thế giới đã có nhiều hệ thống phát hiện gian lận lậu cước trong viễn thông (fraud detection systems). Đây là một hệ thống được tối ưu hóa để có thể phát hiện được hành động lậu cước tại thời điểm tin nhắn, cuộc gọi vừa được đối tượng thực hiện (real-time detection). Trong các nghiên cứu gần đây, các hệ thống phát hiện gian lận lậu cước viễn thông có thể chia ra làm ba chiến lược chính: hệ thống dựa theo luật trên tri thức viễn thông, hệ thống phân tích quan sát phát hiện hành vi bất thường, hệ thống hồ sơ người dùng (user-profiling). Các hệ thống này, cũng được chia ra thành ba loại mô hình: thống kê (statistical), học máy (machine learning) và dựa trên luật (rule-based). Cách mô hình phân loại học quan sát được (supervised-learning) cũng cho kết quả phân loại chính xác như sử dụng các mô hình máy vec-tơ hỗ trợ, mạng nơ-ron nhân tạo, cây quyết định. Bên cạnh đó, nghiên cứu hướng tiếp cận dựa trên trích rút đặc trưng với mục tiêu rút ra được đặc trưng của đối tượng gian lận dựa trên các hành vi viễn thông hàng tuần, sau đó sử dụng mô hình

mạng nơ-ron nhân tạo hai chiều (feed-forward neural networks) cũng giành được sự quan tâm.

Tuy nhiên, trên thực tế hiệu quả phát hiện và chặn lọc các thuê bao gian lận lậu cước trên thị trường hiện nay không cao, bởi đối tượng sử dụng nhiều thiết bị, nhiều sim và nội dung tin nhắn lậu cước biến đổi đa dạng. Với chiến lược truyền thống dựa theo luật viễn thông (telcomunication rule-based) chỉ hiệu quả trong thời gian đầu, về lâu dài đối tượng liên tục thay đổi hành vi khiến hệ thống bị lỗi thời (out-of-dated) tỉ lệ phát hiện chính xác thấp. Ngoài ra, một số cách tiếp cận chặn tin nhắn hoặc sử dụng một tập thuê bao lậu cước dựa trên tập thuê bao gian lận đen (fraud black list) và chặn các tin gửi từ số trong tập này không thể vét cạn được hết đối tượng.

Nhận thấy các hạn chế trong các phương pháp truyền thống trên thị trường, sáng chế sử dụng chiến lược tiếp cận dựa trên phân tích đặc trưng hồ sơ người dùng (user-profiling). Đây là hướng tiếp cận đang dành được nhiều sự quan tâm trong các hệ thống phát hiện người dùng bất thường, bởi tính hiệu quả và sức mạnh của các công nghệ khai phá dữ liệu lớn giúp mô hình có thể tự học được các đặc trưng của đối tượng gian lận lậu cước. Bên cạnh đó, với cơ sở hạ tầng xử lý dữ liệu lớn và tính toán hiệu năng cao, việc tận dụng khai phá văn bản (text mining) với các kỹ thuật xử lý ngôn ngữ tự nhiên (nature language processing) cũng đề xuất các luật từ khóa gian lận lậu cước để tăng hiệu quả phát hiện cho hệ thống.

Bản chất kỹ thuật của sáng chế

Mục đích của sáng chế là đề xuất phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông giải quyết các hạn chế nêu ra ở trên.

Sáng chế đưa ra một hệ thống phát hiện thời gian thực (Real-time detection) thuê bao có hành vi gian lận lậu cước có hiệu quả với bộ dữ liệu người dùng viễn thông các nhà mạng tại Việt Nam và một số nước trên thế giới. Giải pháp được đưa ra nhằm khắc phục các hệ thống phát hiện hiện nay, dựa trên ba kỹ thuật : dựa theo luật (rule-based) trên tri thức viễn thông, dựa theo nội dung (content-based) hành vi lậu cước và khai phá dữ liệu hồ sơ người dùng. Hệ thống đảm bảo phát hiện chính xác thuê bao có hành vi gian lận, lậu

cước, giúp nhà mạng có các tác động phù hợp tới tập đối tượng này, nhằm tăng trải nghiệm cho người dùng dịch vụ.

Cụ thể, sáng chế đề xuất phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông bao gồm:

Bước 1: thu thập, phân tích dữ liệu hồ sơ người dùng, bản tin tin nhắn được mã hóa gian lận lậu cước. Trong đó, phân hệ xử lý dữ liệu lớn tự động thu thập dữ liệu lịch sử hành vi sử dụng viễn thông của thuê bao trong một khoảng thời gian dài (sáu tháng đến một năm), dữ liệu được mã hóa đảm bảo an toàn thông tin, làm sạch dữ liệu, được tổng hợp các đặc trưng, xây dựng bộ dữ liệu hồ sơ người dùng qua các công nghệ xử lý dữ liệu lớn. Bên cạnh đó, phân hệ xử lý dữ liệu lớn cũng thực hiện xử lý bản tin mã hóa người dùng (message of encrypted user) lậu cước và thông thường ở dạng văn bản, phân đoạn, làm sạch, tiền xử lý dữ liệu. Các bộ dữ liệu hành vi viễn thông, đặc trưng hồ sơ người dùng và bản tin mã hóa người dùng được chuẩn bị sẵn sàng cho bước tiếp theo;

Bước 2: trích chọn các đặc trưng. Bước này thực thi đồng thời trích chọn các đặc trưng hồ sơ người dùng và đặc trưng từ khóa lậu cước từ bản tin mã hóa người dùng trên phân hệ xử lý dữ liệu lớn. Với bộ đặc trưng hồ sơ người dùng được tổng hợp từ dữ liệu bản ghi chi tiết CDR (Call Detail Record) cuộc gọi và tin nhắn, thông qua mô hình trích chọn đặc trưng đưa ra bộ đặc trưng tối ưu có ý nghĩa trong việc xác định người dùng có hành vi gian lận lậu cước. Bên cạnh đó, các kỹ thuật xử lý ngôn ngữ tự nhiên (Nature Language Processing) cũng được sử dụng để khai thác các đặc trưng về mặt ngữ nghĩa gồm từ khóa lậu cước, tần suất từ khóa. Các đặc trưng đầu ra trong bước này sẵn sàng chuẩn bị cho mô hình phân loại người dùng lậu cước và tập luật lậu cước cho bước tiếp theo;

Bước 3: xây dựng mô hình phân loại thuê bao nghi ngờ gian lận lậu cước. Trong bước này, mô hình phân loại thuê bao gian lận lậu cước được huấn luyện dựa trên các đặc trưng hồ sơ người dùng được trích chọn ở bước 2. Đầu ra là tập thuê bao nghi ngờ được bổ sung theo từng ngày;

