(19) Евразийское патентное ведомство

(21) 201900470 (13) A2

# (12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОЙ ЗАЯВКЕ

(43) Дата публикации заявки
**2020.06.30**

(22) Дата подачи заявки
**2019.10.07**

(51) Int. Cl. ***G06F 17/00*** (2019.01)

## (54) СИСТЕМА КЛАССИФИКАЦИИ ТРАФИКА

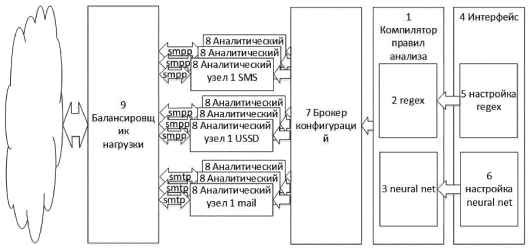
(31) **2018135235**
(32) **2018.10.05**
(33) **RU**

(71) Заявитель:
**ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "АЛГОРИТМ" (RU)**

(72) Изобретатель:
**Горькова Мария Давидовна (RU)**

(74) Представитель:
**Юрченко А.В. (RU)**

(57) Предложены способ и реализующая его система классификации трафика, предоставляющие гибкую и универсальную настройку классификации обычными администраторами сетевых узлов за счет применения авторского неформального языка регулярных выражений irregex. Выполнение системы классификации с применением неформального языка регулярных выражений irregex позволяет снять ограничения на обслуживание системы узконаправленными высококлассными специалистами и предоставить возможность ее эксплуатации обычным телекоммуникационным специалистам. Включение в процесс классификации пользовательских вероятностных коэффициентов и выполнение модуля neural net системы классификации с возможностью корректировки результатов работы нейросети (собственной архитектуры или сторонних инструментов) позволяет администратору системы быстро реагировать на изменение схемы и семантики сообщения и оперативно "подправлять" результаты нейронной классификации. Причем способ классификации как в последовательно-параллельном режиме, так и в параллельном обеспечивает существенный прирост точности классификации за счет синергии алгоритмов regex и neural net. Что особенно важно, способ за счет архитектуры системы классификации трафика обеспечивает опосредованное взаимовлияние пользовательских настроек, притом что администратор системы может настраивать их независимо.




Система классификации траффика TCS (Traffic Classification System)

201900470 A2

201900470 A2

# Система классификации траффика

Предлагаемое изобретение относится к области обработки данных, в частности к системам синтаксической обработки потока текстовых данных, используемым в телекоммуникациях, и предназначено для обработки сообщений, передающихся как между абонентами, так и между абонентами и информационными системами и может использоваться в a2p, p2a, p2p, a2a сервисах.

В задачах обработки данных почти всегда возникает вопрос о последовательном уточнении результатов по мере поступления новых данных "на лету". Общеизвестен рост трафика a2p, p2a, p2p, a2a сервисов. Банки, социальные сети, мобильные операторы генерируют большой объём сообщений для информирования своих клиентов в том числе автоматически. Технологии SMS и USSD стандартизованные в рамках SMPP протокола по-прежнему востребованы в рамках указанных сервисов. В этой связи возникает задача классификации траффика. Задача классификации траффика по типу протокола или приложения - тип траффика SMPP для SMS и USSD, SMTP для mail, и др., на современном уровне техники решена за счет известной технологии анализа сетевых пакетов. Задача распознавания вида траффика по семантике сообщений (рекламный, транзакционный, сервисный, международный, мошеннический траффик и т.д.) для принятия бизнес решений представляет собой сложную техническую задачу. Из уровня техники можно выделить два основных способа классификации текста – семантический анализ текста на основе регулярных выражений и классификация текста эвристическими методами, основу которых составляют методы машинного обучения. У каждого способа есть свои преимущества и недостатки. Специфика a2p, p2p, p2a, a2a траффика в общем случае заключается в том, что большой массив текстовых данных представляет собой множество коротких сообщений. В части машинного распознавания это представляет сложности т.к. в небольшом тексте значительно меньше полезных признаков, и велико влияние порядка слов в сравнении с классификацией больших блоков текста. Это усугубляется тем, что этот траффик сильно подвержен изменениям как в синтаксической, так и в семантической части, причем для устройства распознавания траффика эти изменения происходят беспрерывно (онлайн).

Из уровня техники известны технологии семантического анализа, в основном представленные программно-аппаратными комплексами, выполненными по клиент-

серверной архитектуре. IBM LanguageWare платформа, представляющая собой набор Java, C++ библиотек, обеспечивает разработчиков набором инструментов для обработки естественного языка (Natural Language Processing, NLP). IBM AlchemyAPI – облачный сервис предоставляющий технологию NLP по SaaS модели через свои интерфейсы API, есть набор средств разработки (SDK) поддерживаемый компанией IBM. Оба этих инструмента предоставляют встраиваемых клиентов API, непосредственная обработка текста выполняется на удаленных серверах компании-разработчика. PalitrumLab EUREKA ENGINE – отечественный представитель платформ лингвистического анализа, предоставляется как по внешнему API, так и в качестве встраиваемой системы для бизнес-приложений. У указанных систем есть свои преимущества и недостатки, так LanguageWare и AlchemyAPI более приспособлены для обработки больших объемов текста и показывают меньшую эффективность при обработке потока разнородных сообщений. PalitrumLab, хорошо справляясь с вероятностной обработкой потока текстовых данных, не поддерживает их четкую классификацию. Существуют другие технологии, в целом однотипные по облику и архитектуре.

Существенным недостатком, объединяющим указанные выше и другие известные решения, с точки зрения авторов, является отсутствие возможности на уже запущенной системе применять пользовательские настройки. Настройки анализа прописываются разработчиками в коде при интеграции клиентов или самостоятельных систем в бизнес-приложения и применяются к данным после компиляции. Т.е. конфигурирование известных систем доступно только специалистам, разрабатывающим клиентское ПО для интеграции в бизнес-приложения. Пользователям, обслуживающим бизнес-приложения, настройка правил анализа недоступна. Рост количества гетерогенных узлов в сети общеизвестен, возрастает обмен данными, требующий разнородного анализа. Уровень техники не позволяет формализовать, т.е. положить на логику алгоритма, все текущие и будущие прикладные задачи семантического анализа. В то же время, инженерам и пользователям бизнес-приложений требуется универсальный инструмент, позволяющий без трудоемкого изменения алгоритма оперативно изменять и применять правила обработки текстовых и других данных при существенном изменении схемы текстовых данных (синтаксиса и семантики сообщения). По этой же причине роста гетерогенных узлов в сети важное значение имеет производительность оборудования и алгоритмов управления им.

Авторы предлагают универсальное решение классификации траффика по категориям, т.е. семантическому виду траффика - рекламный, транзакционный,

сервисный, международный, мошеннический и др. Решение совмещает точный разбор текста по семантическим атрибутам и вероятностную классификацию текста на основе машинного обучения. Алгоритм способа классификации и архитектура системы, реализующая его предоставляет администраторам системы в режиме реального времени настраивать категории классификации сообщения, правила его анализа и корректировать результаты сепарабельности (отделимости результатов классификации) нейронной сети. Причем предлагаемые настройки предоставляются широкому кругу пользователей и не требуют глубокой подготовки в вопросах семантического анализа и/или машинного обучения. Дополнительно, способ интегрируется в разнородные программно-аппаратные платформы (узлы сети), и масштабирует свою производительность в соответствии с потоковой нагрузкой онлайн.

Известен патент №US8494987 от 23.07.13г. представляющий систему и способы генерации гипотез, в котором локальная или распределенная система взаимодействует с базой данных и содержит модуль генерации гипотез, модуль хранения концепций, на основании которых генерируются гипотезы, а также модули пользовательского ввода и вывода результатов. А способ генерации гипотез заключается в предварительном вычислении гипотез на основе предустановленных концепций (теме знаний) и данных БД, причем после приема пользовательского ввода система генерирует по крайней мере одну гипотезу, основанную на совместно встречающихся концепциях (теме знаний) имеющих отношение к сущностям, выделенным из пользовательского ввода (предложения, утверждения).

Способ требует пользовательского ввода базы знаний или указание на базу данных по принадлежности к анализируемым сущностям или инструкции по обработке текста, что исключает его использование в режиме постоянно поступающих больших объемов онлайн данных.

Недостатком программно-аппаратной архитектуры является «вертикальная», не масштабируемая модульность программной архитектуры, в которой модули жестко сконфигурированы для последовательного выполнения функций анализа-загрузка данных из БД, предварительного вычисления, приема пользовательских данных, вычисления, вывода.

Патент № PCT/US2007/063983 от 14.03.2007г. описывает программный продукт, систему и способ генерации гипотез из базы данных, по которому строятся и заводятся в структуры различные семантические связи между концепциями и понятиями и/или

концепциями и концепциями по определенным правилам отношений и на основании разбора множества документов в предустановленном репозитории. Фразы назначаются по одному из множества понятий, или идентификатору отношения, связывающего одно понятие со вторым, или присвоение одного понятия семантической категории или расположение концепций в иерархической взаимосвязи.

Недостатками такого решения также является его неприменимость при обработке постоянно поступающих потоковых данных и невозможность масштабирования нагрузки.

Заявка PCT №WO2013/135474 от 22.02.2013г. притязает на способ семантического анализа текста, заключающегося в построении матрицы векторов вхождений слов в обучающую выборку документов, например статей Википедии, обучении 1 нейронной сети (самоорганизующейся карты Кохонена) этой выборкой документов и векторов, сопоставление документов, содержащих целевое слово и множество координат карты (область) вхождений этих документов, определения семантического контекста слова и сохранение соответствия в словарь,карту шаблонов. Обучение второй нейронной сети на документах того же языка, состоящее из построения последовательности предложений из документов обучающей выборки, подсчет коэффициента сложности документа (частота слова в документе и т.д.) и сортировка слов в последовательности. И обучение сети документами выборки с возрастанием коэффициента семантической сложности, сопоставление карты и слов второй выборки, обучение второй (как правило иерархической) нейронной сети картами шаблонов.

Решение основано только на эвристических методах анализа, не предполагает четкой классификации текста, не приспособлено под оперативную настройку правил анализа и не применимо к анализу потоковых данных.

Патент RU2615632 от 20.03.2015г. предлагающий способ распознавания коммуникационных сообщений в части определения имени отправителя и предполагающий наличие специальных символов в сообщении выделяющих имя отправителя или запрос на сервер с предустановленной базой данных имен абонентов и определение имени отправителя по результатам сравнения содержимого заданного определителя и имени пользователя в базе данных. При этом семантический анализ согласно изобретению может быть реализован аппаратно за счет доработки мобильного терминала модулями обработки текста и связи с сервером БД.

