



ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ПАТЕНТУ

(21)(22) Заявка: 2016125278, 24.06.2016

(24) Дата начала отсчета срока действия патента:
24.06.2016

Дата регистрации:
24.10.2017

Приоритет(ы):

(22) Дата подачи заявки: 24.06.2016

(45) Опубликовано: 24.10.2017 Бюл. № 30

Адрес для переписки:

125212, Москва, Ленинградское ш., 39а, стр. 3,
АО Лаборатория Касперского, Управление по
интеллектуальной собственности, Надежда
Васильевна Кащенко

(72) Автор(ы):

Деденок Роман Андреевич (RU)

(73) Патентообладатель(и):

Акционерное общество "Лаборатория
Касперского" (RU)

(56) Список документов, цитированных в отчете
о поиске: RU 2583713 C2, 10.05.2016. RU
85247 U1, 27.07.2009. RU 2355018 C2,
10.05.2009. US 7636716 B1, 22.12.2009. US 2006/
0168035 A1, 27.07.2006.

(54) Система и способ определения сообщения, содержащего спам, по теме сообщения, отправленного по электронной почте

(57) Реферат:

Изобретение относится к области обнаружения спама. Техническим результатом является обнаружение спама в сообщении, отправленном по электронной почте. Раскрыт способ обнаружения спама в сообщении, отправленном по электронной почте, в котором: а) при помощи средства обработки сообщений получают сообщение, отправленное по электронной почте, содержащее в заголовке тему сообщения, в виде текста, который состоит более чем из трех слов; б) при помощи средства обработки сообщений определяют параметры текста темы сообщения, где параметрами текста темы сообщения является по крайней мере одно из: язык, на котором написан текст темы сообщения, количество слов в тексте темы сообщения, количество артиклей в тексте темы сообщения, количество пунктуационных знаков в тексте темы сообщения, количество местоимений в тексте темы сообщения, количество предлогов в тексте темы сообщения; в) при помощи средства определения коэффициентов определяют значения k и n коэффициентов для построения k-skip-n-gram

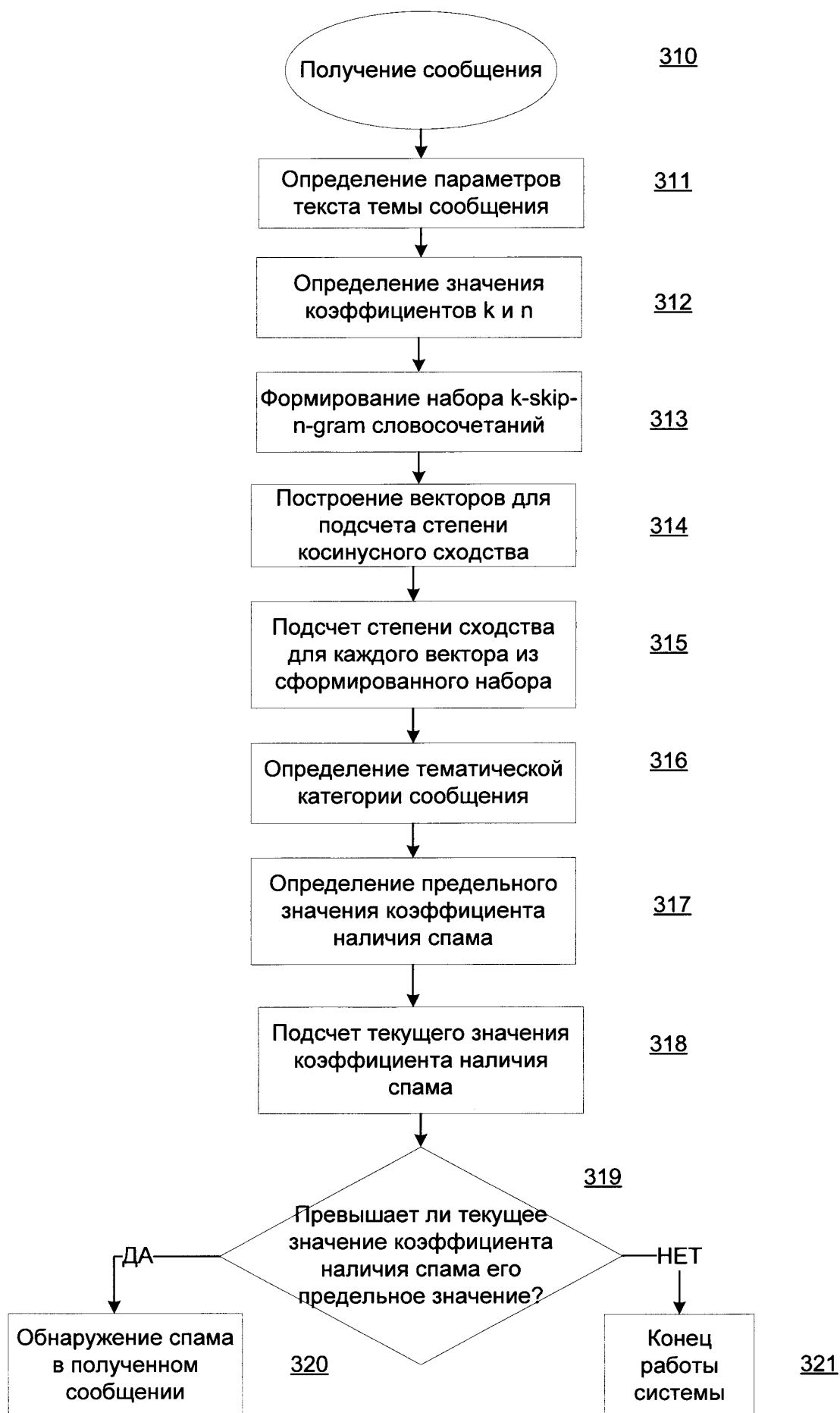
словосочетаний на основе параметров текста темы сообщения с помощью правил определения коэффициентов; г) при помощи средства определения коэффициентов формируют набор k-skip-n-gram словосочетаний от текста темы сообщения с использованием определенных значений k и n коэффициентов; д) при помощи средства построения векторов выполняют построение вектора для подсчета степени косинусного сходства для каждого k-skip-n-gram словосочетания из сформированного набора; е) при помощи средства построения векторов для каждого построенного вектора подсчитывают степень косинусного сходства с известными векторами из базы данных векторов; ж) при помощи средства обнаружения спама определяют тематическую категорию сообщения на основании множества подсчитанных степеней косинусного сходства с известными векторами; з) при помощи средства обнаружения спама подсчитывают текущее значение коэффициента наличия спама на основе множества посчитанных степеней косинусного сходства всех построенных векторов; и) при помощи средства обнаружения

спама при превышении определенного
предельного значения коэффициента наличия

спама обнаруживают спам в полученном
сообщении. 2 н.п. ф-лы, 5 ил.