Bước 4: xây dựng bộ luật phát hiện thuê bao gian lận lậu cước trên bản tin mã hóa người dùng. Trong bước này, được thực hiện trên phân hệ thời gian thực, tập thuê bao nghi ngờ gian lận lậu cước trong bước 3, được qua một bộ lọc dựa trên các luật từ khóa và đặc

trưng của bản tin mã hóa người dùng khi có hành vi gửi tin nhắn tại thời điểm hiện tại. Đầu ra xác nhận thuê bao có hành vi gian lận lậu cước có độ chính xác cao. Tập thuê bao được xác nhận có gian lận lậu cước được đẩy ra cho bước tiếp theo;

Bước 5: đề xuất các phương án chặn lọc thuê bao gian lận lậu cước thời gian thực. Bước cuối cùng, được xử lý trên phân hệ nghiệp vụ, với các thuê bao được phát hiện gian lận lậu cước, sẽ được đưa lên phía xác nhận nghiệp vụ, sáng chế cũng đề xuất các giải pháp ra quyết định cho việc chặn thuê bao thời gian thực khác nhau. Bên cạnh đó, các luật lậu cước tĩnh và động xây dựng ở bước 4, cũng được đề xuất để phía nghiệp vụ thực hiện thống kê, điều chỉnh, tăng tính linh hoạt cho hệ thống.

Trong sáng chế này, ở bước 1, giúp sáng chế đạt được bộ dữ liệu đặc trưng đa dạng bằng cách tập trung vào cả dữ liệu hồ sơ người dùng lẫn bản tin mã hóa người dùng, dựa trên các kỹ thuật khai phá dữ liệu lớn giúp cho hệ thống có nguồn dữ liệu đầy đủ về số lượng đặc trưng lẫn số lượng người dùng, có thể bao quát được hành vi của đối tượng gian lận lậu cước và người dùng thông thường;

Trong sáng chế này, ở bước 2, sáng chế đạt được bộ đặc trưng tối ưu đảm bảo không bị đặc trưng nhiễu và tăng hiệu suất của mô hình phát hiện. Thông qua sử dụng trích chọn đặc trưng theo phương pháp mô hình bao trùm (wrapper method), với mô hình hồi quy đa biến thích ứng đảm bảo bắt được mối quan hệ phức tạp, phi tuyến tính giữa các đặc trưng với đặc điểm người dùng gian lận lậu cước và có tốc độ thực thi nhanh. Tiếp đó qua việc xếp hạng mức độ quan trọng của các đặc trưng và đánh giá tương quan, cho ta một tập con đặc trưng hồ sơ người dùng tối ưu cho mô hình phân loại thuê bao gian lận lậu cước;

Trong sáng chế này, ở bước 3, thực hiện tối ưu mô hình phân loại rừng quyết định ngẫu nhiên (random forest) để xác định thuê bao gian lận lậu cước. Các siêu tham số (hyperparameter) gồm số lượng cây quyết định (number of decision tree), bậc tối đa của cây (maximum depth), số lượng lá tối đa (maximum of leaf nodes) được lựa chọn dựa trên kiểm chứng chéo (cross validation) đảm bảo mô hình có độ chính xác cao và tránh hiện tượng quá khớp (overfitting). Tiêu chuẩn Gini cho từng nút cây quyết định cũng được sử dụng để tăng tốc độ tính toán. Bên cách đó, sử dụng thêm các mô hình vệ tinh phát hiện

người giao hàng, cuộc gọi tự động, người bán hàng chuyên nghiệp, để loại bớt các thuê bao hành vi khác ra khỏi tập thuê bao nghi ngờ gian lận lậu cước;

Trong sáng chế này, ở bước 5, đưa ra các giải pháp chặn lọc đối tượng gian lận lậu cước khác nhau phù hợp với đặc thù các hệ thống với bốn phương án chặn lọc: chặn làm chậm tin, chặn tin, chặn thuê bao khoảng thời gian và chặn thuê bao một chiều áp dụng phù hợp với mức độ của từng đối tượng. Từ đó đem lại hiệu quả chặn lọc cho các hệ thống.

Mô tả vắn tắt các hình vẽ

Hình 1 là sơ đồ hoạt động của giải pháp phát hiện thuê bao gian lận lậu cước;

Hình 2 là sơ đồ chi tiết hoạt động của các phân hệ trong giải pháp phát hiện và chặn lọc thuê bao gian lận lậu cước viễn thông.

Mô tả chi tiết sáng chế

Sáng chế được mô tả chi tiết dưới đây có dựa vào các hình vẽ kèm theo, các hình vẽ này nhằm mục đích minh họa các phương án của sáng chế mà không làm giới hạn phạm vi bảo hộ sáng chế.

Trong bản mô tả sáng chế này, "thuê bao" được hiểu là thuê bao thực hiện gửi tin nhắn sms, "người dùng" chỉ người dùng viễn thông nói chung, thực hiện cả hành vi gửi và nhận tin nhắn. Phương pháp phát hiện và chặn lọc cuộc gọi rác bao gồm các bước sau: Bước 1: thu thập, phân tích dữ liệu hồ sơ người dùng, bản tin mã hóa người dùng gian lận lậu cước. Đầu ra là dữ liệu hồ sơ người dùng và dữ liệu bản tin mã hóa người dùng. Bước 2: trích chọn các đặc trưng của người dùng. Đầu ra là các đặc trưng (feature) tối ưu. Bước 3: xây dựng mô hình phân loại thuê bao nghi ngờ gian lận lậu cước. Đầu ra là tập thuê bao đã được gán nhãn nghi ngờ. Bước 4: xây dựng bộ luật phát hiện thuê bao gian lận lậu cước trên nội dung sms. Đầu ra là các thuê bao xác nhận là có hành vi gian lận lậu cước. Bước 5: đề xuất các phương án chặn lọc thuê bao gian lận lậu cước thời gian thực.

Hệ thống gồm ba phân hệ: phân hệ xử lý dữ liệu lớn, phân hệ xử lý thời gian thực và phân hệ nghiệp vụ. Trung tâm của hệ thống là toàn bộ hạ tầng dữ liệu lớn là nơi lưu trữ và xử lý dữ liệu, cầu nối giữa ba phân hệ. Từ tầng cơ sở dữ liệu viễn thông, thu thập các bộ dữ liệu về hành vi viễn thông và bản tin mã hóa người dùng, dữ liệu người dùng được mã

hóa đảm bảo an toàn thông tin. Luồng xử lý dữ liệu lớn và mô hình học máy được thực thi, kết hợp với luồng phát hiện gian lận lậu cước tại phân hệ thời gian thực. Dữ liệu của hệ thống được lưu trữ tại hạ tầng dữ liệu lớn - nơi kết nối với phân hệ nghiệp vụ gồm hệ thống máy chủ tương tác giữa hệ thống mà quản trị đưa ra quyết định chặn lọc thuê bao gian lận lậu cước.

Phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông bao gồm:

Bước 1: thu thập, phân tích dữ liệu hồ sơ người dùng, bản tin mã hóa người dùng (message of encrypted user) gian lận lậu cước. Trong đó, phân hệ xử lý dữ liệu lớn (Hình 1) tự động thu thập dữ liệu lịch sử hành vi sử dụng viễn thông của thuê bao trong một khoảng thời gian dài (6 tháng đến 1 năm), dữ liệu được mã hóa đảm bảo an toàn thông tin, làm sạch dữ liệu, được tổng hợp các đặc trưng, xây dựng bộ dữ liệu hồ sơ người dùng qua các công nghệ xử lý dữ liệu lớn. Bên cạnh đó, phân hệ cũng thực hiện xử lý bản tin mã hóa người dùng lậu cước và thông thường ở dạng văn bản, phân đoạn, làm sạch, tiền xử lý dữ liệu. Các bộ dữ liệu hành vi viễn thông, đặc trưng hồ sơ người dùng và bản tin mã hóa người dùng được chuẩn bị sẵn sàng cho bước tiếp theo. Dữ liệu hành vi sử dụng viễn thông bao gồm : dữ liệu lưu lượng cuộc gọi - tin nhắn hai chiều, dữ liệu đăng ký dịch vụ, dữ liệu sử dụng 3G/4G,... Dữ liệu hồ sơ người dùng có các đặc trưng như: hành vi sử dụng nhiều thiết bị, lịch sử vị trí và di chuyển, mối quan hệ giữa các người dùng, tần suất thực hiện cuộc gọi trong khoảng thời gian,...Dữ liệu bản tin mã hóa người dùng gồm các đoạn tin nhắn được gán nhãn lậu cước và nhãn thông thường ứng với hai tập đối tượng. Việc phân tích hồ sơ người dùng sử dụng các kỹ thuật phân tích dữ liệu như mô hình thống kê (statistical model), phân tích đồ thị (graph analysis), phân tích dòng thời gian (time-series analysis). Đối với bộ dữ liệu lớn (của 120 triệu người dùng với dung lượng khoảng 100 TB) được lưu trữ và xử lý trong phân hệ xử lý dữ liệu lớn với công nghệ lưu trữ và thực thi song song thời gian thực (Hadoop, Spark).

Bước 2: trích chọn các đặc trưng của người dùng. Bước này thực thi đồng thời trích chọn các đặc trưng hồ sơ người dùng và đặc trưng từ khóa lậu cước từ bản tin mã hóa người dùng trên phân hệ xử lý dữ liệu lớn. Với bộ đặc trưng hồ sơ người dùng được tổng hợp từ dữ liệu bản ghi chi tiết CDR (Call Detail Record) cuộc gọi và tin nhắn, thông qua mô hình

trích chọn đặc trưng đưa ra bộ đặc trưng tối ưu có ý nghĩa trong việc xác định người dùng có hành vi gian lận lậu cước. Bên cạnh đó, các kỹ thuật xử lý ngôn ngữ tự nhiên (nature language processing) cũng được sử dụng để khai thác các đặc trưng về mặt ngữ nghĩa gồm từ khóa lậu cước, tần suất từ khóa. Các đặc trưng đầu ra này, sẵn sàng chuẩn bị cho mô hình phân loại người dùng lậu cước và tập luật lậu cước cho bước tiếp theo. Với bộ dữ liệu hồ sơ người dùng được tổng hợp (65 đặc trưng của 120 triệu người dùng), thông qua mô hình trích chọn đặc trưng phù hợp, để lọc ra những đặc trưng quan trọng có ý nghĩa trong việc phát hiện thuê bao lậu cước. Mô hình học thống kê giám sát (supervised learning) được sử dụng, cụ thể mô hình hồi quy đa biến thích ứng (multivariate adaptive regression spline) cho đầu ra với bộ hệ số hồi quy tối ưu của mô hình hồi quy với biến là các đặc trưng và véc-tơ đầu ra là nhãn người dùng gian lận lậu cước. Đánh giá tỉ trọng hệ số hồi quy theo thứ tự giảm dần, loại bỏ các đặc trưng có độ tương quan cao ($|corr| > 0,75$) cho ta bộ đặc trưng tinh gọn (khoảng 20 đặc trưng) về tần suất thực hiện cuộc gọi - tin nhắn, thay đổi vị trí - thiết bị và mối quan hệ viễn thông giúp tăng hiệu quả và độ chính xác cho mô hình phân loại ở bước tiếp theo.

Với bộ dữ liệu bản tin mã hóa người dùng đã được tiền xử lý, tiếp tục thông qua các bước xử lý ngôn ngữ tự nhiên: tách cụm từ (word segmentation), gán nhãn từ loại (part of speech tagging), trích từ khóa (keyword extracting), vec-tơ hóa văn bản (word2vec). Dựa trên tần suất các từ khóa xuất hiện trong bản tin mã hóa người dùng gian lận lậu cước sắp xếp giảm dần, với mỗi nội dung ta trích được các đặc trưng ngữ nghĩa véc-tơ hóa văn bản (word2vec) và từ khóa lậu cước.

Bước 3: xây dựng mô hình phân loại thuê bao nghi ngờ gian lận lậu cước. Trong bước này, mô hình phân loại thuê bao gian lận lậu cước được huấn luyện dựa trên các đặc trưng hồ sơ người dùng được trích chọn ở bước 2. Đầu ra là tập thuê bao nghi ngờ được bổ sung theo từng ngày. Tổng hợp các đặc trưng hồ sơ người dùng đã được trích chọn ở bước 2, ta có đầu vào cho mô hình học máy phân loại sử dụng thuật toán học kết hợp (ensemble learning) rừng quyết định ngẫu nhiên (random forest). Qua mô hình phân loại hai lớp, mỗi thuê bao có xác suất nhãn gian lận lậu cước (từ 0,0 đến 1,0), thuê bao có xác suất hơn ngưỡng sẽ được xếp vào tập nghi ngờ gian lận lậu cước, ngưỡng tối ưu trong phạm vi 0,75.