В данном решении очевидными недостатками является предопределение набора семантических признаков отправителя, вынужденное поддержание базы данных с

именами пользователей, решение предусматривает аппаратную доработку конечного устройства пользователя устройством распознавания. Отсутствие алгоритмов нечеткой логики (машинного обучения), невозможность настройки правил также отмечается как недостаток.

Наиболее близким по технической сути и принятый за прототип является патент RU 2429533 от 29.06.2009г. который описывает механизм динамического синтаксического анализа/компоновки на основе схем для синтаксического анализа мульти форматных сообщений выполняющий прием сообщения в различных форматах, преобразование его в общий формат, определение грамматической структуры данных формата сообщения и на ее основании указания на подходящий обработчик. Причем обработчики выполнены модулями с отдельной компиляцией и могут добавляться в архитектуру по мере появления новых форматов сообщений. За счет подгрузки отдельных модулей обработки достигается выигрыш в производительности системы анализа в целом. Дополнительный выигрыш в производительности достигается за счет индексации внутреннего формата сообщений. Маршрутизация сообщений к обработчикам, правила обработки и новые спецификации загружаются динамически в БД при регистрации новой услуги, за счет этого достигается универсальность решения.

Такое решение применимо для обработки данных, поступающих в потоке онлайн, однако очевидным недостатком является процесс лексического и синтаксического разбора сообщений только по правилам анализа описания синтаксиса формальных языков (парсинг). Представляется, что этого недостаточно для сематического анализа, осмысления текстовых сообщений естественного языка для принятия бизнес решений. Хотя американскими авторами указано, что обработчики загружаются динамически в БД, и новые обработчики не влияют на работу системы в целом, это не избавляет авторов решения от необходимости разработки, компиляции и тестирования каждого нового обработчика для загрузки.

Авторы отмечают интересное решение индексации формата сообщений, вытекающего из архитектуры такой системы, тем не менее отмечают и недостатки прототипа:

1. Невозможность ручной «горячей» пользовательской настройки правил анализа онлайн без необходимости заблаговременно разрабатывать, тестировать и подключать новый модуль обработчик семантического анализа для новой грамматики.
2. Отсутствие вероятностной оценки сообщения.

3. Плохая аппаратная масштабируемость - отсутствие решения для автоматической настройки производительности системы при повышении нагрузки - существенного увеличения входящих сообщений.

Технической задачей предлагаемой системы классификации траффика по способу классификации a2p, p2a, p2p, a2a траффика по предопределенным семантическим категориям траффика является:

1. повышение гибкости классификации сообщений за счет ручного пользовательского определения интуитивно понятных неформальных шаблонов-строк, применяемых к анализу текста и определения вероятностных коэффициентов отнесения сообщения к определенной категории;
2. повышение надежности классификации сообщений за счет объединения алгоритмов четкой логики на основе применения регулярных выражений, эвристических алгоритмов нейросетей и дополнительной гибкой настройки этих алгоритмов в соответствии с пользовательскими определениями шаблонов-строк и вероятностных коэффициентов;
3. повышение производительности анализа сообщений за счет динамического генерирования (воспроизведения) и распределения конфигурации правил анализа по множеству узлов, анализирующих сообщение.

Внедрение в процесс аналитической обработки текста указанных выше пользовательских настроек выводит ее на новый универсальный уровень, на которым процесс обработки не только подвластен обычным администраторам (пользователям) системы, не являющимся специалистами в области семантического анализа, языков и диалектов регулярных выражений, но и неограничен существующими и будущими прикладными задачами анализа, обусловленными изменениями схемы сообщения. Более того, совместное их применение к анализируемым сообщениям обеспечивает существенное повышение достоверности анализа. Взаимовлияние настроек пользователя на результат классификации предопределено последовательностью выполнения способа, предполагающего сначала применение неформальных шаблонов-строк к тексту сообщения алгоритмами четкой логики, и последующим применением вероятностных коэффициентов к результатам эвристических алгоритмов нейросети. При этом способ предусматривает не зависимость определения пользовательских настроек.

Технический результат способа классификации трафика достигается за счет системы классификации трафика, в соответствии с которым принимают сообщение естественного языка, определяют схему сообщения, делят сообщение на значащие синтаксические

(семантические) поля, предоставляют множество узлов для анализа значащих полей сообщения, к которым применяют правила анализа текста в зависимости от прикладного протокола сообщения, причем для классификации сообщения по определенным категориям определяют неформальный шаблон-строку неформального языка поиска и обработки подстроки в строке и вероятностный коэффициент отнесения сообщения к определенной категории, преобразуют неформальный шаблон-строку в регулярное выражение стандартного языка поиска и обработки подстроки в строке, преобразуют регулярное выражение стандартного языка в программный модуль конфигурации правил анализа текста содержащий данные и методы обработки строк, динамически создают множество программных модулей конфигурации правил анализа и распределяют их по множеству узлов для анализа значащих полей сообщения, применяют конфигурацию правил анализа к значащим полям сообщения, подают сообщение на вход нейронной сети, причем результат классификации сообщения нейронной сетью корректируют в соответствии с определенным вероятностным коэффициентом отнесения сообщения к определенной категории. Определение неформального шаблона-строки неформального языка поиска и обработки подстроки в строке и вероятностного коэффициента отнесения сообщения к определенной категории производят в процессе обработки сообщения. Множество программных модулей конфигурации правил анализа генерируют и распределяют по множеству узлов для анализа значащих полей сообщения пропорционально сетевой нагрузке, обусловленной количеством анализируемых сообщений. А результат классификации нейронной сети корректируют в соответствии с определенным вероятностным коэффициентом отнесения сообщения к определенной категории, трансформацией пороговой величины классификации нейросети или установлением границы достоверности выходных элементов нейросети.

Технической задачей устройства системы классификации a2p, p2a, p2p, a2a трафика для осуществления способа классификации траффика, является:

1. улучшение эксплуатационных свойств, за счет включения в архитектуру системы классификации трафика модуля 4 интерфейса пользователя, выполненного с возможностью ручного пользовательского определения семантических категорий классификации сообщения, неформальных шаблонов-строк, применяемых к анализу сообщения для определения его категории и вероятностных коэффициентов отнесения сообщения к определенной категории;
2. повышение надежности классификации, за счет включения в архитектуру системы модуля 2 regex, в памяти которого содержится и процессором которого выполняется алгоритм

неформального языка регулярных выражений irregex (irregular expressions) и алгоритм одного или нескольких формальных языков регулярных выражений regex (regular expressions). Причем алгоритм irregex позволяет получать неформальные пользовательские шаблоны-строки и преобразовывать их в регулярные выражения стандартных языков regex, а модуль нейросети 3 neural net выполнен с возможностью алгоритмического преобразования пользовательских вероятностных коэффициентов в трансформацию пороговой величины нейросети и/или установления границы достоверности значений ее выходных элементов;

3. улучшение эксплуатационных свойств, за счет возможности использования множества инструментов семантического анализа, которое достигается «инкапсуляцией» программной архитектуры-пользователю системы предоставляется только модуль 4 интерфейса пользователя, а полиморфная работа (интеграция с широкой номенклатурой стандартных библиотек regex и программ-нейроимитаторов) достигается за счет определения множества программных API интерфейсов согласования языка irregex с различными библиотеками regex и доработкой результатов работы программ-нейроимитаторов за счет программной установки границы достоверности в соответствии с определенным пользователем коэффициентом вероятности отнесения сообщения к определенной категории;
4. повышение производительности классификации, за счет включения в архитектуру системы модуля 7 брокера конфигурации, выполненного с возможностью динамического воспроизведения - генерирования множества программных модулей конфигурации правил анализа и распределения их по множеству узлов для применения правил анализа к анализируемому тексту.

Техническая задача решается за счет системы классификации траффика для осуществления способа классификации траффика включающей координирующий модуль, модуль определения правил анализа, модуль машинного обучения, по крайней мере один модуль применения правил анализа к значащим полям сообщения, и дополнительно включающей интерфейс администратора (пользователя) системы который выполнен с возможностью определения категорий классификации сообщения, неформальных шаблонов-строк применяемых к анализу сообщения, вероятностных коэффициентов отнесения сообщения к определенной категории и дополнительно включает модуль компилятора правил анализа выполненный с возможностью создания программного модуля конфигурации правил анализа и состоящий из модуля языка регулярных выражений и модуля машинного обучения, причем модуль языка регулярных выражений

содержит в памяти и выполняет алгоритмы неформального языка регулярных выражений и по крайней мере одного формального языка регулярных выражений и выполнен с возможностью получения неформальной шаблона-строки и преобразования ее в регулярное выражение формального языка регулярных выражений, а модуль нейросети выполнен с возможностью получения вероятностного коэффициента отнесения сообщения к определенной категории и корректировки на его основе результата работы нейросети, а координирующий модуль выполнен с возможностью получения программного модуля конфигурации правил анализа, создания множества программных модулей конфигураций правил анализа и распределения их по множеству модулей применения правил анализа к значащим полям сообщения. При этом модуль машинного обучения содержит в памяти и выполняет алгоритм трансформации пороговой величины нейросети и установления границы достоверности выходных элементов нейросети в соответствии с полученным коэффициентом вероятности отнесения сообщения к определенной категории.

Изобретение поясняется чертежами:

На фиг.1 изображена функциональная схема системы классификации траффика.

На фиг.2 изображена общая схема a2p сервиса.

На фиг.2а приведена обобщенная последовательность SMPP TCP/IP вызовов способа классификации трафика в рамках a2p сервиса.

На фиг.2б приведена обобщенная последовательность SMPP TCP/IP вызовов способа классификации трафика в рамках p2a сервиса.

На фиг.3а приведен обобщенный алгоритм последовательно-параллельного режима работы системы классификации трафика.

На фиг.3б приведен обобщенный алгоритм параллельного режима работы системы классификации трафика.

На фиг.4 приведен обобщенный процесс преобразования данных системой классификации трафика.

Функциональная схема устройства реализующего предлагаемый способ сетевого семантического анализа представлена на фиг.1. Модуль 1 компилятора правил анализа включает модуль 2 regex и модуль 3 neural net. Основной функцией модуля 1 является сборка правил анализа в программный модуль конечной конфигурации правил анализа и

передача его в модуль 7 брокера конфигураций для дальнейшего воспроизведения заданного количества программных модулей и передачи их в аналитические узлы 8 для применения к сообщению.