R U 2 6 3 4 1 8 0 C 1

R U 2 6 3 4 1 8 0 C 1



Фиг. 3



FEDERAL SERVICE
FOR INTELLECTUAL PROPERTY

(12) **ABSTRACT OF INVENTION**(21)(22) Application: **2016125278, 24.06.2016**(24) Effective date for property rights:
24.06.2016Registration date:
24.10.2017

Priority:

(22) Date of filing: **24.06.2016**(45) Date of publication: **24.10.2017** Bull. № 30

Mail address:

**125212, Moskva, Leningradskoe sh., 39a, str. 3, AO
Laboratoriya Kasperskogo, Upravlenie po
intellektualnoj sobstvennosti, Nadezhda Vasilevna
Kashchenko**

(72) Inventor(s):

Dedenok Roman Andreevich (RU)

(73) Proprietor(s):

**Aktsionernoe obshchestvo "Laboratoriya
Kasperskogo" (RU)**

(54) **SYSTEM AND METHOD FOR DETERMINING SPAM-CONTAINING MESSAGE BY TOPIC OF MESSAGE SENT VIA E-MAIL**

(57) Abstract:

FIELD: information technology.

SUBSTANCE: method for detecting spam in a message sent via e-mail is disclosed, wherein: a) by means a message processing means the message sent via e-mail is received and where the message header contains a message in the form of a text which comprises more than three words; b) the message processing means determines the text parameters of the message subject, where the text parameters of the message topic are at least one of: a language on which the text of message subject is written, the number of words in the text of the message subject, the number of the articles in the text of the message subject, the number of punctuation characters in the text of the message subject, the number of pronouns in the text of the message subject, the number of prepositions in the text of the message subject; b) by means of a coefficient determining means, k and n coefficients are determined for constructing k-skip-and-gram of word combinations based on text parameters of the message subject by rules defining the coefficients; g) using the coefficient determining means, a set of k-skip-n-gram of the word

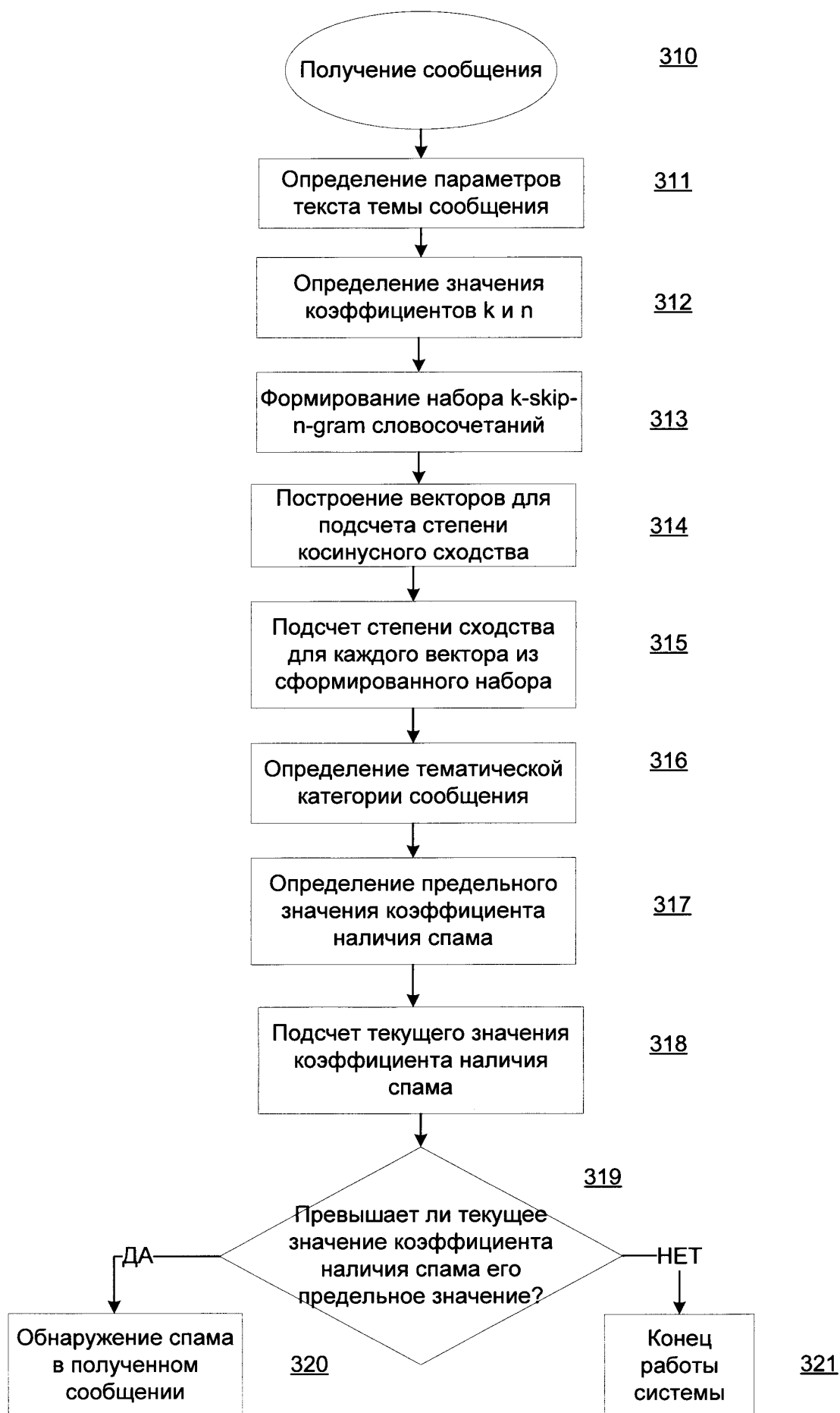
combinations from the text of the message subject using certain values of k and n coefficients; d) using a vector construction means, the vector is constructed to calculate the degree of cosine similarity for each k-skip-n-gram of the word combination from the generated set; e) using the vector construction means for each constructed vector, the degree of cosine similarity with known vectors from the vector database is calculated; g) using a spam detection means, a theme category of the message is determined based on a plurality of calculated degrees of cosine similarity with known vectors; h) by means of a spam detection means, the current value of the spam coefficient is calculated based on the plurality of counted degrees of cosine similarity of all constructed vectors; and i) by means of the spam detection means, when a certain threshold value of the spam coefficient is exceeded, the spam in the received message is detected.

EFFECT: spam detection in the message sent via e-mail.

2 cl, 5 dwg

RU 2 634 180 C1

RU 2 634 180 C1



Фиг. 3

Область техники

Изобретение относится к области защиты информации, в частности к системам и способам определения сообщения, содержащего спам.

Уровень техники

5 Реклама в Интернете является одним из самых дешевых видов рекламы. Спам-сообщения, как основной и наиболее массовый вид рекламы в современном мире, занимает от 70-90% от общего объема почтового трафика.

Спам - массовая рассылка рекламы или иного вида информации лицам, не выражавшим желания их получать. К спаму относятся сообщения, передаваемые по
10 электронной почте, протоколам мгновенных сообщений, в социальных сетях, блогах, сайтах знакомств, форумах, а также SMS- и MMS-сообщения.

Ввиду постоянного роста объемов рассылки спама возникают проблемы технического, экономического и криминального характера. Нагрузка на аппаратуру и каналы передачи данных, затраты времени пользователей на обработку сообщений,
15 изменение направленности сообщений в сторону мошенничества и воровства - эти и другие аспекты показывают острую необходимость непрерывной борьбы со спамом.