Bên cạnh đó, mô hình cũng sử dụng các mô hình vệ tinh, phát hiện người giao hàng, cuộc gọi tự động (robot call), người bán hàng chuyên nghiệp (saleman call) để loại bớt các thuê bao hành vi khác ra khỏi tập thuê bao nghi ngờ gian lận lậu cước.

Bước 4: xây dựng bộ luật phát hiện thuê bao gian lận lậu cước trên bản tin mã hóa người dùng. Trong bước này, được thực hiện trên phân hệ thời gian thực, tập thuê bao nghi ngờ gian lận lậu cước trong bước 3, được qua một bộ lọc dựa trên các luật từ khóa và đặc trưng của bản tin mã hóa người dùng khi có hành vi gửi tin nhắn tại thời điểm hiện tại. Đầu ra xác nhận thuê bao có hành vi gian lận lậu cước có độ chính xác cao. Tập thuê bao được xác nhận có gian lận lậu cước được đẩy ra cho bước tiếp theo. Với kết quả của bước 2, thông qua khai khá dữ liệu văn bản đã trích chọn từ khóa của bản tin mã hóa người dùng được gán nhán nội dung lậu cước, sáng chế thu thập được các bộ từ khóa lậu cước với tần suất xuất hiện lớn gồm các từ khóa mã OTP (One Time Password), từ khóa mã hóa, từ khóa tin nhắn ứng dụng,... Các thuê bao trong tập nghi ngờ gian lận lậu cước, có hành vi gửi tin nhắn, sẽ được qua bộ luật từ khóa dựa trên kỹ thuật biểu thức chính quy (regular expression) để xác nhận các từ khóa tĩnh và từ khóa động, cuối cùng các ngưỡng tối ưu về tần suất từ khóa, tần suất gửi tin nhắn, tỉ lệ nhận tin nhắn, cũng được kết hợp để tăng độ tin cậy cho hệ thống phát hiện thuê bao gian lận lậu cước.

Bước 5: đề xuất các phương án chặn lọc thuê bao gian lận lậu cước thời gian thực. Bước cuối cùng, được xử lý trên phân hệ nghiệp vụ, với các thuê bao được phát hiện gian lận lậu cước, sẽ được đưa lên phía xác nhận nghiệp vụ, sáng chế cũng đề xuất các giải pháp ra quyết định cho việc chặn thuê bao thời gian thực khác nhau. Bên cạnh đó, các luật lậu cước tĩnh và động xây dựng ở bước 4, cũng được đề xuất để phía nghiệp vụ thực hiện thống kê, điều chỉnh, tăng tính linh hoạt cho hệ thống. Đối với đặc tính từng nhà mạng, sáng chế đưa ra đồng thời bốn giải pháp chặn lọc các thuê bao phát hiện có hành vi gian lận lậu cước. Các giải pháp chặn lọc theo mức độ tăng dần : chặn làm chậm tin, chặn tin, chặn thuê bao tạm thời, chặn thuê bao một chiều. Đối với, chặn làm chậm, hành vi gửi tin của đối tượng qua chức năng phát hiện thuê bao gian lận lậu cước tích hợp tại trung tâm dịch vụ tin nhắn (SMSC), nếu xác nhận là lậu cước, tin nhắn đến người dùng sẽ bị chậm hơn so với thời gian thực. Giải pháp chặn tin, chặn các tin nhắn gửi đi từ thuê bao được phát hiện gian lận

lậu cước và bản tin mã hóa người dùng thỏa mãn luật từ khóa lậu cước. Giải pháp chặn thuê bao tạm thời hoặc chặn thuê bao một chiều đảm bảo đối tượng có hành vi gian lận lậu cước không thể thực hiện hành vi gửi tin trong một khoảng thời gian đủ lớn.

Trong sáng chế này, bước 1 giúp sáng chế đạt được bộ dữ liệu đặc trưng đa dạng. Bằng cách tập trung vào cả dữ liệu hồ sơ người dùng lẫn bản tin mã hóa người dùng, dựa trên các kỹ thuật khai phá dữ liệu lớn giúp cho hệ thống có nguồn dữ liệu đầy đủ về số lượng đặc trưng lẫn số lượng người dùng, có thể bao quát được hành vi của đối tượng gian lận lậu cước và người dùng thông thường.

Trong sáng chế này, ở bước 2, sáng chế đạt được bộ đặc trưng tối ưu đảm bảo không bị đặc trưng nhiễu và tăng hiệu suất của mô hình phát hiện. Thông qua sử dụng trích chọn đặc trưng theo phương pháp mô hình bao trùm (wrapper method), với mô hình hồi quy đa biến thích ứng đảm bảo bắt được mối quan hệ phức tạp, phi tuyến tính giữa các đặc trưng với đặc điểm người dùng gian lận lậu cước và có tốc độ thực thi nhanh. Tiếp đó qua việc xếp hạng mức độ quan trọng của các đặc trưng và đánh giá tương quan, cho ta một tập con đặc trưng hồ sơ người dùng tối ưu cho mô hình phân loại thuê bao gian lận lậu cước.

Trong sáng chế này, ở bước 3, thực hiện tối ưu mô hình phân loại rừng quyết định ngẫu nhiên (random forest) để xác định thuê bao gian lận lậu cước. Các siêu tham số (hyperparameter) gồm số lượng cây quyết định (number of decision tree), bậc tối đa của cây (maximum depth), số lượng lá tối đa (maximum of leaf nodes) được lựa chọn dựa trên phương pháp kiểm chứng chéo (cross validation) đảm bảo mô hình có độ chính xác cao và tránh hiện tượng quá khớp (overfitting). Tiêu chuẩn Gini cho từng nút cây quyết định cũng được sử dụng để tăng tốc độ tính toán. Bên cách đó, sử dụng thêm các mô hình vệ tinh phát hiện người giao hàng, cuộc gọi tự động, người bán hàng chuyên nghiệp, để loại bớt các thuê bao hành vi khác ra khỏi tập thuê bao nghi ngờ gian lận lậu cước.

Trong sáng chế này, ở bước 5, đưa ra các giải pháp chặn lọc đối tượng gian lận lậu cước khác nhau phù hợp với đặc thù các hệ thống với bốn phương án chặn lọc: chặn làm chậm tin, chặn tin, chặn thuê bao khoảng thời gian và chặn thuê bao một chiều áp dụng phù hợp với mức độ của từng đối tượng. Từ đó đem lại hiệu quả chặn lọc cho các hệ thống.