В памяти модуля 2 regex хранятся данные и машинные команды с инструкциями процессорам скомпилированного неформального языка регулярных выражений - irregex и динамические и/или статические единицы трансляции, содержащие команды и данные по крайней мере одного известного из уровня техники стандартного языка регулярных выражений - regex. В качестве языков regex память может содержать несколько стандартных языков – это могут быть известные библиотеки regex языков Perl, Java, C++ и других известных из уровня техники. Согласование данных между языком irrigex и regex достигается переопределением программных интерфейсов (полиморфизмом) языка irregex и программной логикой преобразования сгенерированных структур irregex в структуры regex. Программная логика преобразования структур irrigex в структуры regex включена в функциональность неформального языка irrigex разработанного заявителем и являющимся его ноу-хау. Обобщенный метод преобразования структур irrigex в структуры regex заключается в согласовании API интерфейсов данных и методов их обработки языка irrigex и доступных в памяти модуля языков regex при сохранении семантики неформального шаблона-строки языка irrigex. Выбор конкретного языка regex может осуществляется администратором системы в модуле 5 настройки regex.

В памяти модуля 3 нейросети neural net хранятся данные нейросетевой структуры (нейронная сеть). Существует большое количество программных продуктов (программ-нейроимитаторов), предоставляющих инструменты для проектирования нейросетевых структур анализа текста – платформы Keras, SyntaxNet, Microsoft Cognitive Toolkit (CNTK) и другие известные из уровня техники. Нейронная сеть может быть спроектирована самостоятельно на известных языках программирования, такой вариант предпочтителен и используется заявителем т.к. предоставляет возможность программного преобразования (трансформации) пороговой величины нейросети в соответствии с пользовательским коэффициентом вероятности Pf полученным из модуля 6 настройки neural net т.к. не все программы-нейроимитаторы предоставляют возможность такого преобразования. В контексте предлагаемого способа классификации пороговая величина — это величина, определяющая разделение множества входных данных системы классификации траффика на подмножества, каждое из которых является определенной администратором системы семантической категорией классификации сообщения (рекламное, транзакционное, сервисное, международное, мошенническое и др.) $category_1$ фиг.4 1b. Как известно

разделение множества на классы (подмножества) должно подчиняться правилу попадания элемента входных данных только в одно подмножество, а объединение всех подмножеств должно совпадать с множеством входных данных. В зависимости от количества предустановленных категорий классификации или размерности пространства классификации выходные элементы нейронной сети в соответствии с функцией активации формируют пороговую величину разделения этого пространства. Например, в случае двух категорий классификации – рекламное и сервисное сообщение, пространство категорий представляет собой плоскость, а пороговая величина представляет собой линию разделяющую эту плоскость и формирующую подмножества категорий классификаций. Значения выходных нейронов определяют координаты распределения сообщений на плоскости категорий (например, рекламный траффик 0,5, сервисный 0,3). Если администратор системы предполагает, что в соответствии с шаблоном-строкой сообщение должно относится к рекламному траффику то он может установить для него преобладающий коэффициент вероятности $P_f$ фиг.4 1c. Программный алгоритм преобразования пороговой величины выполняет трансформацию линии - например сдвиг в пространстве двухмерных координат (категорий классификации) с учетом преобладающего влияния категории классификации для которой администратор системы установил больший коэффициент вероятности (доверия) Pf. После выполнения сдвига большее количество сообщений попадают в подмножество рекламной категории. Большее количество категорий классификации предполагает более сложную логику трансформации в общем случае аналогичную описанной выше. Трансформация пороговой величины возможна известными методами: преобразованием функции активации Fa (X, Pf) фиг.4 1c, 5a и/или изменением значений весов нейронов смещения, других параметров нейросети в соответствии с Pf, включает сложную логику, разработанную авторами, и является ноу-хау заявителя. Алгоритм преобразования пороговой величины предоставляет администратору системы возможность оперативного уточнения результатов классификации онлайн без необходимости переобучения и тестирования нейронной сети. В другом исполнении, например при использовании нейросетей сторонних вендоров на архитектуру которых в части преобразования пороговой величины влиять невозможно корректировка результатов классификации нейросети реализуется программным установлением допустимой границы достоверности классификации фиг.4 1c, 5b. В этом случае под доверительный интервал попадают результаты, превышающие установленную границу достоверности Pf. Обучение нейросети осуществляется по наборам учебных данным, также передающимся на вход нейросети из модуля 6.

Модуль 4 пользовательского интерфейса предназначен для взаимодействия с администратором (пользователем) системы и в различных вариантах может иметь интерфейс командной строки и/или интерфейс, исполненный в виде графических изображений, включает визуальные интерактивные средства контроля различных каналов трафика (SMPP, SMTP, HTTP и т.д.). Модуль 4 пользовательского интерфейса включает модуль 5 настройки regex предоставляющий администратору системы определять категории классификации сообщений фиг.4 1b например: транзакционная категория для выявления SMS запроса получения логина/пароля, рекламная категория для выявлений SMS с рекламными предложениями (оповещениями), сервисная категория для SMS взаимодействия пользователя с оператором СПРС, международная и др. Модуль 5 настройки regex предоставляет администратору определять упрощенные неформальные шаблоны-строки user template фиг.4 1a которые ему предпочтительны для применения к сообщению с учетом текущего формата синтаксиса и семантики сообщений для каждой определенной категории. Модуль 5 настройки regex предоставляет администратору определять коэффициент вероятности отнесения сообщения к определенной категории user factor Pf фиг.4 1c. Администратор системы определяет неформальные шаблоны-строки user template фиг.4 1a вне правил задания формата регулярных выражений языков regex. Коэффициент вероятности Pf жестко не «привязан» к неформальному шаблону-строке, т.к. при последующей обработке user template приводится в формат языков regex фиг.4 3., а Pf преобразуется в трансформацию пороговой величины фиг.4 5a, 5b. Применяются данные настройки к сообщению независимо фиг.4 8,10. Опосредованная связь их взаимовлияния на итоговую классификацию сообщения определяется тем, что обработка сообщения в модуле 3 neural net фиг.4 9 производится после применения правил модуля 2 regex фиг.4 8. Т.е. после классификации сообщения алгоритмами четкой логики regex по шаблону пользователя фиг.3а 6-10, 3б 5-9, последующий результат классификации neural net уточняется фиг. 3а 11, 3б 10 по Pf «коэффициенту доверия» фиг.3а, 15-18, 3б 14-17 пользователя к применению своего шаблона. Возможность ввода произвольного формата шаблонов-строк интуитивно по собственному усмотрению администратора системы, избавляет его от необходимости вникать в правила и диалекты языков regex, а также предоставляет возможность оперативного изменения параметра классификации (шаблона-строки) и установления доверия к нему Pf при мало прогнозируемом изменении формата сообщения (синтаксиса, семантики, схемы сообщения). В дальнейшем после преобразований в модуле 2 regex измененный шаблон-строка применяются к значащим полям field (атрибутам) сообщения фиг.4 8. Такими полями могут быть например данные отправителя, получателя, текст сообщения и могут

включать редко употребляемые фразы, выражения, символы, и т.д. Модуль настройки neural net 6 дополнительно включает интерфейс для задания настроек обучения и тестирования нейронной сети, а также хранилище данных для хранения исторического массива сообщений и обучающих данных. Строки-шаблоны и вероятностные коэффициенты модуль 4 интерфейса передает соответственно в модули 2, 3 модуля 1 компилятора анализа для дальнейшей обработки.

Модуль 7 - брокер конфигураций выполняющий функции координирующего модуля получает программный модуль конфигурации правил анализа сгенерированный компилятором правил анализа 1, воспроизводит необходимое множество конфигураций и распределяет их по аналитическим узлам 8. Программно-аппаратная логика брокера конфигураций создает и поддерживает требуемое количество аналитических узлов в зависимости от анализируемого системой типа трафика (SMS, USSD, mail, др.) и распределяет аналитические узлы в зависимости от нагрузки в дополнение к балансировщику нагрузки 9, распределяет между аналитическими узлами 8 программные модули конфигурации правил анализа, осуществляет их запись/перезапись в память аналитических узлов. Программно-аппаратная логика брокера конфигураций допускает обработку сообщения в рамках одной сетевой сессии (отправитель-получатель) несколькими аналитическими узлами и/или обработку одним аналитическим узлом сообщений разных установленных сетевых сессий. В общем случае логика распределения конфигурации правил анализа по аналитическим узлам 8 определяется типом анализируемого траффика, нагрузкой (количеством поступающих сообщений) по каждому типу траффика и синтаксической и семантической принадлежностью сгенерированной конфигурации. Например, менее подверженная изменениям сервисная структура SMS траффика (направление одноразового пароля-сервисная категория трафика) может анализироваться одним аналитическим узлом 8 в рамках разных сетевых сессий.

Аналитический узел 8 обрабатывает сообщения в соответствии с конфигурацией правил анализа, предоставляет интерфейсы для приема определенного типа трафика в соответствии с протоколом (SMPP для SMS, USSD, SMTP для mail, и др.). Сбор сообщения из сетевых пакетов, разбор по значащим полям и применение конфигурации правил анализа производится по командам и данным процесса-обработчика. Аналитический узел содержит пул таких процессов, логику их создания, размещения в памяти и удаления. Процесс-обработчик представляет собой совокупность выполняемых машинных команд, хранящихся в памяти и являющейся частью набора инструкций

процессорному устройству аналитического узла. Пул процессов обработчиков дополнительно обеспечивает балансировку нагрузки к мерам, применяемым логикой балансировщика нагрузки 9 и модуля конфигураций 7. После обработки сообщения аналитический узел классифицирует сообщение по принадлежности к предустановленной категории (рекламное, транзакционное, сервисное, международное, мошенническое и т.д.) и направляет сообщение получателю.

Балансировщик нагрузки 9 обеспечивает распределение сетевой нагрузки (переадресацию клиентских сетевых пакетов) между аналитическими узлами 8. Переадресация может быть реализована динамически на транспортном уровне известной технологией DNS-балансировки предусматривающей назначение нескольких IP-адресов одному доменному имени и выбора определенного IP-адреса в зависимости от правил балансировки (например алгоритм циклического обхода IP-адресов.)». В другой реализации балансировщик нагрузки 9 статически назначает клиентским IP адресам конкретные IP адреса аналитических узлов 8 (пример - Таненбаум Э., Уэзеролл Д. Компьютерные сети. 5-е изд. — СПб.: Питер, 2012. стр.783, 787.). В преимущественной реализации аналитические узлы 8 разнесены по разным узлам сети для обеспечения отказоустойчивости сервиса по технологии HA (High Availability) кластеров высокой доступности. В этом случае балансировщик нагрузки 9 может быть построен на основе прокси-серверов (маршрутизаторов), устанавливающих резервные сетевые соединения к аналитическим узлам 8 для обеспечения равномерного распределения клиентских запросов. Дополнительно балансировщик нагрузки 9 реализует переадресацию сообщений по типу прикладного протокола фиг.1 - сетевые пакеты SMTP протокола маршрутизируются на аналитические узлы 8 обрабатывающие SMS, USSD сообщения, пакеты SMPP адресуются на аналитические узлы обрабатывающие сообщения электронной почты. Клиентские сетевые узлы направляют сообщения на сетевой адрес балансировщика нагрузки 9 через обобщенный шлюз (SMS агрегатор для a2p сервиса или шлюз оператора PCEF для p2a). Физически балансировщик нагрузки 9 реализуется топологией сети включением в ее состав дополнительных маршрутизаторов с программной логикой маршрутизации сетевых пакетов. С целью упрощения фиг.1 эта архитектура функционально включена в модуль 9 и на фиг.1 не раскрыта.