Существует много способов противодействия спам-рассылкам. Одним из самых эффективных является использование антиспам-приложений - программ, предназначенных для обнаружения и удаления нежелательных спам-сообщений.
20 Антиспам-приложения используют методы, с помощью которых происходит фильтрация и удаление спама. Методы основаны на анализе словосочетаний и контрольных сумм от словосочетаний текста сообщения.

Например, патенте US 755523 B1 описана система, в которой анализируют последовательности букв с использованием различающихся по длине n-gram. Вывод о
25 том, содержит ли сообщение спам, делают на основе поиска похожих последовательностей из базы данных спам-содержащих последовательностей.

Указанные решения осуществляют анализ текста сообщения (body). Настоящее изобретение позволяет эффективно решить задачу обнаружения спама в сообщениях, отправленных по электронной почте, на основании текста полей заголовка (header).

30 Раскрытие изобретения

Изобретение относится к системам и способам определения письма, содержащего спам, по теме сообщения, отправленного по электронной почте. Технический результат настоящего изобретения заключается в обеспечении защиты пользователя от получения спама в сообщениях, отправленных по электронной почте, которые имеют в заголовке
35 тему сообщения, в виде текста, который состоит более чем из трех слов. Указанный технический результат достигается за счет обнаружения спама в сообщении, отправленном по электронной почте при подсчете коэффициента наличия спама на основе k-skip-n-gram словосочетаний, выстроенных от текста темы заголовка сообщения, отправленного по электронной почте. Превышения предельного значения коэффициента
40 наличия спама является признаком того, что сообщение, отправленное по электронной почте, является спамом.

В одном из вариантов реализации предоставляется система обнаружения спама в сообщении, отправленном по электронной почте, которая содержит: средство обработки сообщений, предназначенное для: получения сообщения, отправленного по электронной
45 почте, содержащего в заголовке тему сообщения, в виде текста, который состоит более чем из трех слов, определения параметров текста темы сообщения, где параметрами текста темы сообщения являются: язык, на котором написан текст темы сообщения, количество слов в тексте темы сообщения, количество артиклей в тексте темы

сообщения, количество пунктуационных знаков в тексте темы сообщения, количество местоимений в тексте темы сообщения, количество предлогов в тексте темы сообщения; передачи текста и параметров текста темы сообщения средству определения коэффициентов; средство определения коэффициентов, предназначенное для:

5 определения значения k и n коэффициентов для построения k -skip- n -gram словосочетаний на основе параметров текста темы сообщения с помощью правил определения коэффициентов, формирования набора k -skip- n -gram словосочетаний от текста темы сообщения с использованием определенных значений k и n коэффициентов, передачи сформированного набора k -skip- n -gram словосочетаний средству построения векторов;

10 базу данных правил, предназначенную для хранения правил определения коэффициентов; средство построения векторов, предназначенное для: построения вектора для подсчета степени косинусного сходства для каждого k -skip- n -gram словосочетания из сформированного набора; для каждого построенного вектора подсчета степени косинусного сходства с известными векторами из базы данных

15 векторов; определения тематической категории сообщения на основании подсчитанных степеней косинусного сходства с известными векторами из базы данных векторов; передачи данных о подсчитанных степенях косинусного сходства и тематической категории сообщения средству обнаружения спама; базу данных векторов, предназначенную для хранения известных векторов для подсчета степени косинусного

20 сходства k -skip- n -gram словосочетаний; средство обнаружения спама, предназначенное для: определения предельного значения коэффициента наличия спама на основании тематической категории сообщения, подсчета текущего значения коэффициента наличия спама на основе степеней косинусного сходства для всех векторов, при превышении определенного предельного значения коэффициента наличия спама обнаружение спама

25 в полученном сообщении.

В другом варианте реализации предоставляется способ обнаружения спама в сообщении, отправленном по электронной почте, в котором: при помощи средства обработки сообщений получают сообщение, отправленное по электронной почте, содержащее в заголовке тему сообщения, в виде текста, который состоит более чем из

30 трех слов; при помощи средства обработки сообщений определяют параметры текста темы сообщения, где параметрами текста темы сообщения является по крайней мере одно из: язык, на котором написан текст темы сообщения, количество слов в тексте темы сообщения, количество артиклей в тексте темы сообщения, количество пунктуационных знаков в тексте темы сообщения, количество местоимений в тексте

35 темы сообщения, количество предлогов в тексте темы сообщения; при помощи средства определения коэффициентов определяют значения k и n коэффициентов для построения k -skip- n -gram словосочетаний на основе параметров текста темы сообщения с помощью правил определения коэффициентов; при помощи средства определения коэффициентов формируют набор k -skip- n -gram словосочетаний от текста темы сообщения с

40 использованием определенных значений k и n коэффициентов; при помощи средства построения векторов выполняют построение вектора для подсчета степени косинусного сходства для каждого k -skip- n -gram словосочетания из сформированного набора; при помощи средства построения векторов для каждого построенного вектора подсчитывают степень косинусного сходства с известными векторами из базы данных

45 векторов; при помощи средства обнаружения спама определяют тематическую категорию сообщения на основании подсчитанных степеней косинусного сходства с известными векторами; при помощи средства обнаружения спама подсчитывают текущее значение коэффициента наличия спама на основе степеней косинусного сходства

всех построенных векторов; при помощи средства обнаружения спама при превышении определенного предельного значения коэффициента наличия спама обнаруживают спам в полученном сообщении.

Краткое описание чертежей

5 Дополнительные цели, признаки и преимущества настоящего изобретения будут очевидными из прочтения последующего описания осуществления изобретения со ссылкой на прилагаемые чертежи, на которых:

Фиг. 1А отображает поля заголовка и пример темы произвольного сообщения, отправленного по электронной почте.

10 Фиг. 1Б изображает формулу вычисления и векторное отображение косинусного сходства.

Фиг. 2 отображает структурную схему системы обнаружения спама в сообщении, отправленном по электронной почте.

15 Фиг. 3 иллюстрирует алгоритм работы системы обнаружения спама в сообщении, отправленном по электронной почте.

Фиг. 4 представляет пример компьютерной системы общего назначения.

20 Хотя изобретение может иметь различные модификации и альтернативные формы, характерные признаки, показанные в качестве примера на чертежах, будут описаны подробно. Следует понимать, однако, что цель описания заключается не в ограничении изобретения конкретным его воплощением. Наоборот, целью описания является охват всех изменений, модификаций, входящих в рамки данного изобретения, как это определено в приложенной формуле.

Описание вариантов осуществления изобретения

25 Объекты и признаки настоящего изобретения, способы для достижения этих объектов и признаков станут очевидными посредством отсылки к примерным вариантам осуществления. Однако настоящее изобретение не ограничивается примерными вариантами осуществления, раскрытыми ниже, оно может воплощаться в различных видах. Сущность, приведенная в описании, является ничем иным, как конкретными деталями, необходимыми для помощи специалисту в области техники в исчерпывающем понимании изобретения, и настоящее изобретение определяется в объеме приложенной формулы.