Theo khía cạnh tiếp theo, sáng chế đề xuất phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông trong đó, tại bước 1, dữ liệu hành vi viễn thông người dùng bao gồm các chỉ số về tần suất nhắn tin, gọi điện, dữ liệu di động được sử dụng; các bộ đặc trưng trong dữ liệu hồ sơ người dùng bao gồm đặc đặc trưng vị trí người dùng, đặc trưng mối quan hệ giữa người dùng, đặc trưng sử dụng thiết bị; kỹ thuật khai phá dữ liệu hồ sơ người dùng sử dụng các kỹ thuật: mô hình thống kê (statistical model), phân tích đồ thị (graph analysis), phân tích dòng thời gian (time-series analysis); việc phân tích dữ liệu, xử lý NLP được lưu trữ, thực thi trong hạ tầng dữ liệu lớn được lưu trữ phân tán, thực thi tính toán song song tại thời gian thực trên các công cụ xử lý dữ liệu lớn Apache Spark, Apache Hadoop.

Theo khía cạnh tiếp theo, sáng chế đề xuất phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông trong đó, tại bước 2: kỹ thuật trích chọn đặc trưng hồ sơ người dùng sử dụng phương pháp trích chọn đặc trưng (feature extraction) dựa theo cách thức mô hình hóa (modeling); với mô hình hồi quy đa biến thích ứng (multivariate adaptive regression spline) với bộ đặc trưng đầu vào lớn, đã đưa ra bộ đặc trưng tinh gọn, có ý nghĩa trong việc dự đoán đối tượng gian lận lậu cước, tăng độ chính xác đồng thời tăng hiệu năng cho hệ thống. Kỹ thuật trích chọn đặc trưng bản tin mã hóa người dùng sử dụng phương pháp xử lý ngôn ngữ tự nhiên theo trình tự rút ra các đặc trưng về mặt ngữ nghĩa. Các kỹ thuật khai phá văn bản (text mining) xây dựng trên toàn bộ từ điển bản tin mã hóa người dùng bao gồm: tách cụm từ (word segmentation), n-gram, véc-tơ hóa văn bản (count vectorization, word2vec). Kỹ thuật trích chọn đặc trưng bản tin mã hóa người dùng về mặt từ khóa lậu cước kết hợp với hai kỹ thuật tách cụm từ (word segmentation), gán nhãn từ loại (part of speech tagging), mô hình tách từ khóa (keyword extraction) cũng được sử dụng để rút ra các từ khóa lậu cước với tần suất xuất hiện lớn. Cũng tại bước này, xử lý ngôn ngữ tự nhiên hiệu năng cao trích chọn đặc trưng bản tin mã hóa người dùng dựa trên cơ sở hạ tầng lưu trữ phân tán Apache Hadoop và các thuật toán tính toán song song Mllib (Machine Learning Library) trên Apache Spark.

Theo khía cạnh tiếp theo, sáng chế đề xuất phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông trong đó, tại bước 3, mô hình học máy phân loại rừng quyết định

ngẫu nhiên (random forest) được tối ưu tham số (parameters tuning), đưa ra mô hình phân loại hai lớp mỗi thuê bao có xác suất nhãn gian lận lậu cước (từ 0,0 đến 1,0) ,thuê bao có xác suất hơn ngưỡng lậu cước sẽ được xếp vào tập nghi ngờ gian lận lậu cước. Cũng tại bước này, mô hình học máy phân loại rừng quyết định ngẫu nhiên (random forest) với ngưỡng phân loại được tự động cập nhập dựa trên độ đo chính xác của mô hình. các nhãn được xác nhận là chính xác người dùng gian lận lậu cước sẽ được cho vào tập kiểm tra (test set). Ngưỡng phân loại tối ưu được cập nhập tự động theo ROC (Receiver operation characteristic) để đạt được độ đo AUC (Area Under Curve) lớn nhất. Bước này cũng sử dụng các mô hình vệ tinh giúp tăng độ chính xác cho tập thuê bao gian lận lậu cước bao gồm các mô hình phát hiện người giao hàng, cuộc gọi tự động, người bán hàng chuyên nghiệp. Các tập thuê bao nghi ngờ gian lận lậu cước của mô hình phân loại sẽ được lọc qua để loại bỏ các thuê bao nằm trong các thuê bao của mô hình vệ tinh. .

Theo khía cạnh tiếp theo, sáng chế đề xuất phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông trong đó, tại bước 4, bộ từ khóa được trích chọn từ bản tin mã hóa người dùng là các từ khóa lậu cước qua bước 2, với các từ khóa được trích chọn trong các tin được gán nhãn tin từ đối tượng lậu cước, tạo nên một bộ từ khóa lậu cước. Tại bước này, thông qua bộ lọc từ khóa lậu cước, đảm bảo độ chính xác cao cho hệ thống phát hiện đối tượng gian lận lậu cước. Với đầu ra từ bước 3, thuê bao được phân loại trên hồ sơ người dùng là thuê bao gian lận lậu cước, sẽ được tự động thu thập bản tin mã hóa người dùng tại thời gian thực qua hệ thống trung tâm dịch vụ tin nhắn (SMSC), nếu nội dung thỏa mãn các luật từ khóa lậu cước sẽ được đưa đến bước 5, xử lý chặn lọc. Cũng tại bước này, bộ luật từ khóa động được xây dựng để tăng tính linh hoạt; với một tập con bộ luật từ khóa sẽ được cấu hình động, tăng tính sẵn sàng cho hệ thống giải quyết hiện trạng đối tượng gian lận lậu cước liên tục thay đổi hành vi. .

Theo khía cạnh tiếp theo, sáng chế đề xuất phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông trong đó, tại bước 5, các thuê bao đã được xác nhận và bị chặn lọc gian lận lậu cước được lưu trữ đặc trưng tiến hành phân tích. Trong đó, hồ sơ người dùng và các bản tin mã hóa người dùng của đối tượng gian lận lậu cước được đưa vào phân hệ xử lý dữ liệu lớn, tiến hành phân tích dữ liệu, khai phá thêm các đặc trưng để cải tiến

mô hình. Phương án chặn lọc thuê bao gian lận lậu cước hướng chặn thuê bao kết hợp với chặn hoàn toàn một chiều, với phương pháp chặn thuê bao tạm thời trong một khoảng thời gian từ lúc có hành vi gửi tin nhắn gian lận lậu cước. Cũng tại bước này phương án chặn lọc thuê bao gian lận lậu cước hướng chặn tin làm chậm được đề xuất đối với các nhà mạng trong và ngoài nước không cho phép chặn một chiều. đối tượng thông qua chặn làm chậm, khi có hành vi gửi tin nhắn, xử lý trên hệ thống trung tâm dịch vụ tin nhắn (SMSC) qua một khoảng thời gian đủ lớn (10-30 phút) thì người dùng mới nhận được tin nhắn. Phương án này đảm bảo vô hiệu hóa mục đích nhắn tin của đối tượng gian lận lậu cước.