Аппаратно модули 1, 4, 7, 8, 9 могут выполняться на одном сетевом узле или разных, могут присутствовать в базовой сети СПРС и/или вне ее. Сами модули 1,4,7,8,9 представляют собой комплекс программно-аппаратных средств (ЭВМ) фон Неймановской архитектуры универсального или специализированного вида и в общем случае содержат:

Контроллер сетевой карты реализованный в виде платы, содержащей процессор память и вспомогательную логику, выполненный с возможностью обработки входящих и исходящих сетевых пакетов со скоростью их передачи. Центральное процессорное устройство, может быть реализовано в одно/многопроцессорном варианте, управляет работой узла, выполняет инструкции и обрабатывает данные, хранимые в памяти. Память соответствующих узлов содержит преобразованные в внутреннее машинное представление (скомпилированные) блоки исходного кода (единицы компиляции): интерфейсного модуля, модуля неформального и формального языков регулярных выражений regex, нейронной сети neural net, брокера конфигураций, аналитических узлов. В дополнение к этому модуль 4 интерфейса содержит видеоконтроллер - микросхему, формирующую видеоизображение на мониторе администратора системы. Монитор выполнен с возможностью отображать изображение включающее графический интерфейс и/или интерфейс командной строки предоставляющий инструменты управления каналами трафика (SMPP, SMTP, другими известными из уровня техники), определения категорий классификации сообщений, задания и отображения упрощенных пользовательских строк-шаблонов поиска подстроки в строке, отображения ввода администратором системы вероятностных коэффициентов, визуального контроля обучением и тестированием нейронной сети. Устройства ввода (клавиатура, мышь, др.) и контроллер устройства ввода, отвечающий за обработку данных, введенных администратором с устройств ввода. Хранилище данных (жесткий диск) модуля 4 интерфейса предназначен для хранения исторических данных с целью их постобработки модулем 3 neural net, обучающей выборки и других данных, и контроллер жесткого диска, предназначенный для приема-передачи и обработки данных от хранилища данных. Шина-совокупность проводников соединяющих ЦПУ с памятью, устройствами хранения данных и ввода-вывода.

Фиг.2 иллюстрирует общую схему a2p сервиса, как понятно специалисту, не ограничиваясь фиг.2 изобретение может быть использовано в p2a, a2a, p2p сервисах. На фиг.2 в качестве SMS направляемых абоненту могут выступать банковские SMS рассылки, рекламные SMS предложения, SMS уведомления социальных сетей и любые другие известные из уровня техники. TSC (Traffic Classification System) – система классификации траффика мониторит все SMS сообщения переданные в сеть программами генерирующими SMS рассылку и выполняющихся на сетевых узлах - $ESME_1$ – $ESME_{n-1}$. В качестве сетевых узлов могут выступать сервера всевозможных хозяйствующих субъектов - банки, магазины и др. (ESMEn).

На фиг.2а приведена обобщенная последовательность SMPP TCP/IP вызовов a2p траффика для получения, анализа и отправки SMS сообщений SMPP протокола (спецификация 3GPP 3G TS 23.039). Пример иллюстрирует SMS сообщения в направлении от внешнего сетевого узла ESME на MS абонента системы подвижной радиосвязи (СПРС).

Пунктирной линией показаны сигнальные вызовы, сплошной вызовы TCP/IP.

На этапах 1-2 происходит стандартный сигнальный диалог установления соединения между ESME и SMSC абонента СПРС сигнальные сообщения SMPP протокола по известным правилам укладываются в стэк протоколов OSI и передаются по TCP/IP протоколу в сеть. После установления соединения внешний узел ESME генерирующий SMS сообщение по тем же правилам укладывает SMS сообщение deliver_sm, data_sm в стек протоколов OSI, фрагментирует на TCP/IP пакеты и направляет в сеть. По определению TCP/IP пакеты направляются на IP адрес балансировщика нагрузки системы классификации траффика фиг.1 9. Обычно перенаправление SMS сообщений от различных ESME на IP адрес балансировщика нагрузки может быть реализовано через сетевого посредника, который агрегирует SMS сообщения от различных ESME (т.н. SMS агрегаторы, на фиг. не показано). Балансировщик нагрузки в соответствии с алгоритмом переадресации направляет пакеты на IP адрес соответствующего аналитического узла фиг.1 8. Процесс обработчик аналитического TCS применяет алгоритм анализа к SMS сообщению (SMS PDU analysis and processing) после чего передает SMS в сеть на SMSC (SMS Center) СПРС абонента 4, а признак классификации сообщения направляет на заинтересованный узел СПРС, например на узел биллинга SCP 5. Далее происходит стандартный сигнальный обмен доставки сообщения 6-7 и подтверждения доставки 8. Признак классификации сообщения classification mark к определенному виду траффика оператор СПРС использует для последующих бизнес решений. В другой реализации TCS сохраняет все полученные SMS сообщения в хранилище данных интерфейса пользователя 4 и осуществляет их постобработку модулями 2 regex и 3 neural net 3, после чего передает результаты классификации на заинтересованный узел оператора. Этот режим может применяться в дополнение к онлайн классификации для уточнения результатов классификации, полученных в режиме реального времени на фиг.2а не показан.

На фиг.2б приведен примерSMPP TCP/IP вызовов p2a траффика для получения, анализа и отправки SMS сообщений SMPP протокола (спецификация 3GPP 3G TS 23.039).

На примере SMS сообщения в направлении от SMSC оператора СПРС на внешний узел ESME.

Пунктирной линией показаны сигнальные вызовы, сплошной вызовы TCP/IP.

Клиентами системы классификации трафика в варианте классификации SMS сообщений выступают преимущественно операторы СПРС, а сетевыми узлами являются мобильные станции абонентов (MS Mobil Station), узлы обслуживания коротких сообщений (SMSC Short Message Service Center) оператора СПРС, службы или приложения ДВО (VAS Value Added Services) операторов СПРС либо иные внешние узлы (ESME External Short Message Entities). Клиенты направляют SMS сообщения на сетевой IP адрес балансировщика нагрузки 9 который выполняет алгоритм маршрутизации сетевых пакетов для балансировки нагрузки (на фиг.3а, 3б не показано). Перенаправление клиентских пакетов может быть реализовано на шлюзовом узле PCEF оператора СПРС.

Абонент СПРС с своего MS отправляет SMS сообщение, производится ряд вызовов 1-2 по сигнальным протоколам для доставки исходящего SMS в SMSC. SMSC начинает типичную SMPP последовательность запроса/ответа в качестве отправителя SMS на внешний узел ESME 3-4. После установления соединения 3-4 SMSC укладывает SMS сообщение deliver_sm, data_sm в стек протоколов OSI, фрагментирует на TCP/IP пакеты и направляет в сеть через шлюз СПРС PCEF 5. PCEF на транспортном уровне заменяет адрес получателя IP TCP пакета на IP адрес балансировщика нагрузки 9 системы классификации траффика 6. После маршрутизации пакетов балансировщиком нагрузки 9 пакеты приходят на сетевой интерфейс аналитического узла 8 TCS где собираются в SMS PDU и далее производится анализ и обработка (SMS PDU analysis and processing). После классификации сообщения аналитический узел направляет сообщение на IP адрес ESME 7, а в качестве адреса отправителя sender IP устанавливает адрес шлюза СПРС PCEF. После получения ESME SMS сообщения производится сигнальный диалог подтверждения получения SMS 8, а система классификации траффика TCS (Traffic Classification System) направляет в адрес заинтересованного узла (например, на узел биллинга SCP) признак классификации SMS сообщения classification mark 8. Постобработка сохраненного массива сообщений в рамках p2a траффика может проводится аналогично описанной выше для a2p траффика.

Фиг.3а иллюстрирует обобщенный алгоритм работы системы классификации траффика на примере последовательно-параллельного режима работы системы классификации трафика. На этапе 1 назначенный для обработки SMS сообщений,

аналитический узел 8 принимает SMS сообщение - выполняется захват сетевых пакетов, агрегирование пакетов в потоки, классификация пакетов (потоков) по протоколу прикладного уровня, извлечение данных и другие стандартные операции DPI анализа траффика. На этапе 2 менеджер пула процессов инициирует процесс-обработчик, который проверяет размер сообщения, разбирает сообщение по значащим полям для анализа (текст, отправитель получатель, служебные данные и др.) формирует соответствующую структуру данных с разделением на значащие поля и размещает ее в памяти аналитического узла 8. На этапе 3 программная логика аналитического узла предусматривает передачу структуры сообщения в модуль 4 интерфейса для записи в хранилище данных 4 с целью накопления исторических данных и постобработки. Этап 5 предлагает администратору выбор алгоритма классификации алгоритмами четкой логики Regex или эвристическим алгоритмом Neural net. На этапе 6 администратору системы предоставляется опция изменения правил анализа сообщения. Например, в случае если в сообщениях в случайном порядке, не возможном для формализации (выражения в программной логике алгоритма) изменяется синтаксис или семантика сообщения, появляются символы другого алфавита и/или появляются новые слова и фразы, заимствованные из других языков (например дедлайн, пиар, зум, др.). В этом случае на этапе 7 администратор системы по своему усмотрению вводит упрощенные интуитивно понятные ему шаблоны-строки, которые с его точки зрения приемлемы для распознавания изменившейся семантики сообщения. Ввод неформальных шаблонов-строк осуществляется техническими средствами модуля 5 настройки regex интерфейса пользователя - администратор определяет неформальные строки-шаблоны user template. Одновременно он может определять вероятностные коэффициенты user factor Pf отнесения сообщений к предустановленной категории - виду траффика (category $_I$), фиг.3а, 3б. Контроль корректности ввода администратор осуществляет на мониторе по визуальным средствам контроля (графическому интерфейсу) отрисованным видеоконтроллером модуля интерфейса пользователя 4. На этапе 8 преобразованные в машинное представление данные пользовательских шаблонов передаются и размещаются в памяти модуля 2 regex и далее по набору машинных инструкций неформального языка регулярных выражений irrigex модуль 2 выполняет преобразование пользовательских шаблонов в регулярные выражения уже формального языка regex. Преобразование включает программное генерирование структуры данных irrigex struct неформального языка irrigex и преобразование ее в структуру regex struct соответствующую шаблонам (регулярным выражениям) уже формального языка регулярных выражений regex. На этапе 9 из структуры regex struct компилятор правил анализа 1 формирует конечную

конфигурацию правил анализа config - программный модуль представляющий собой файл объектного кода с инструкциями по анализу значащих полей сообщения и передает его для записи в память брокеру конфигураций 7 который воспроизводит назначенное множество копий файла объектного кода конфигурации и передает ее назначенным аналитическим узлам 8. На этапе 10 аналитический узел 8 размещает конфигурацию в памяти соответствующего процесса-обработчика, который применяет ее для обработки текста значащих полей сообщения после чего генерирует признак отнесения сообщения к предопределенной категории, виду трафика – этап 11. Далее на этапе 12 значащие поля сообщения string могут передаваться в модуль 3 neural net на вход нейронной сети (этапы 14-18) для уточнения классификации (этап 11) или сообщение передается в сеть для отправки получателю 13. Такая «петля» позволяет уточнить категорию сообщения онлайн средствами neural net после применения правил regex. В случае передачи значащих полей сообщения string на вход нейронной сети, модуль 3 neural net использует машинное представление коэффициента вероятности Pf полученный из модуля 6 настройки neural net этап 15. После отработки алгоритма классификации нейронной сетью этапы 16, 17, выполняется программный алгоритм преобразования пороговой величины нейронной сети по коэффициенту Pf – этап 18. Далее по сгенерированному модулем 3 neural net признаку классификации процесс обработчик уточняет категорию сообщения и далее передает в сеть для отправки - этапы 11,12. В рамках одного процесса - обработчика при выполнении этапов 14-18 этапы 1-10 применяются к очередному сообщению, этим достигается последовательно-параллельная обработка сообщений.