30 Согласно RFC 5322, сообщение состоит из полей заголовков (совокупность этих полей называют разделом заголовков сообщения), за которыми может следовать тело сообщения. Раздел заголовков представляет собой последовательность символьных строк, синтаксис которых описан в данной спецификации. Тело сообщения представляет собой последовательность символов, которая следует после раздела заголовков и отделена от него пустой строкой (строкой, содержащей только CRLF).

40 Упомянутые в уровне техники публикации выполняют обнаружение спама путем анализа тела сообщения (body). Помимо тела сообщения, предметом анализа может быть заголовок. Особый интерес, в частности, может представлять поле заголовка - тема сообщения (subject). Тема сообщения - короткое предложение, которое описывает цель написания и содержание сообщения. Фиг. 1А отображает поля заголовка и пример темы произвольного сообщения, отправленного по электронной почте.

45 Для того чтобы на основе данных о теме сообщения определить сообщение, содержащее спам, используют систему обнаружения спама в сообщении, отправленном по электронной почте. Система обнаружения спама в сообщении, отправленном по электронной почте, содержит средство обработки сообщений 220, средство определения коэффициентов 230, средство построения векторов 240, средство обнаружения спама

250, базу данных правил 260, базу данных векторов 270.

Средство обработки сообщений 220 предназначено для получения сообщения, отправленного по электронной почте, содержащего в заголовке тему сообщения в виде текста, который состоит более чем из трех слов 210.

5 В одном случае может быть получено сообщение, тема которого состоит из бессмысленной нераздельной или неразборной последовательности символов. Подобные сообщения могут появляться, например, когда была допущена ошибка при выборе кодировки. Помимо этого, подобные сообщения зачастую являются признаком сообщения, содержащего спам. Анализ темы подобных сообщений затруднителен. В
10 другом случае может быть получено сообщение, тема которого состоит из нескольких слов. Текст темы сообщения, состоящий более чем из трех слов, наиболее предпочтителен для анализа.

Помимо этого, средство обработки сообщений 220 предназначено для определения параметров текста темы сообщения.

15 Параметры текста темы сообщения - величина, характеризующая основные существенные особенности текста темы сообщения. Параметрами текста темы сообщения может быть, например:

- язык, на котором написан текст темы сообщения,
- количество слов в тексте темы сообщения,
- 20 - количество артиклей слов в тексте темы сообщения,
- количество пунктуационных знаков слов в тексте темы сообщения,
- количество местоимений слов в тексте темы сообщения,
- количество предлогов слов в тексте темы сообщения и т.д.

Так же средство обработки сообщений 220 предназначено для передачи текста и
25 параметров текста темы сообщения средству определения коэффициентов 230.

Средство определения коэффициентов 230 предназначено для определения значения k, n коэффициентов для построения k-skip-n-gram словосочетаний на основе параметров текста темы сообщения с использованием правил определения коэффициентов из базы данных правил 260.

30 N-gram - это последовательность словосочетаний, состоящих из n-слов. Коэффициент n - количество слов, которое будет содержать одно словосочетание из набора. Например, в случае, если n=2, набор 2-gram (bi-gram), построенный от предложения: «Привет! Не забудь купить билеты по акции!» - будет выглядеть следующим образом: «привет не; не забудь; забудь купить; купить билеты; билеты по; билеты акции; по акции.»

35 K-skip-n-gram - это последовательность словосочетаний, состоящая из n-слов, между которыми опускают до k-слов. Таким образом, коэффициент k показывает, что в словосочетание, помимо соседних слов, входят слова через от одного до k слов от начального слова.

$$40 \quad \{w_{i_1}, w_{i_2}, \dots, w_{i_n} \mid \sum_{j=1}^n i_j - i_{j-1} < k\}$$

Например, в случае, если n=2, k=2, набор 2-skip-bi-gram, построенный от упомянутого предложения, будет выглядеть следующим образом: «привет не; привет забудь; привет
45 купить; не забудь; не купить; не билеты; забудь купить; забудь билеты; забудь по; купить билеты; купить по; купить акции; билеты по; билеты акции; по акции.»

Правило определения коэффициентов - набор условий, при выполнении которых средство определения коэффициентов осуществляет выбор или подсчет наиболее подходящего значения коэффициента. Условие может быть основано на параметрах текста темы сообщения. В большинстве европейских языков на коэффициент k может

влиять количество артиклей и предлогов: чем их больше в языке, тем меньше должно быть значение k . Но и количество артиклей в тексте сообщения определенного языка так же влияет на коэффициент k . На коэффициент n имеют особое влияние такие параметры, например как количество местоимений или количество предлогов в тексте

Например, правилом определения коэффициента k может быть следующий набор условий: количество слов в тексте темы сообщения - от 3 до 10, язык текста темы сообщения - английский, количество пунктуационных знаков - 3, в итоге коэффициент k равен 2. Другим примером правила определения коэффициента k может быть следующий набор условий: количество слов в тексте темы сообщения - больше 10, язык темы сообщения - русский, количество пунктуационных знаков - 0, коэффициент k равен 3.

Например, правилом определения коэффициента n может быть следующий набор условий: коэффициент k - равен 2, язык текста темы сообщения - испанский, количество местоимений в тексте темы сообщения - 3, в итоге коэффициент n равен 2 и т.д.

Помимо этого, средство определения коэффициентов 230 предназначено для формирования набора k -skip- n -gram словосочетаний от текста темы сообщения с использованием определенных ранее значений k , n коэффициентов. Так же средство определения коэффициентов 230 предназначено для передачи сформированного набора k -skip- n -gram словосочетаний средству построения векторов 240.

База данных правил 260 предназначена для хранения правил определения коэффициента k , правила определения коэффициента n .

Средство построения векторов 240 предназначено для построения вектора для подсчета степени косинусного сходства для каждого k -skip- n -gram словосочетания из сформированного набора.

Фиг. 1Б изображает формулу вычисления и векторное отображение косинусного сходства. Косинусное сходство - это мера сходства между двумя векторами предгильбертового пространства, которая используется для измерения косинуса угла между ними. Для каждого k -skip- n -gram словосочетания из сформированного набора выстраивают вектор для подсчета степени косинусного сходства. Например, для 2-skip-
bi-gram словосочетания от упомянутого текста «билеты акции» вектор для подсчета степени сходства будет следующим: -2.622624 1.091368 1.221946 1.118406 0.286586 -0.477737 0.925635 -0.179525 0.212215 -1.804560 1.452010 -1.630697 -0.030821 0.525848 -1.306217 -0.847145 -0.209074 -2.020271.

Помимо этого, средство построения векторов 240 предназначено для подсчета степени косинусного сходства каждого построенного вектора с известными векторами из базы данных векторов 270.

Для упомянутого вектора рассчитывают косинусное сходство с известными векторами из базы данных векторов 270 для выявления наиболее похожих словосочетаний. Степень косинусного сходства вектора словосочетания «билеты акции» с известным из базы данных векторов 270 вектором от словосочетания «билеты скидки» после подсчета имеет значение 0,75, а со словосочетанием «билеты концерт» может быть равно 0.79.

Помимо этого, средство построения векторов 240 предназначено для определения тематической категории сообщения на основе посчитанной степени сходства по крайней мере одного построенного вектора и известных векторов из базы данных векторов 270.