Ví dụ thực hiện sáng chế

Sáng chế đã đi vào thực nghiệm cho kết quả chặn thuê bao lậu cước lũy kế (số lượng thuê bao) trong các tháng từ đầu năm 2021 như sau:

- Tháng 1: 19465;
- Tháng 2: 8623;
- Tháng 3: 9028;
- Tháng 4: 23952;
- Tháng 5: 23098;
- Tháng 6: 56068;
- Tháng 7: 27933;
- Tháng 8: 133792;
- Tháng 9: 73102.

Hiệu quả đạt được của sáng chế

Phương pháp chặn lọc thuê bao gian lận lậu cước đem đến những hiệu quả như sau: Giảm tình trạng đối tượng gửi tin nhắn gian lận lậu cước. Đưa ra các biện pháp chặn lọc thuê bao có hành vi gian lận lậu cước, từ đó dẫn tới việc sử dụng các đầu số thương hiệu (brandname) để nhắn các tin nhắn ứng dụng, giúp tăng trải nghiệm cho khách hàng.

Mặc dù các mô tả nêu trên chứa nhiều chi tiết cụ thể, chúng không được coi là giới hạn về phương án thực thi sáng chế, mà chỉ nhằm mục đích minh họa một số phương án thực thi được ưu tiên.

Yêu cầu bảo hộ

1. Phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông bao gồm:

bước 1: thu thập, phân tích dữ liệu hồ sơ người dùng, bản tin mã hóa người dùng gian lận lậu cước; trong đó, phân hệ xử lý dữ liệu lớn tự động thu thập dữ liệu lịch sử hành vi sử dụng viễn thông của thuê bao trong một khoảng thời gian dài (sáu tháng đến một năm), dữ liệu được mã hóa đảm bảo an toàn thông tin, làm sạch dữ liệu, được tổng hợp các đặc trưng, xây dựng bộ dữ liệu hồ sơ người dùng qua các công nghệ xử lý dữ liệu lớn; bên cạnh đó, phân hệ cũng thực hiện xử lý bản tin mã hóa người dùng lậu cước và thông thường ở dạng văn bản, phân đoạn, làm sạch, tiền xử lý dữ liệu; các bộ dữ liệu hành vi viễn thông, đặc trưng hồ sơ người dùng và bản tin mã hóa người dùng được chuẩn bị sẵn sàng cho bước tiếp theo; trong đó, ở bước này, giúp sáng chế đạt được bộ dữ liệu đặc trưng đa dạng bằng cách tập trung vào cả dữ liệu hồ sơ người dùng lẫn bản tin mã hóa người dùng, dựa trên các kỹ thuật khai phá dữ liệu lớn giúp cho hệ thống có nguồn dữ liệu đầy đủ về số lượng đặc trưng lẫn số lượng người dùng, có thể bao quát được hành vi của đối tượng gian lận lậu cước và người dùng thông thường;

bước 2: trích chọn các đặc trưng của người dùng; bước này thực thi đồng thời trích chọn các đặc trưng hồ sơ người dùng và đặc trưng từ khóa lậu cước từ bản tin mã hóa người dùng trên phân hệ Xử lý dữ liệu lớn; với bộ đặc trưng hồ sơ người dùng được tổng hợp từ dữ liệu bản ghi chi tiết CDR (call detail record) cuộc gọi và tin nhắn, thông qua mô hình trích chọn đặc trưng đưa ra bộ đặc trưng tối ưu có ý nghĩa trong việc xác định người dùng có hành vi gian lận lậu cước; bên cạnh đó, các kỹ thuật xử lý gônn ngữ tự nhiên (nature language processing) cũng được sử dụng để khai thác các đặc trưng về mặt ngữ nghĩa gồm từ khóa lậu cước, tần suất từ khóa; các đặc trưng đầu ra này, sẵn sàng chuẩn bị cho mô hình phân loại người dùng lậu cước và tập luật lậu cước cho bước tiếp theo; trong đó, ở bước này, sáng chế đạt được bộ đặc trưng tối ưu đảm bảo không bị đặc trưng nhiễu và tăng hiệu suất của mô hình phát hiện; thông qua sử dụng trích chọn đặc trưng theo phương pháp mô hình bao trùm (wrapper method), với mô hình hồi quy đa biến thích ứng đảm bảo bắt được mối quan hệ phức tạp, phi tuyến tính giữa các đặc trưng với đặc điểm người dùng gian lận

lậu cước và có tốc độ thực thi nhanh; tiếp đó qua việc xếp hạng mức độ quan trọng của các đặc trưng và đánh giá tương quan, cho ta một tập con đặc trưng hồ sơ người dùng tối ưu cho mô hình phân loại thuê bao gian lận lậu cước;

bước 3: xây dựng mô hình phân loại thuê bao nghi ngờ gian lận lậu cước; trong bước này, mô hình phân loại thuê bao gian lận lậu cước được huấn luyện dựa trên các đặc trưng hồ sơ người dùng được trích chọn ở bước 2; đầu ra là tập thuê bao nghi ngờ được bổ sung theo từng ngày; trong đó, ở bước này, thực hiện tối ưu mô hình phân loại rừng quyết định ngẫu nhiên (random forest) để xác định thuê bao gian lận lậu cước; các siêu tham số (hyperparameter) gồm số lượng cây quyết định (number of decision tree), bậc tối đa của cây (maximum depth), số lượng lá tối đa (maximum of leaf nodes) được lựa chọn dựa trên phương pháp kiểm chứng chéo (cross validation) đảm bảo mô hình có độ chính xác cao và tránh hiện tượng quá khớp (overfitting); tiêu chuẩn gini cho từng nút cây quyết định cũng được sử dụng để tăng tốc độ tính toán; bên cách đó, sử dụng thêm các mô hình vệ tinh phát hiện người giao hàng, cuộc gọi tự động, người bán hàng chuyên nghiệp, để loại bớt các thuê bao hành vi khác ra khỏi tập thuê bao nghi ngờ gian lận lậu cước;

bước 4: xây dựng bộ luật phát hiện thuê bao gian lận lậu cước trên bản tin mã hóa người dùng; trong bước này, được thực hiện trên phân hệ xử lý thời gian thực, tập thuê bao nghi ngờ gian lận lậu cước trong bước 3, được qua một bộ lọc dựa trên các luật từ khóa và đặc trưng của bản tin mã hóa người dùng khi có hành vi gửi tin nhắn tại thời điểm hiện tại; đầu ra xác nhận thuê bao có hành vi gian lận lậu cước có độ chính xác cao; tập thuê bao được xác nhận có gian lận lậu cước được đẩy ra cho bước tiếp theo;