Фиг.3б иллюстрирует блок схему параллельной обработки сообщения. В этом варианте после этапа 3 к значащим полям сообщения string параллельно применяются алгоритмы четкой классификации на основе правил Regex – этапы 5-9 и эвристические алгоритмы классификации нейронной сети Neural net – этапы 13-17. На этапе 10 программный алгоритм процесса обработчика дожидается завершения выполнения обоих веток алгоритма, этапы 13-17 предполагают передачу (копирование) значащих полей сообщения string в модуль 3 neural net и ожидание выполнения алгоритма классификации нейросети. После чего формируется признак классификации сообщения (программная логика этапов 5-9 и 13-17 выполняются аналогично варианту последовательно-параллельной обработки). Далее сообщение и признак классификации направляются получателям – этапы 10-12.

Фиг.4 иллюстрирует схему обработки данных системой классификации траффика. Поток сообщений SMS pdu flow после отработки логики балансировки сетевой нагрузки

модулем 9 (на фиг.4 не показана) поступает на сетевой интерфейс аналитического узла 8. По программному интерфейсу сокетов (адрес сетевого узла и назначенный порт процесса-обработчика) аналитический узел принимает сетевой пакет 6 выполняет стандартные операции (DPI Deep Packet Inspection) классификации пакетов, агрегирования их в потоки и по программному алгоритму процесса обработчика генерирует структуру данных SMS struct 7 представляющую собой машинное представление значащих данных SMS сообщения – длина length, отправитель sender, получатель receiver, текст сообщения - строка string, включающая значимые для анализа и классификации семантические поля $field_n$. Модуль интерфейса пользователя 4 содержит в памяти определенные администратором системы данные: категории классификации сообщения – $category_l$ 1b, неформальный шаблон-строку – user template 1a, коэффициент вероятности отнесения сообщения к определенной категории – user factor Pf 1c. Модуль 2 regex на основе неформального шаблона-строки 1a по программному алгоритму неформального языка регулярных выражений irrigex генерирует структуру irrigex struct 2 включающую данные и методы для различных операций таких как поиск вхождения шаблона, выполнение замены строки и др. и сохраняет ее в памяти модуля. В соответствии с настройками пользователя по выбранной библиотеке стандартного языка regex, программный алгоритм модуля 2 преобразует данные структуры неформального языка irrigex struct 2 в структуру данных соответствующего уже стандартного языка регулярных выражений regex struct 3 определяющую правила анализа текста. Программный алгоритм компилятора правил анализа 1 собирает файл config 4 - программный модуль конфигурации правил анализа включающий данные и методы их обработки $methods_k(..)$ структуры regex struct 3. Брокер конфигурации 7 воспроизводит (копирует) заданное программным алгоритмом количество файлов конфигурации config и распределяет их по аналитическим узлам 8 (на фиг.4 не показано). Процесс-обработчик по программному алгоритму применяет правила анализа файла конфигурации config к значащим полям сообщения string 8 и по результату генерирует признак классификации, т.е. относит сообщение к одной из предустановленных категорий $category_l$ 1б. Для классификации сообщения нейронной сетью процесс обработчик передает значащие поля сообщения string в модуль 3 neural net. Значащие поля сообщения $field_n$ подаются на вход входным элементам нейросети 9, после выполнения логики классификации выходные элементы $X_l$ нейронной сети генерируют соответствующие коэффициенты вероятности $Fa_l$ ($X_l$) отнесения сообщения к определенной категории l. Программный алгоритм преобразования пороговой величины модуля 3 neural net преобразует функцию выходных элементов $X_l$ (пороговую величину) с учетом коэффициента Pf 1c - $Fa_l$ ($X_l$, Pf) 5a. Этим обеспечивается влияние

пользовательского коэффициента вероятности Pf отнесения сообщения к определенной категории на результат классификации. После чего формируется признак классификации сообщения, отнесения его к определенной категории 10. В другом варианте работы 5b программный алгоритм модуля 3 neural net присваивает признак классификации сообщения той категории $category_l$, пороговая величина которой $P_{l-1}$ соответствующего выходного нейрона $X_l$ преобладает и превышает доверительный интервал, установленный пользовательским коэффициентом Pf.

Представленная архитектура системы классификации траффика, предполагает разнесение модулей по разным узлам сети. Быстродействие системы достигается за счет обработки сообщений правилами regex только на аналитическом узле 8. Масштабирование производительности наряду с известными мерами, принимаемыми балансировщиком нагрузки достигается воспроизведением и перераспределением конфигурации правил анализа брокером конфигураций по аналитическим узлам и поддержанием их необходимого количества. Архитектура системы позволяет администратору выполнять этапы 6-10, 14-18 фиг.3а и 5-9, 13-17 фиг.3б в последовательно-параллельном и параллельном режиме соответственно. Архитектура системы позволяет применять пользовательские настройки - изменение режима работы, отключение/включение обработки regex и/или neural net, ввод шаблонов и коэффициентов вероятности, без прерывания анализа потока сообщений онлайн.

Основное свойство и преимущество предлагаемой авторами системы состоит в гибкости и универсальности настройки классификации обычными администраторами сетевых узлов за счет применения авторского неформального языка регулярных выражений irregex, выполняющего свою логику в модуле 3 regex. Функционал упрощенного ввода шаблонов распознавания текста позволяет снять ограничения на обслуживание устройства узконаправленными высококлассными специалистами и предоставить возможность его эксплуатации обычным телекоммуникационным специалистам. В настоящее время в известных решениях анализа текста при изменении схемы сообщения разработчики решений семантического анализа неизбежно вынуждены заниматься трудоемкой, не всегда возможной доработкой архитектуры решения семантического анализа. Случайное изменение схемы и структуры сообщения в начальных настройках сервера по умолчанию определено быть не может, а предусмотреть формализовать все возможные форматы сообщений не представляется возможным по понятным причинам. Доступный для настройки функционал модуля 4 интерфейса пользователя переводит предлагаемый авторами сервер из разряда специализированного

оборудования в разряд узлов телекоммуникационной сети доступных для настройки обычными специалистами в телекоммуникациях. Дополнительно включение в архитектуру системы неформального языка irregex позволило взаимодействовать с любыми различными стандартными языками regex, подключать их для анализа в соответствии с предпочтениями пользователя. Более того, включение в процесс классификации пользовательских вероятностных коэффициентов, и выполнение модуля neural net с возможностью корректировки результатов работы нейросети (собственной архитектуры или сторонних инструментов) позволяет администратору системы быстро реагировать на изменение схемы и семантики сообщения и оперативно «подправлять» результаты нейронной классификации. Причем способ классификации, как в последовательно-параллельном режиме, так и в параллельном, обеспечивает существенный прирост точности классификации за счет синергии алгоритмов regex и neural net. Что особенно важно способ и архитектура системы классификации траффика обеспечивает опосредованное взаимовлияния пользовательских настроек, при том, что администратор системы может настраивать их независимо.

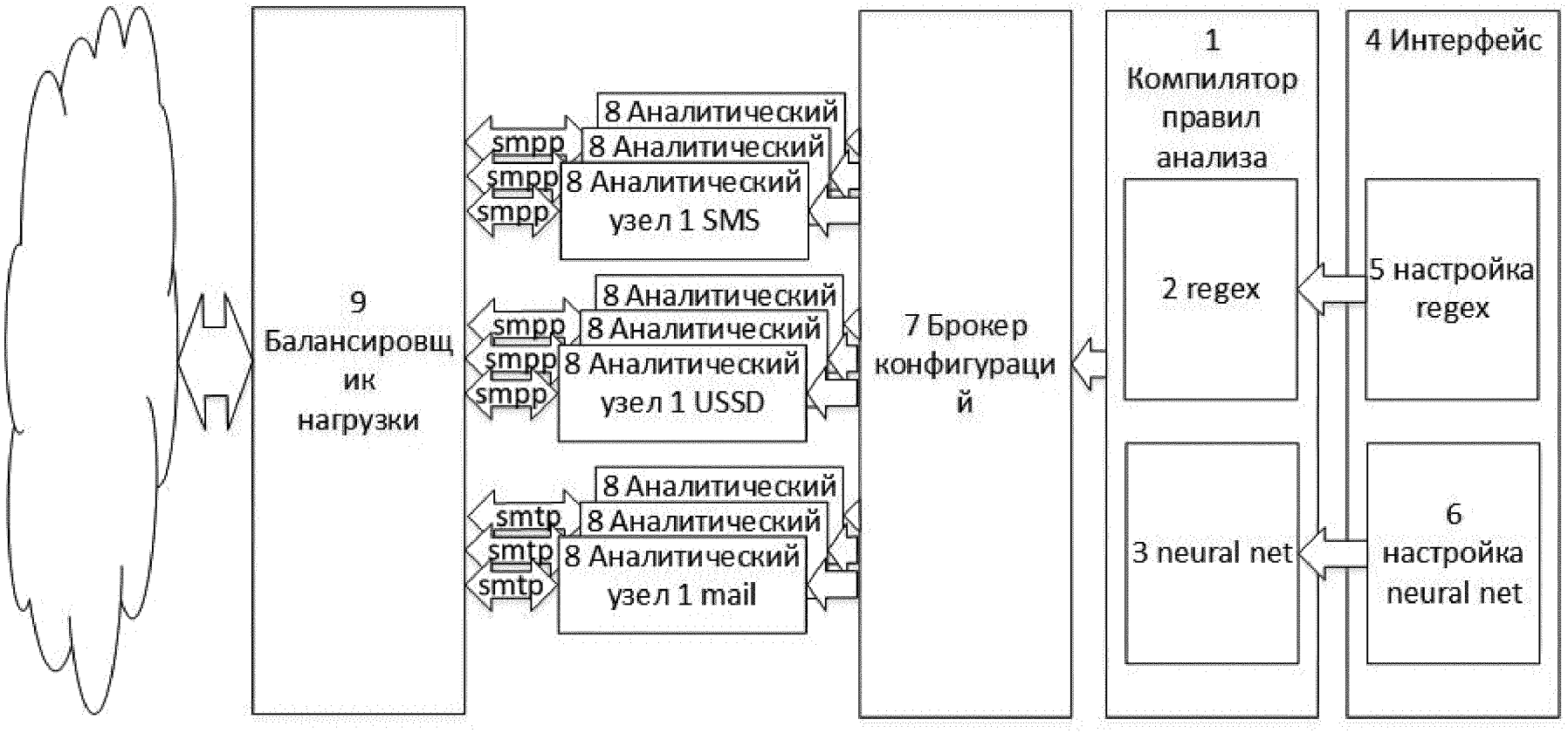
Предлагаемая система классификации траффика продемонстрирована выше на примере SMPP траффика в рамках взаимодействия a2p, p2a. Как понятно специалистам предлагаемое решение может быть использовано для семантического анализа потока сообщений p2p, a2a сервисов, преимущественно небольших объемов текста. А также иных стандартов и приложений (в том числе например спецификации SMPT для сервисов электронной почты mail), не выходя при этом за рамки правовой охраны предлагаемого способа классификации траффика и архитектуры устройства для его осуществления.