Тематическая категория сообщения - категория, содержащая сообщения, отдельные словосочетания или векторы, схожие по содержанию или цели написания, в которых

использованы аналогичные по смыслу слова, сочетания слов или сочетания слов и символов, характерные для конкретной тематики. Множество спам писем заранее подвергаются анализу, например при помощи NMF (Non-negative matrix factorization) или с использованием латентного размещения Дирихле, разбиваются на тематические категории. Примером может быть категория спам писем, «финансовый заработок». В эту категорию входят сообщения, которые содержат сведения о дополнительном заработке, сумме заработка в месяц, возможности заработка помимо работы и т.д. Тематическую категорию может иметь как сообщение, так и отдельное словосочетание или построенный от него вектор.

Средство построения векторов 240 определяет тематическую категорию сообщения по вектору или совокупности векторов, например, имеющих наивысшую степень схожести с известными векторами из базы данных векторов 270 одной тематики. Соответственно, если, например, известный вектор из базы данных векторов 270 относится к определенной тематической категории, например, «интимные услуги», и один из построенных от текста темы заголовка векторов имеет высокую степень косинусного сходства с ним, то именно эту категорию и будет иметь полученное сообщение.

Дополнительно, средство построения векторов 240 предназначено для передачи данных о посчитанных степенях косинусного сходства и определенной тематической категории сообщения средству обнаружения спама 250.

База данных векторов 270 предназначена для хранения известных векторов для подсчета степени косинусного сходства k-skip-n-gram словосочетаний. Известные векторы в базе данных векторов 270 распределены по тематическим категориям.

В качестве базы данных правил 260 и базы данных векторов 270 могут использоваться различные виды баз данных, а именно: иерархические (IMS, TDMS, System 2000), сетевые (Cerebrum, Cronospro, DBVist), реляционные (DB2, Informix, Microsoft SQL Server), объектно-ориентированные (Jasmine, Versant, POET), объектно-реляционные (Oracle Database, PostgreSQL, FirstSQL/J, функциональные и т.д. Обновление баз данных также может быть осуществлено при помощи антивирусного сервера.

Текущее значение коэффициента наличия спама - количественный показатель наличия спама в полученном сообщении, определяемый на основе посчитанных степеней косинусного сходства всех векторов из набора, сформированного от текста темы сообщения. Текущее значение коэффициента наличия спама может быть посчитано с использованием, например, среднего арифметического от степеней косинусного сходства всех построенных векторов.

Предельное значение коэффициента наличия спама - значение коэффициента наличия спама, при котором полученное сообщение считается сообщением, содержащим спам. Предельное значение коэффициента наличия спама зависит от тематической категории сообщения. Для каждой тематической категории сообщения определяют предельное значение коэффициента наличия спама опытным путем или путем регрессивного анализа тестовых коллекций сообщений и заносят в базу данных векторов 270. Далее упомянутое значение корректируют в зависимости от количества k-skip-n-gram в сформированном наборе, количестве известных векторов из базы данных векторов, с которыми по крайней мере один посчитанный вектор имеет высокую степень схожести, и т.д.

В базе данных векторов 270 для каждой тематической категории сообщений хранят предельное значение коэффициента наличия спама.

Средство обнаружения спама 250 предназначено для определения предельного значения коэффициента наличия спама на основе тематической категории сообщения.

В одном случае, например, сообщения принадлежит к тематической категории «интимные услуги». Тематическая категория сообщения определена на основе, например, 2-skip-2-gram «интимные цена», от которого построен вектор для подсчета степени схожести. 2-skip-2-gram «интимные цена» выбран в качестве определяющего, поскольку построенный вектор от этого 2-skip-2-gram имеет максимальную степень схожести с 3-мя известными векторами из базы данных векторов 270. С первым вектором, от k-skip-n-gram «интимные рублей» упомянутый вектор имеет степень схожести 0,78 и является максимальной среди степеней схожести для всех построенных векторов из сформированного набора. Со вторым вектором от k-skip-n-gram «интимная бесплатно» упомянутый вектор имеет степень схожести 0,69, также является высокой. С третьим вектором от k-skip-n-gram от «интимной обстановке» упомянутый вектор имеет степень схожести 0.53. В сформированном наборе 24 2-skip-2-gram, что означает, что в тексте темы сообщения 10 слов. Таким образом, предельный коэффициент наличия спама для тематической категории «интимные услуги», известный из базы данных векторов 270, равен 0,80, при определении средством обнаружения спама 250 с учетом уточненных данных (3 вектора, 10 слов) будет равен 0,77. Уточнение может быть посчитано при помощи арифметических формул.

Так же средство обнаружения спама 250 предназначено для подсчета текущего значения коэффициента наличия спама на основе степеней косинусного сходства всех построенных векторов.

Помимо этого, средство обнаружения спама 250 предназначено для обнаружения спама в полученном сообщении при превышении определенного предельного значения коэффициента наличия спама.

В случае, если предельное значение коэффициента наличия спама не превышено, в полученном сообщении спам не может быть обнаружен и система прекращает работу

Фиг. 3 представляет схему алгоритма работы системы обнаружения спама в сообщении, отправленном по электронной почте. На этапе 310 средство обработки сообщений 220 получает сообщение, отправленное по электронной почте, содержащее в заголовке тему сообщения, в виде текста, который состоит более чем из трех слов 210. На этапе 311 средство обработки сообщений 220 определяет параметры текста темы сообщения и передает текст и параметры текста темы сообщения средству определения коэффициентов 230. На этапе 312 средство определения коэффициентов 230 определяет значения k и n коэффициентов для построения k-skip-n-gram словосочетаний на основе параметров текста темы сообщения с помощью правил определения коэффициентов. На этапе 313 средство определения коэффициентов 230 формирует набор k-skip-n-gram словосочетаний от текста темы сообщения с использованием определенных значений k и n коэффициентов и передает сформированный набор k-skip-n-gram словосочетаний средству построения векторов. На этапе 314 средство построения векторов 240 выполняет построение вектора для подсчета степени косинусного сходства для каждого k-skip-n-gram словосочетания из сформированного набора. На этапе для каждого построенного вектора 315 средство построения векторов 240 подсчитывает степень косинусного сходства с известными векторами из базы данных векторов 270. На этапе 316 средство построения векторов 240 определяет тематическую категорию сообщения на основании подсчитанных степеней косинусного сходства с известными векторами и передает данные о подсчитанных степенях косинусного сходства и тематической категории сообщения средству обнаружения спама 250. На этапе 317 средство обнаружения спама 250 определяет предельное значение коэффициента наличия спама на основании

тематической категории сообщения. На этапе 318 средство обнаружения спама 250 выполняет подсчет значения коэффициента наличия спама на основе степеней косинусного сходства всех построенных векторов. На этапе 319 средство обнаружения спама 250 проверяет, превышает ли посчитанное значение коэффициента наличия спама его определенное предельное значение. При превышении определенного предельного значения коэффициента наличия спама на этапе 320 средство обнаружения спама выполняет обнаружение спама в полученном сообщении. Если определенное предельное значение коэффициента наличия спама не превышено, на этапе 321 система заканчивает работу.