bước 5: đề xuất các phương án chặn lọc thuê bao gian lận lậu cước thời gian thực; đây là bước cuối cùng, được xử lý trên phân hệ nghiệp vụ, với các thuê bao được phát hiện gian lận lậu cước, sẽ được đưa lên phía xác nhận nghiệp vụ, sáng chế cũng đề xuất các giải pháp ra quyết định cho việc chặn thuê bao thời gian thực khác nhau; bên cạnh đó, các luật lậu cước tĩnh và động xây dựng ở bước 4, cũng được đề xuất để phía nghiệp vụ thực hiện thống kê, điều chỉnh, tăng tính linh hoạt cho hệ thống; trong đó, ở bước này, đưa ra các giải pháp chặn lọc đối tượng gian lận lậu cước khác nhau phù hợp với đặc thù các hệ thống; bốn phương án chặn lọc được đưa ra gồm: chặn làm chậm tin, chặn tin, chặn thuê bao khoảng

thời gian và chặn thuê bao một chiều áp dụng phù hợp với mức độ của từng đối tượng, từ đó đem lại hiệu quả chặn lọc cho các hệ thống.

2. Phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông theo điểm 1, trong đó, tại bước 1: dữ liệu hành vi viễn thông người dùng bao gồm các chỉ số về tần suất nhắn tin, gọi điện, dữ liệu di động sử dụng.

3. Phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông theo điểm 1 hoặc điểm 2, trong đó, tại bước 1: các bộ đặc trưng trong dữ liệu hồ sơ người dùng bao gồm đặc đặc trưng vị trí người dùng, đặc trưng mối quan hệ giữa người dùng, đặc trưng sử dụng thiết bị.

4. Phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông theo điểm bất kỳ trong số các điểm từ điểm 1 đến điểm 3, trong đó, tại bước 1: kỹ thuật khai phá dữ liệu hồ sơ người dùng sử dụng các kỹ thuật mô hình thống kê (statistical model), phân tích đồ thị (graph analysis), phân tích dòng thời gian (time-series analysis).

5. Phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông theo điểm bất kỳ trong số các điểm từ điểm 1 đến điểm 4, trong đó, tại bước 1: việc phân tích dữ liệu, xử lý ngôn ngữ tự nhiên được lưu trữ, thực thi trong hạ tầng dữ liệu lớn; được lưu trữ phân tán, thực thi tính toán song song tại thời gian thực trên các công cụ xử lý dữ liệu lớn apache spark, apache hadoop.

6. Phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông theo điểm bất kỳ trong số các điểm từ điểm 1 đến điểm 5, trong đó, tại bước 2: kỹ thuật trích chọn đặc trưng hồ sơ người dùng sử dụng phương pháp trích chọn đặc trưng (feature extraction) dựa theo cách thức mô hình hóa (modeling).

7. Phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông theo điểm bất kỳ trong số các điểm từ điểm 1 đến điểm 6, trong đó, tại bước 2: kỹ thuật trích chọn đặc trưng bản tin mã hóa người dùng sử dụng phương pháp xử lý ngôn ngữ tự nhiên theo trình tự rút ra các đặc trưng về mặt ngữ nghĩa; kỹ thuật trích chọn đặc trưng bản tin mã hóa người dùng

về mặt từ khóa lậu cước kết hợp với hai kỹ thuật tách cụm từ (word segmentation), gán nhãn từ loại (part of speech tagging), mô hình tách từ khóa (keyword extraction) cũng được sử dụng để rút ra các từ khóa lậu cước với tần suất xuất hiện lớn.

8. Phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông theo điểm bất kỳ trong số các điểm từ điểm 1 đến điểm 7, trong đó, tại bước 2: xử lý ngôn ngữ tự nhiên hiệu năng cao trích chọn đặc trưng bản tin mã hóa người dùng dựa trên cơ sở hạ tầng lưu trữ phân tán apache hadoop và các thuật toán tính toán song song MLLIB (machine learning library) trên apache spark.

9. Phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông theo điểm bất kỳ trong số các điểm từ điểm 1 đến điểm 8, trong đó, tại bước 3: mô hình học máy phân loại rừng quyết định ngẫu nhiên (random forest) được tối ưu tham số (parameters tuning), đưa ra mô hình phân loại hai lớp mỗi thuê bao có xác suất nhãn gian lận lậu cước (từ 0,0 đến 1,0) ,thuê bao có xác suất hơn ngưỡng lậu cước sẽ được xếp vào tập nghi ngờ gian lận lậu cước; mô hình học máy phân loại rừng quyết định ngẫu nhiên (random forest) với ngưỡng phân loại được tự động cập nhập dựa trên độ đo chính xác của mô hình.

10. Phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông theo điểm bất kỳ trong số các điểm từ điểm 1 đến điểm 9, trong đó, tại bước 3: sử dụng các mô hình vệ tinh giúp tăng độ chính xác cho tập thuê bao gian lận lậu cước bao gồm các mô hình phát hiện người giao hàng, cuộc gọi tự động, người bán hàng chuyên nghiệp; các tập thuê bao nghi ngờ gian lận lậu cước của mô hình phân loại sẽ được lọc qua để loại bỏ các thuê bao nằm trong các thuê bao của mô hình vệ tinh.

11. Phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông theo điểm bất kỳ trong số các điểm từ điểm 1 đến điểm 10, trong đó, tại bước 4: bộ từ khóa được trích chọn từ bản tin mã hóa người dùng là các từ khóa lậu cước qua bước 2, với các từ khóa được trích chọn trong các tin được gán nhãn tin từ đối tượng lậu cước, tạo nên một bộ từ khóa lậu cước.

12. Phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông theo điểm bất kỳ trong số các điểm từ điểm 1 đến điểm 11, trong đó, tại bước 4: thông qua bộ lọc từ khóa lậu cước, đảm bảo độ chính xác cao cho hệ thống phát hiện đối tượng gian lận lậu cước; với đầu ra từ bước 3, thuê bao được phân loại trên hồ sơ người dùng là thuê bao gian lận lậu cước, sẽ được tự động thu thập bản tin mã hóa người dùng tại thời gian thực qua hệ thống trung tâm dịch vụ tin nhắn (SMSC), nếu bản tin mã hóa người dùng thỏa mãn các luật từ khóa lậu cước sẽ được đưa đến bước 5, xử lý chặn lọc.