Предлагаемое изобретение испытано в ПАО Мегафон и ПАО ВымпелКом, в результате работы подтвержден заявленный технический эффект.

# Формула

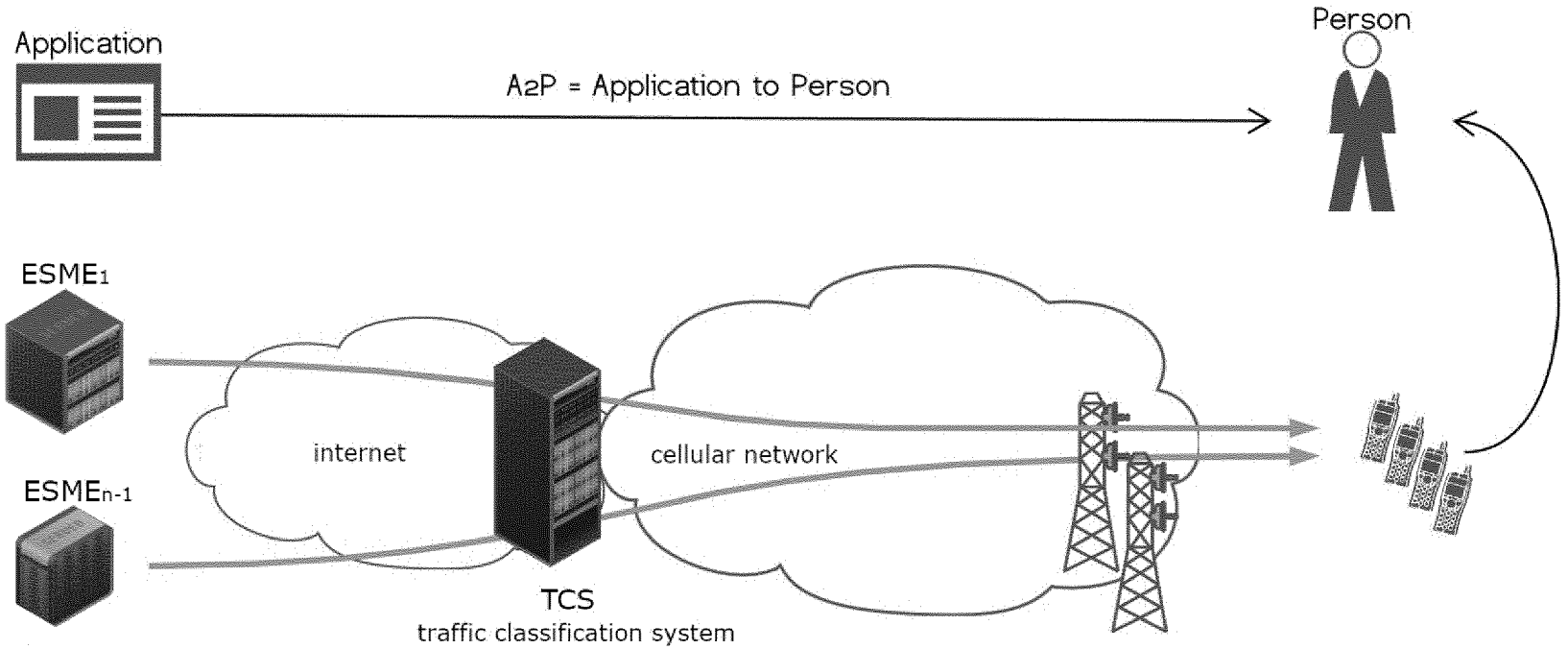
1. Способ классификации трафика, в соответствии с которым принимают сообщение естественного языка, определяют схему сообщения, делят сообщение на значащие синтаксические и семантические поля, предоставляют множество узлов для анализа значащих полей сообщения, к которым применяют правила анализа текста в зависимости от прикладного протокола сообщения, отличающийся тем, что для классификации сообщения по определенным категориям определяют неформальный шаблон-строку неформального языка поиска и обработки подстроки в строке и вероятностный коэффициент отнесения сообщения к определенной категории, преобразуют неформальный шаблон-строку в регулярное выражение стандартного языка поиска и обработки подстроки в строке, преобразуют регулярное выражение стандартного языка в программный модуль конфигурации правил анализа текста содержащий данные и методы обработки строк, динамически создают множество программных модулей конфигурации правил анализа и распределяют их по множеству узлов для анализа значащих полей сообщения, применяют конфигурацию правил анализа к значащим полям сообщения, подают сообщение на вход нейронной сети, причем результат классификации сообщения нейронной сетью корректируют в соответствии с определенным вероятностным коэффициентом отнесения сообщения к определенной категории.
2. Способ классификации трафика по п.1 отличающийся тем, что определение неформального шаблона-строки неформального языка поиска и обработки подстроки в строке и вероятностного коэффициента отнесения сообщения к определенной категории производят в процессе обработки сообщения.
3. Способ классификации трафика по п.1 отличающийся тем, что множество программных модулей конфигурации правил анализа генерируют и распределяют по множеству узлов для анализа значащих полей сообщения пропорционально сетевой нагрузке, обусловленной количеством анализируемых сообщений.
4. Способ классификации трафика по п.1 отличающийся тем, что результат классификации нейронной сети корректируют в соответствии с определенным вероятностным коэффициентом отнесения сообщения к определенной категории, трансформацией пороговой величины классификации нейросети.

5. Способ классификации трафика по п.1 отличающийся тем, что результат классификации нейронной сети, корректируют в соответствии с определенным вероятностным коэффициентом отнесения сообщения к определенной категории, установлением границы достоверности выходных элементов нейросети.
6. Система классификации траффика включающая координирующий модуль, модуль определения правил анализа, модуль машинного обучения, по крайней мере один модуль применения правил анализа к значащим полям сообщения, отличающаяся тем, что дополнительно включает интерфейс администратора или пользователя системы выполненный с возможностью определения категорий классификации сообщения, неформальных шаблонов-строк применяемых к анализу сообщения, вероятностных коэффициентов отнесения сообщения к определенной категории и дополнительно включает модуль компилятора правил анализа выполненный с возможностью создания программного модуля конфигурации правил анализа и состоящий из модуля языка регулярных выражений и модуля нейросети, причем модуль языка регулярных выражений содержит в памяти и выполняет алгоритмы неформального языка регулярных выражений и по крайней мере одного формального языка регулярных выражений и выполнен с возможностью получения неформальной шаблона-строки и преобразования ее в регулярное выражение формального языка регулярных выражений, а модуль нейросети выполнен с возможностью получения вероятностного коэффициента отнесения сообщения к определенной категории и корректировки на его основе результата работы нейросети, а координирующий модуль выполнен с возможностью получения программного модуля конфигурации правил анализа, создания множества программных модулей конфигураций правил анализа и распределения их по множеству модулей применения правил анализа к значащим полям сообщения.
7. Система классификации траффика по п.6 отличающаяся тем, что модуль нейросети содержит в памяти и выполняет алгоритм трансформации пороговой величины нейросети в соответствии с полученным коэффициентом вероятности отнесения сообщения к определенной категории.
8. Система классификации траффика по п.6 отличающаяся тем, что модуль нейросети содержит в памяти и выполняет алгоритм установления границы достоверности выходных элементов нейросети в соответствии с полученным коэффициентом отнесения сообщения к определенной категории.