Фиг. 4 представляет пример компьютерной системы общего назначения, персональный компьютер или сервер 20, содержащий центральный процессор 21, системную память 22 и системную шину 23, которая содержит разные системные компоненты, в том числе память, связанную с центральным процессором 21. Системная шина 23 реализована, как любая известная из уровня техники шинная структура, содержащая в свою очередь память шины или контроллер памяти шины, периферийную шину и локальную шину, которая способна взаимодействовать с любой другой шинной архитектурой. Системная память содержит постоянное запоминающее устройство (ПЗУ) 24, память с произвольным доступом (ОЗУ) 25. Основная система ввода/вывода (BIOS) 26 содержит основные процедуры, которые обеспечивают передачу информации между элементами персонального компьютера 20, например, в момент загрузки операционной системы с использованием ПЗУ 24.

Персональный компьютер 20 в свою очередь содержит жесткий диск 27 для чтения и записи данных, привод магнитных дисков 28 для чтения и записи на сменные магнитные диски 29 и оптический привод 30 для чтения и записи на сменные оптические диски 31, такие как CD-ROM, DVD-ROM и иные оптические носители информации. Жесткий диск 27, привод магнитных дисков 28, оптический привод 30 соединены с системной шиной 23 через интерфейс жесткого диска 32, интерфейс магнитных дисков 33 и интерфейс оптического привода 34 соответственно. Приводы и соответствующие компьютерные носители информации представляют собой энергонезависимые средства хранения компьютерных инструкций, структур данных, программных модулей и прочих данных персонального компьютера 20.

Настоящее описание раскрывает реализацию системы, которая использует жесткий диск 27, сменный магнитный диск 29 и сменный оптический диск 31, но следует понимать, что возможно применение иных типов компьютерных носителей информации 56, которые способны хранить данные в доступной для чтения компьютером форме (твердотельные накопители, флеш карты памяти, цифровые диски, память с произвольным доступом (ОЗУ) и т.п.), которые подключены к системной шине 23 через контроллер 55.

Компьютер 20 имеет файловую систему 36, где хранится записанная операционная система 35, а также дополнительные программные приложения 37, другие программные модули 38 и данные программ 39. Пользователь имеет возможность вводить команды и информацию в персональный компьютер 20 посредством устройств ввода (клавиатуры 40, манипулятора «мышь» 42). Могут использоваться другие устройства ввода (не отображены): микрофон, джойстик, игровая консоль, сканер и т.п. Подобные устройства ввода по своему обычаю подключают к компьютерной системе 20 через последовательный порт 46, который в свою очередь подсоединен к системной шине, но могут быть подключены иным способом, например, при помощи параллельного порта, игрового порта или универсальной последовательной шины (USB). Монитор

47 или иной тип устройства отображения также подсоединен к системной шине 23 через интерфейс, такой как видеоадаптер 48. В дополнение к монитору 47, персональный компьютер может быть оснащен другими периферийными устройствами вывода (не отображены), например, колонками, принтером и т.п.

5 Персональный компьютер 20 способен работать в сетевом окружении, при этом используется сетевое соединение с другим или несколькими удаленными компьютерами 49. Удаленный компьютер (или компьютеры) 49 являются такими же персональными компьютерами или серверами, которые имеют большинство или все упомянутые элементы, отмеченные ранее при описании существа персонального компьютера 20, 10 представленного на Фиг. 4. В вычислительной сети могут присутствовать также и другие устройства, например, маршрутизаторы, сетевые станции, пиринговые устройства или иные сетевые узлы.

Сетевые соединения могут образовывать локальную вычислительную сеть (LAN) 50 и глобальную вычислительную сеть (WAN). Такие сети применяются в корпоративных 15 компьютерных сетях, внутренних сетях компаний и, как правило, имеют доступ к сети Интернет. В LAN- или WAN-сетях персональный компьютер 20 подключен к локальной сети 50 через сетевой адаптер или сетевой интерфейс 51. При использовании сетей персональный компьютер 20 может использовать модем 54 или иные средства обеспечения связи с глобальной вычислительной сетью, такой как Интернет. Модем 20 54, который является внутренним или внешним устройством, подключен к системной шине 23 посредством последовательного порта 46. Следует уточнить, что сетевые соединения являются лишь примерными и не обязаны отображать точную конфигурацию сети, т.е. в действительности существуют иные способы установления соединения техническими средствами связи одного компьютера с другим.

25

(57) Формула изобретения

1. Система обнаружения спама в сообщении, отправленном по электронной почте, которая содержит:

а) средство обработки сообщений, предназначенное для: 30 получения сообщения, отправленного по электронной почте, содержащего в заголовке тему сообщения, в виде текста, который состоит более чем из трех слов, определения параметров текста темы сообщения, где параметрами текста темы сообщения являются:

язык, на котором написан текст темы сообщения, 35 количество слов в тексте темы сообщения, количество артиклей в тексте темы сообщения, количество пунктуационных знаков в тексте темы сообщения, количество местоимений в тексте темы сообщения, количество предлогов в тексте темы сообщения;

40 передачи текста и параметров текста темы сообщения средству определения коэффициентов;

б) средство определения коэффициентов, предназначенное для: определения значения k и n коэффициентов для построения k-skip-n-gram словосочетаний на основе параметров текста темы сообщения с помощью правил 45 определения коэффициентов,

формирования набора k-skip-n-gram словосочетаний от текста темы сообщения с использованием определенных значений k и n коэффициентов, передачи сформированного набора k-skip-n-gram словосочетаний средству построения

векторов;

в) базу данных правил, предназначенную для хранения правил определения коэффициентов;

г) средство построения векторов, предназначенное для:

5 построения вектора для подсчета степени косинусного сходства для каждого k-skip-n-gram словосочетания из сформированного набора;

для каждого построенного вектора подсчета степени косинусного сходства с известными векторами из базы данных векторов;

определения тематической категории сообщения на основании множества

10 подсчитанных степеней косинусного сходства с известными векторами из базы данных векторов;

передачи данных о подсчитанных степенях косинусного сходства и тематической категории сообщения средству обнаружения спама;

15 д) базу данных векторов, предназначенную для хранения известных векторов для подсчета степени косинусного сходства k-skip-n-gram словосочетаний;

е) средство обнаружения спама, предназначенное для:

определения предельного значения коэффициента наличия спама на основании тематической категории сообщения,

подсчета текущего значения коэффициента наличия спама на основе множества

20 посчитанных степеней косинусного сходства для всех векторов,

при превышении определенного предельного значения коэффициента наличия спама обнаружение спама в полученном сообщении.