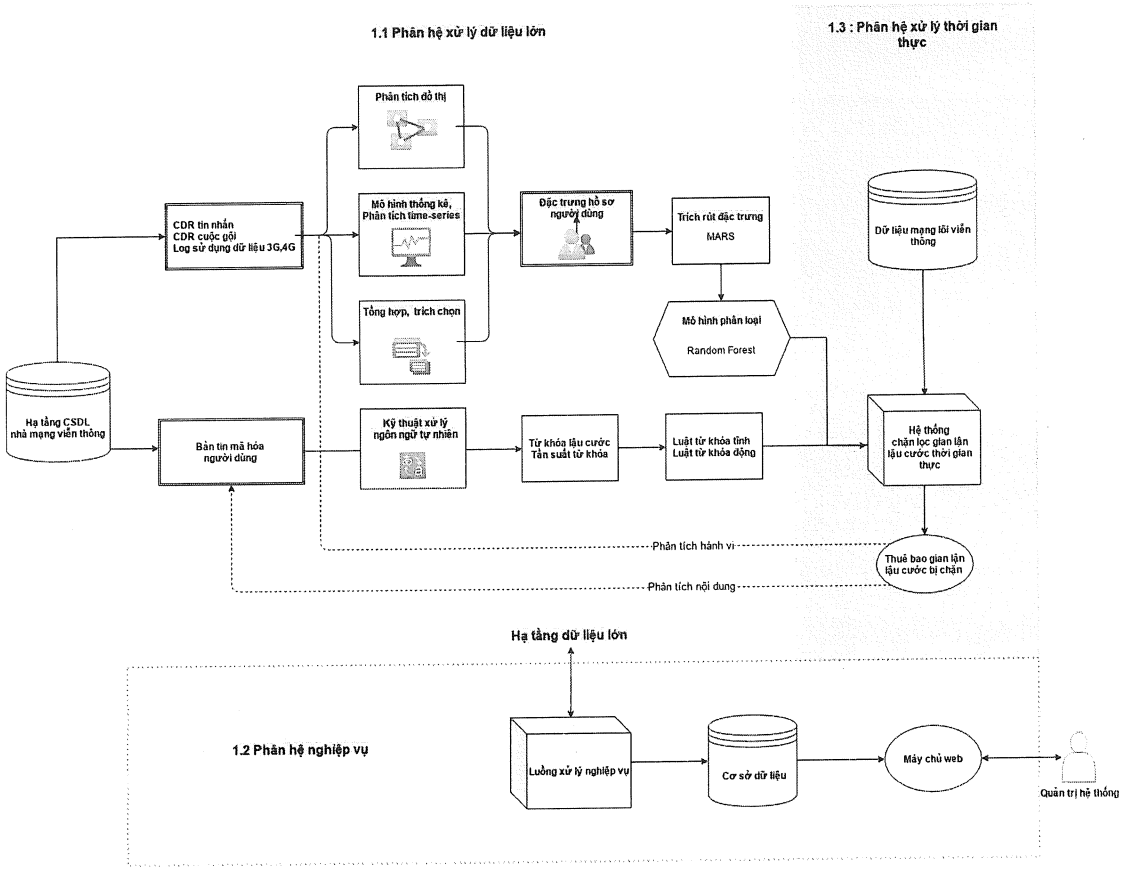
13. Phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông theo điểm bất kỳ trong số các điểm từ điểm 1 đến điểm 12, trong đó, tại bước 4: bộ luật từ khóa động được xây dựng để tăng tính linh hoạt; với một tập con bộ luật từ khóa sẽ được cấu hình động, tăng tính sẵn sàng cho hệ thống giải quyết hiện trạng đối tượng gian lận lậu cước liên tục thay đổi hành vi.

14. Phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông theo điểm bất kỳ trong số các điểm từ điểm 1 đến điểm 13, trong đó, tại bước 5: các thuê bao đã được xác nhận và bị chặn lọc gian lận lậu cước được lưu trữ đặc trưng tiến hành phân tích; trong đó, hồ sơ người dùng và các bản tin mã hóa người dùng của đối tượng gian lận lậu cước được đưa vào phân hệ xử lý dữ liệu lớn, tiến hành phân tích dữ liệu, khai phá thêm các đặc trưng để cải tiến mô hình.

15. Phương pháp phát hiện thuê bao gian lận lậu cước trong viễn thông theo điểm bất kỳ trong số các điểm từ điểm 1 đến điểm 14, trong đó, tại bước 5: phương án chặn lọc thuê bao gian lận lậu cước hướng chặn thuê bao kết hợp với chặn hoàn toàn một chiều, với phương pháp chặn thuê bao tạm thời trong một khoảng thời gian từ lúc có hành vi gửi tin nhắn gian lận lậu cước; phương án chặn lọc thuê bao gian lận lậu cước hướng chặn tin làm chậm được đề xuất đối với các nhà mạng trong và ngoài nước không cho phép chặn một chiều.


Thu thập, phân tích dữ liệu
Dữ liệu hồ sơ người dùng
Dữ liệu bản tin mã hóa
Trích chọn đặc trưng người dùng
Bộ đặc trưng tối ưu cho phát hiện lậu cước
Xây dựng mô hình phân loại thuê bao gian lận lậu cước
Tập thuê bao nghi ngờ gian lận lậu cước
Xây dựng bộ luật phát hiện thuê bao gian lận lậu cước dựa trên nội dung sms
Tập thuê bao phát hiện có hành vi gian lận lậu cước
Đề xuất phương án chặn lọc thuê bao gian lận lậu cước thời gian thực

Hình 1


1.1 Phân hệ xử lý dữ liệu lớn
1.3 : Phân hệ xử lý thời gian thực
Phân tích đồ thị
CDR tin nhắn
CDR cuộc gọi
Log sử dụng dữ liệu 3G,4G
Mô hình thống kê, Phân tích time-series
Đặc trưng hồ sơ người dùng
Trích rút đặc trưng MARS
Dữ liệu mạng lõi viễn thông
Tổng hợp, trích chọn
Mô hình phân loại Random Forest
Hạ tầng CSDL nhà mạng viễn thông
Bản tin mã hóa người dùng
Kỹ thuật xử lý ngôn ngữ tự nhiên
Từ khóa lậu cước Tần suất từ khóa
Luật từ khóa tĩnh Luật từ khóa động
Hệ thống chặn lọc gian lận lậu cước thời gian thực
Phân tích hành vi
Thuê bao gian lận lậu cước bị chặn
Phân tích nội dung
Hạ tầng dữ liệu lớn
1.2 Phân hệ nghiệp vụ
Luồng xử lý nghiệp vụ
Cơ sở dữ liệu
Máy chủ web
Quản trị hệ thống

Hình 2