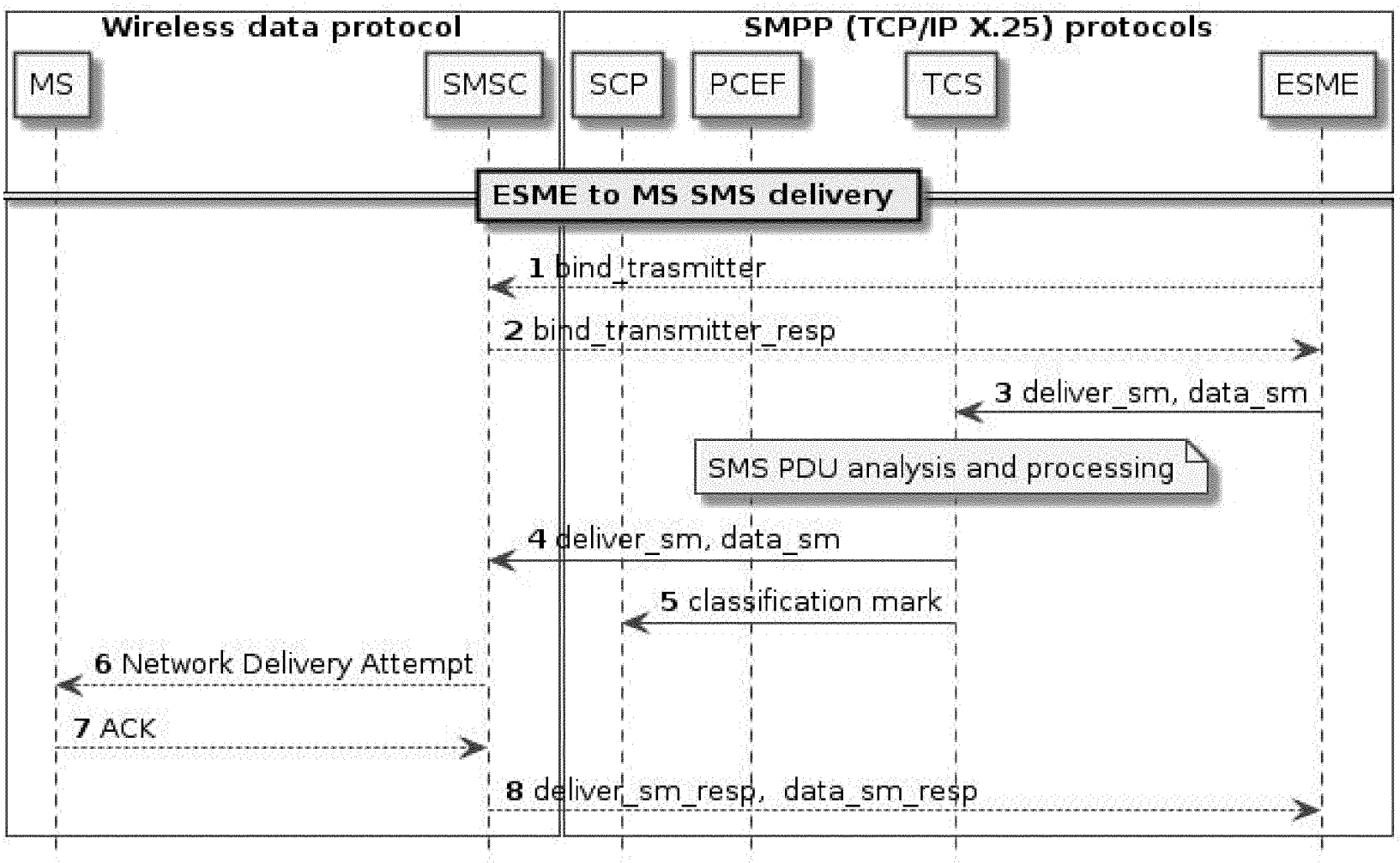



Система классификации траффика TCS (Traffic Classification System)

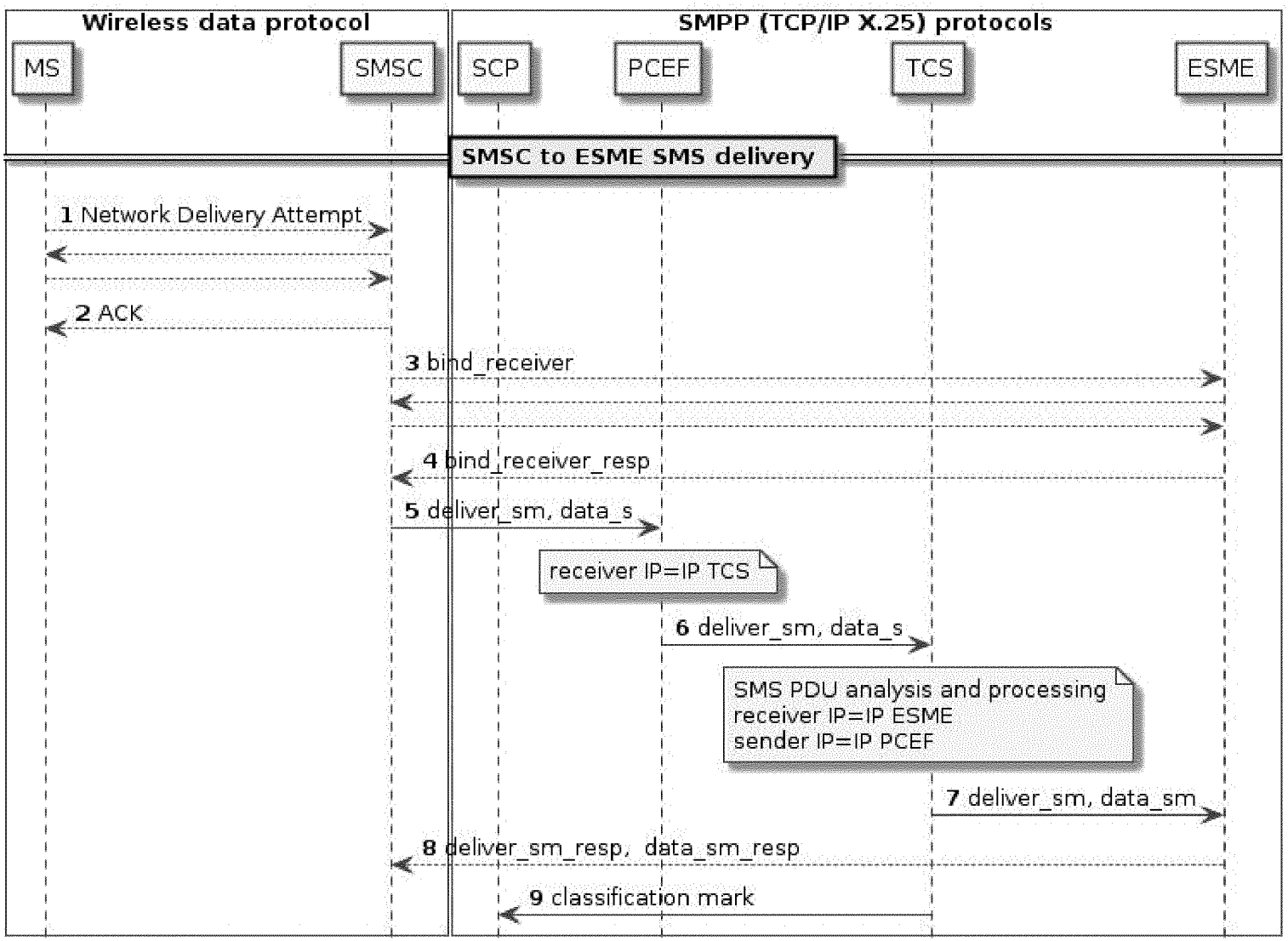
Фиг.1


Application
A2P = Application to Person
Person
ESME1
ESMEn-1
internet
cellular network
TCS
traffic classification system

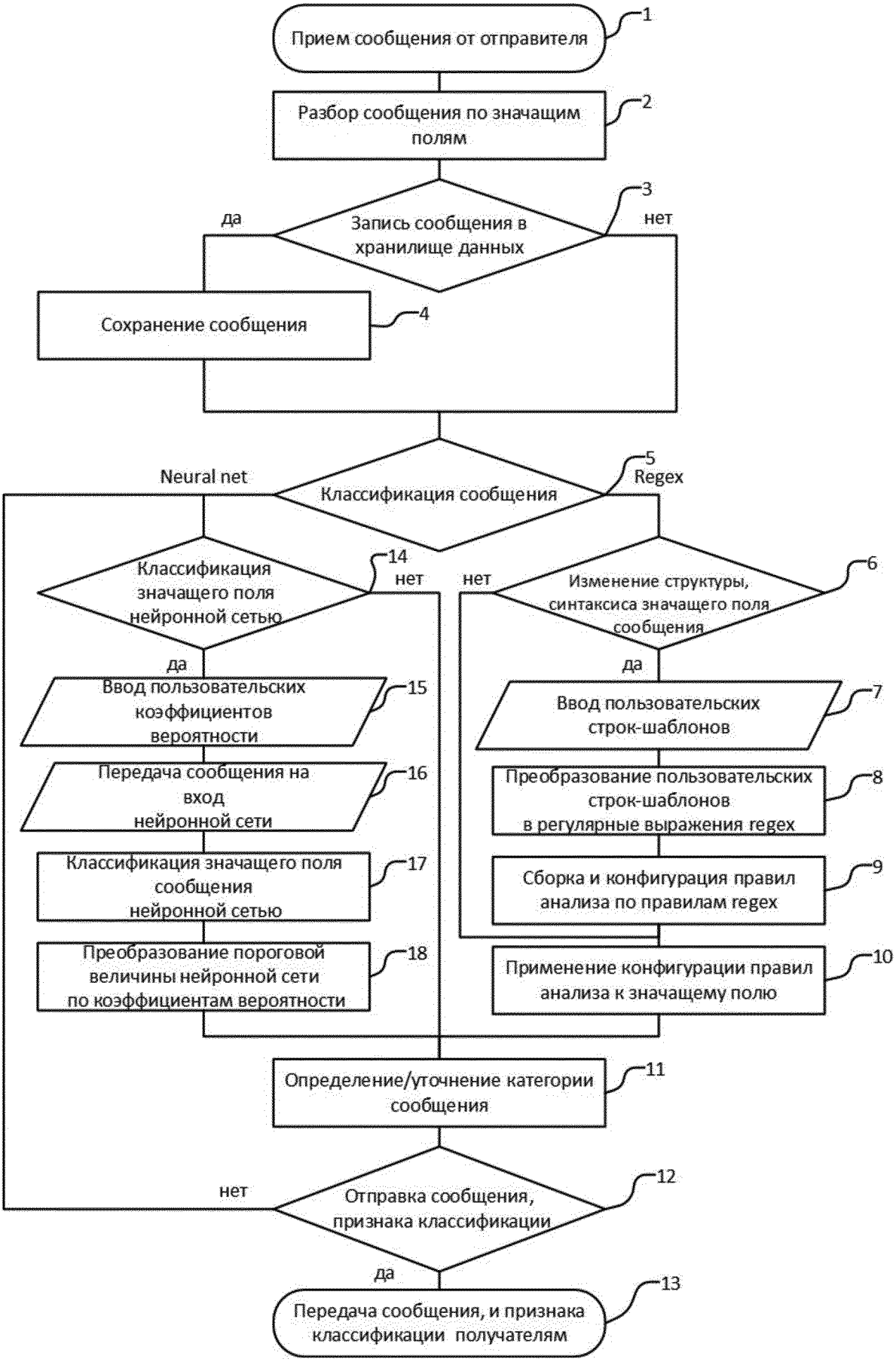
Фиг.2


Wireless data protocol
SMPP (TCP/IP X.25) protocols
MS
SMSC
SCP
PCEF
TCS
ESME
ESME to MS SMS delivery
1 bind_trasmitter
2 bind_transmitter_resp
3 deliver_sm, data_sm
SMS PDU analysis and processing
4 deliver_sm, data_sm
5 classification mark
6 Network Delivery Attempt
7 ACK
8 deliver_sm_resp, data_sm_resp