2. Способ обнаружения спама в сообщении, отправленном по электронной почте, в котором:

25 а) при помощи средства обработки сообщений получают сообщение, отправленное по электронной почте, содержащее в заголовке тему сообщения, в виде текста, который состоит более чем из трех слов;

б) при помощи средства обработки сообщений определяют параметры текста темы сообщения, где параметрами текста темы сообщения является по крайней мере одно

30 из:

язык, на котором написан текст темы сообщения,

количество слов в тексте темы сообщения,

количество артиклей в тексте темы сообщения,

количество пунктуационных знаков в тексте темы сообщения,

35 количество местоимений в тексте темы сообщения,

количество предлогов в тексте темы сообщения;

в) при помощи средства определения коэффициентов определяют значения k и n коэффициентов для построения k-skip-n-gram словосочетаний на основе параметров текста темы сообщения с помощью правил определения коэффициентов;

40 г) при помощи средства определения коэффициентов формируют набор k-skip-n-gram словосочетаний от текста темы сообщения с использованием определенных значений k и n коэффициентов;

д) при помощи средства построения векторов выполняют построение вектора для подсчета степени косинусного сходства для каждого k-skip-n-gram словосочетания из сформированного набора;

45 е) при помощи средства построения векторов для каждого построенного вектора подсчитывают степень косинусного сходства с известными векторами из базы данных векторов;

ж) при помощи средства обнаружения спама определяют тематическую категорию сообщения на основании множества подсчитанных степеней косинусного сходства с известными векторами;

5 з) при помощи средства обнаружения спама подсчитывают текущее значение коэффициента наличия спама на основе множества посчитанных степеней косинусного сходства всех построенных векторов;

и) при помощи средства обнаружения спама при превышении определенного предельного значения коэффициента наличия спама обнаруживают спам в полученном сообщении.

10

15

20

25

30

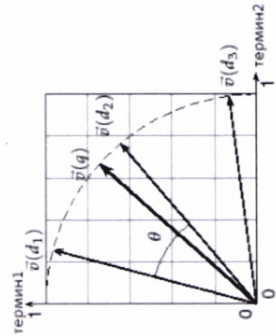
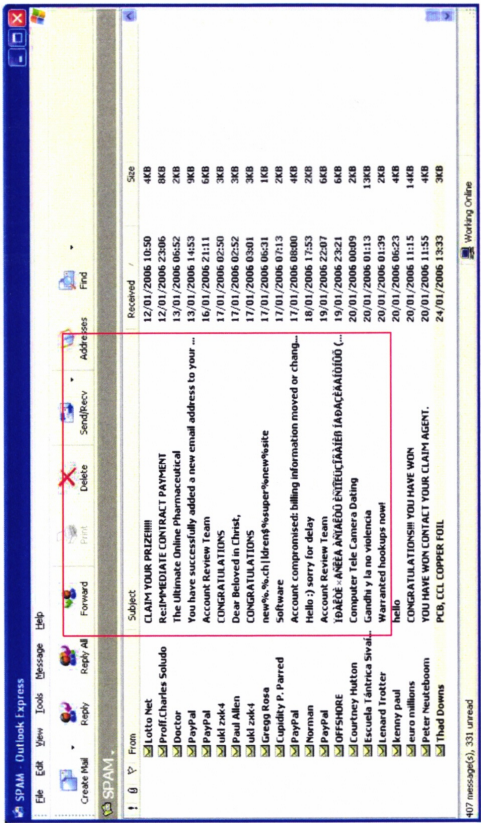
35

40

45

Field	Min number	Max number	Notes
trace	0	unlimited	Block prepended - see 3.6.7
resent-date	0*	unlimited*	One per block, required if other resent fields are present - see 3.6.6
resent-from	0	unlimited*	One per block - see 3.6.6
resent-sender	0*	unlimited*	One per block, MUST occur with multi-address
resent-to	0	unlimited*	Resent-from - see 3.6.6
resent-cc	0	unlimited*	One per block - see 3.6.6
resent-bcc	0	unlimited*	One per block - see 3.6.6
resent-msg-id	0	unlimited*	One per block - see 3.6.6
orig-date	1	1	
from	1	1	See sender and 3.6.2
sender	0*	1	MUST occur with multi-address from - see 3.6.2
reply-to	0	1	
to	0	1	
cc	0	1	
bcc	0	1	
message-id	0*	1	SHOULD be present - see 3.6.4
in-reply-to	0*	1	SHOULD occur in some replies - see 3.6.4
references	0*	1	SHOULD occur in some replies - see 3.6.4
subject	0	1	
comments	0	unlimited	
keywords	0	unlimited	
optional-field	0	unlimited	

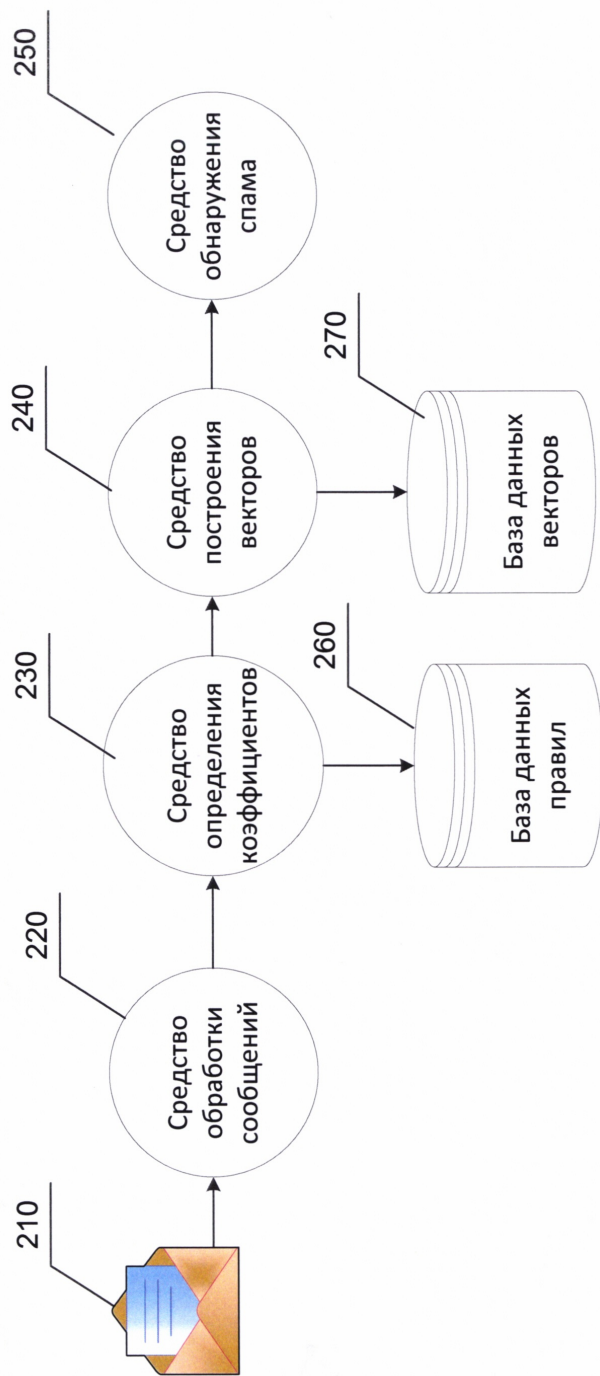
Фиг. 1А



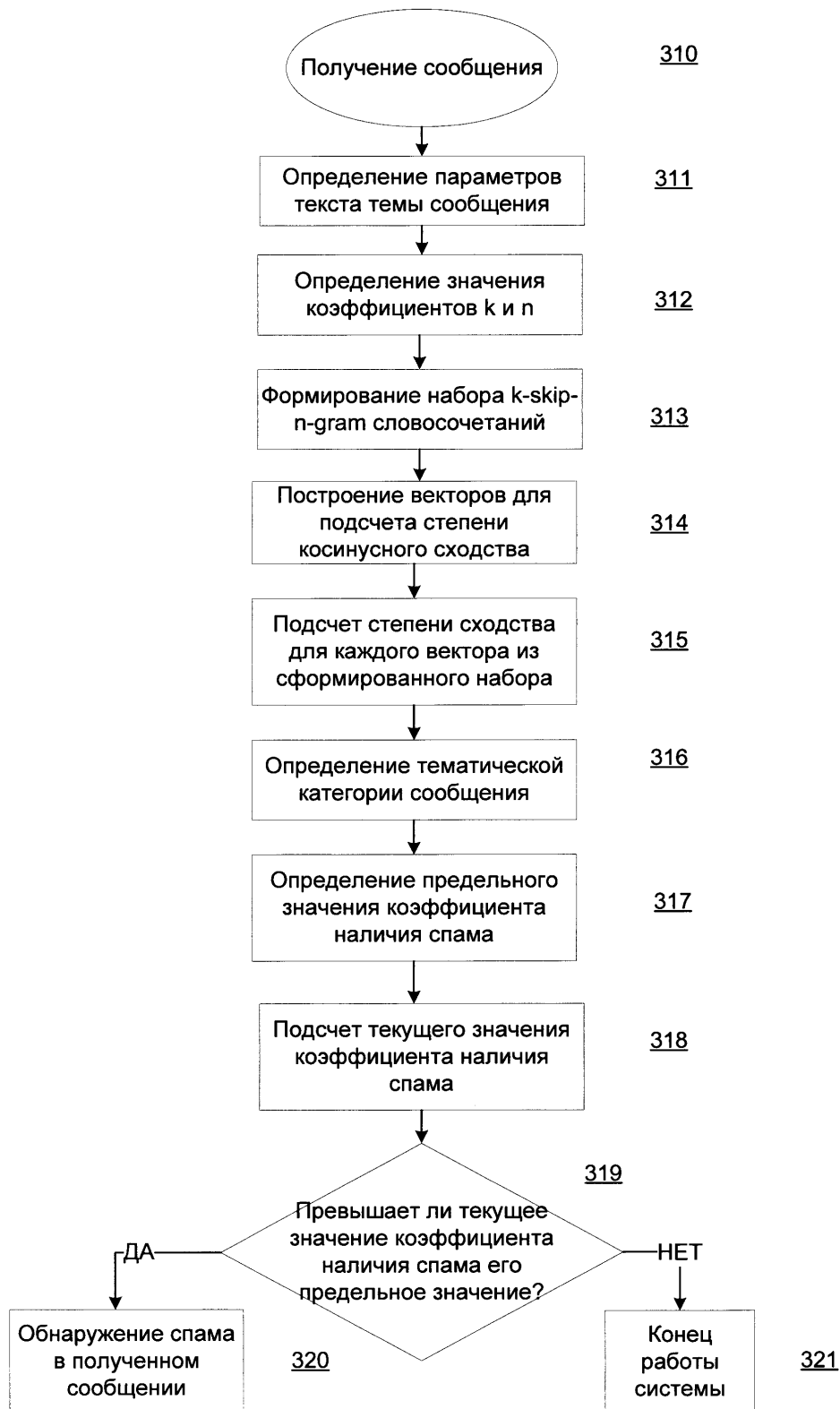
Фиг. 1Б

$sim(d,q) = \cos \theta$

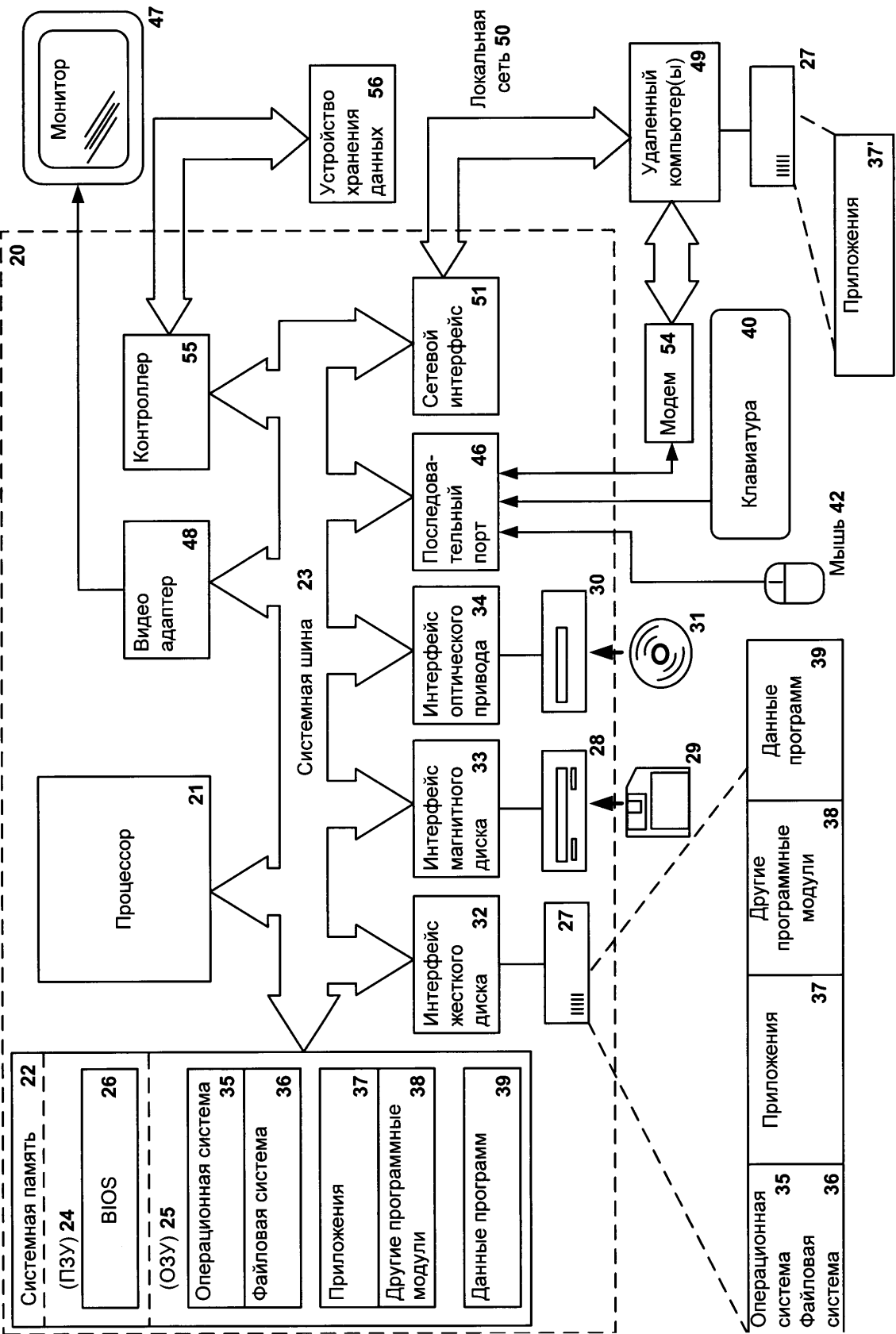
$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$



Фиг. 2



Фиг. 3



Фиг. 4