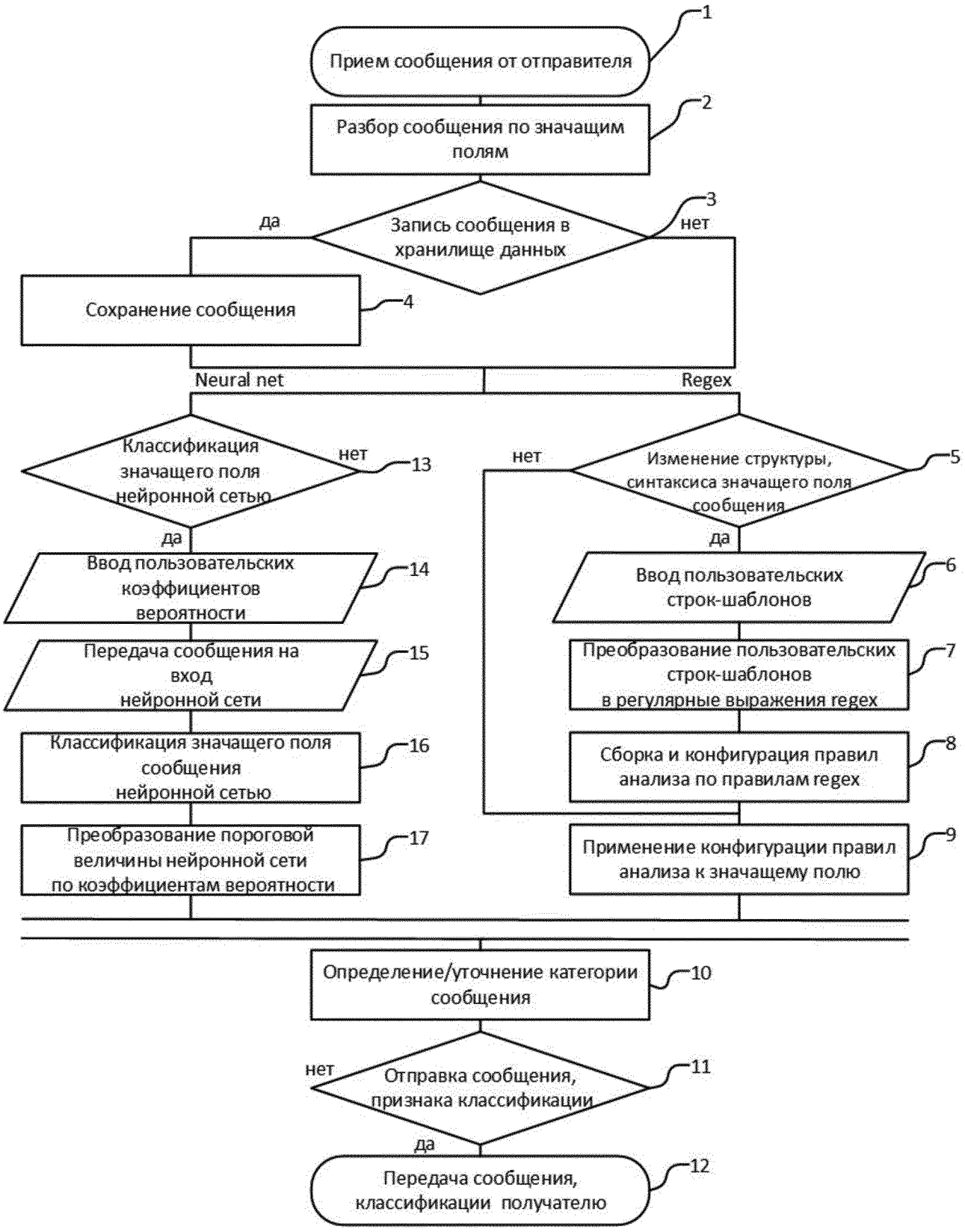
Фиг.2а


Wireless data protocol
SMPP (TCP/IP X.25) protocols
MS
SMSC
SCP
PCEF
TCS
ESME
SMSC to ESME SMS delivery
1 Network Delivery Attempt
2 ACK
3 bind_receiver
4 bind_receiver_resp
5 deliver_sm, data_s
receiver IP=IP TCS
6 deliver_sm, data_s
SMS PDU analysis and processing
receiver IP=IP ESME
sender IP=IP PCEF
7 deliver_sm, data_sm
8 deliver_sm_resp, data_sm_resp
9 classification mark

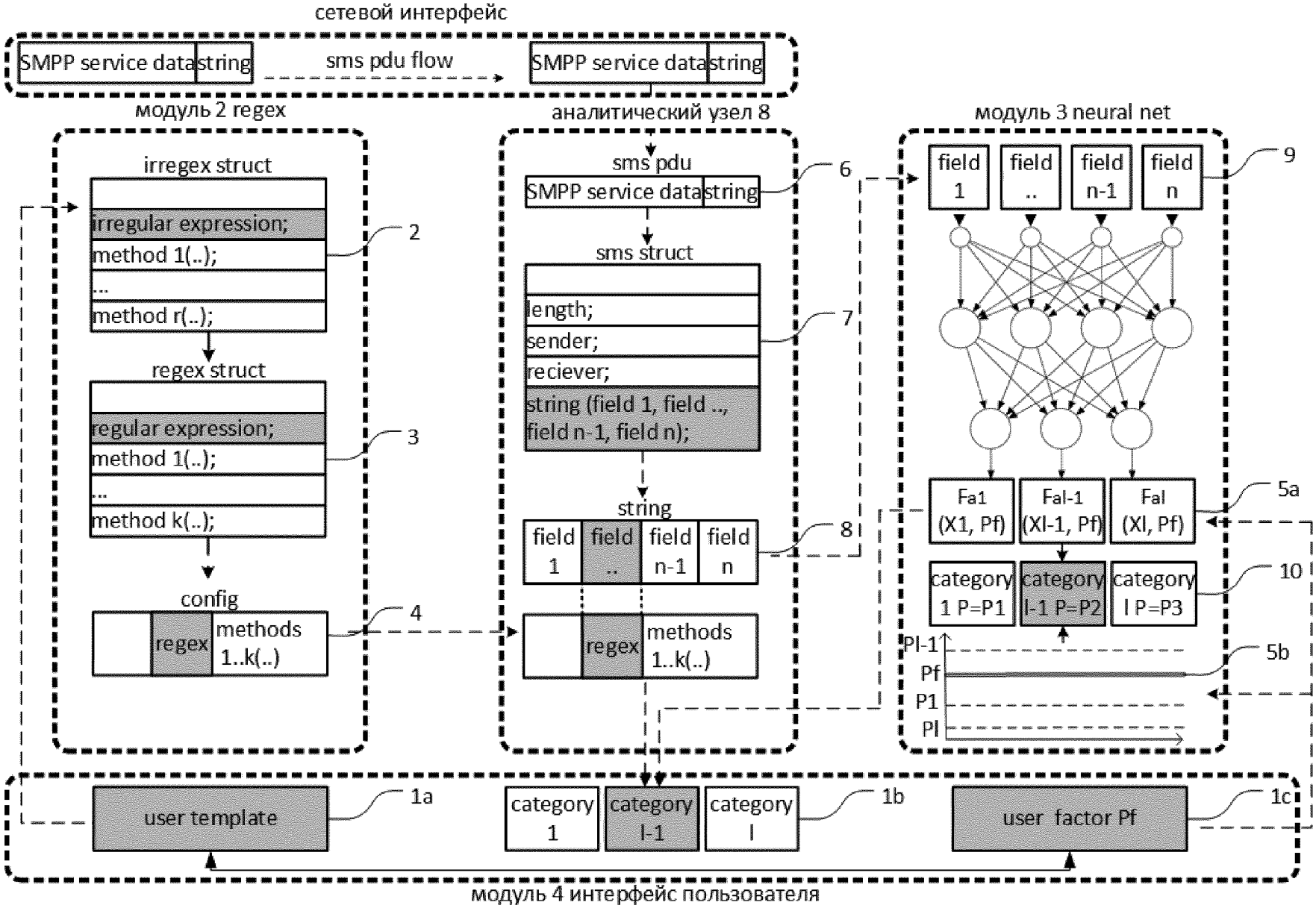
Фиг.2б


Прием сообщения от отправителя
1
Разбор сообщения по значащим полям
2
Запись сообщения в хранилище данных
3
да
нет
Сохранение сообщения
4
Классификация сообщения
5
Neural net
Regex
Классификация значащего поля нейронной сетью
14
нет
нет
Изменение структуры, синтаксиса значащего поля сообщения
6
да
да
Ввод пользовательских коэффициентов вероятности
15
Ввод пользовательских строк-шаблонов
7
Передача сообщения на вход нейронной сети
16
Преобразование пользовательских строк-шаблонов в регулярные выражения regex
8
Классификация значащего поля сообщения нейронной сетью
17
Сборка и конфигурация правил анализа по правилам regex
9
Преобразование пороговой величины нейронной сети по коэффициентам вероятности
18
Применение конфигурации правил анализа к значащему полю
10
Определение/уточнение категории сообщения
11
нет
Отправка сообщения, признака классификации
12
да
Передача сообщения, и признака классификации получателям
13

Фиг.3а


1
Прием сообщения от отправителя
2
Разбор сообщения по значащим полям
3
Запись сообщения в хранилище данных
да
нет
Сохранение сообщения
4
Neural net
Regex
Классификация значащего поля нейронной сетью
нет
13
да
Ввод пользовательских коэффициентов вероятности
14
Передача сообщения на вход нейронной сети
15
Классификация значащего поля сообщения нейронной сетью
16
Преобразование пороговой величины нейронной сети по коэффициентам вероятности
17
нет
Изменение структуры, синтаксиса значащего поля сообщения
5
да
Ввод пользовательских строк-шаблонов
6
Преобразование пользовательских строк-шаблонов в регулярные выражения regex
7
Сборка и конфигурация правил анализа по правилам regex
8
Применение конфигурации правил анализа к значащему полю
9
Определение/уточнение категории сообщения
10
нет
Отправка сообщения, признака классификации
11
да
Передача сообщения, классификации получателю
12

Фиг.3б


сетевой интерфейс
SMPP service data
string
sms pdu flow
SMPP service data
string
модуль 2 regex
irregex struct
irregular expression;
method 1(..);
...
method r(..);
2
regex struct
regular expression;
method 1(..);
...
method k(..);
3
config
regex
methods 1..k(..)
4
аналитический узел 8
sms pdu
SMPP service data
string
6
sms struct
length;
sender;
reciever;
string (field 1, field .., field n-1, field n);
7
string
field 1
field ..
field n-1
field n
8
regex
methods 1..k(..)
модуль 3 neural net
field 1
field ..
field n-1
field n
9
Fa1 (X1, Pf)
Fal-1 (Xl-1, Pf)
Fal (Xl, Pf)
5a
category 1 P=P1
category l-1 P=P2
category l P=P3
10
Pl-1
Pf
P1
Pl
5b
user template
1a
category 1
category l-1
category l
1b
user factor Pf
1c
модуль 4 интерфейс пользователя

Фиг.4