РОССИЙСКАЯ ФЕДЕРАЦИЯ

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

(19) **RU** (11) **2 711 125**(13) **C2**

(51) МПК
G06K 9/62 (2006.01)
G06N 20/00 (2019.01)
G06N 3/02 (2006.01)

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ПАТЕНТУ

(52) СПК
G06K 9/62 (2019.02); G06N 20/00 (2019.02); G06N 3/02 (2019.02); G06F 17/40 (2019.02); G06F 15/00 (2019.02); G06T 1/20 (2019.02)

(21)(22) Заявка: **2017142709, 07.12.2017**

(24) Дата начала отсчета срока действия патента:
07.12.2017

Дата регистрации:
15.01.2020

Приоритет(ы):
(22) Дата подачи заявки: **07.12.2017**

(43) Дата публикации заявки: **10.06.2019** Бюл. № **16**

(45) Опубликовано: **15.01.2020** Бюл. № **2**

Адрес для переписки:
119021, Москва, ул. Льва Толстого, 16, ООО "Яндекс", отдел правовой охраны технологий

(72) Автор(ы):
Лахман Константин Викторович (RU),
Чигорин Александр Александрович (RU),
Юрченко Виктор Сергеевич (RU)

(73) Патентообладатель(и):
ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "ЯНДЕКС" (RU)

(56) Список документов, цитированных в отчете о поиске: **US 2016/0140438 A1, 19.05.2016. US 2011/0258149 A1, 20.10.2011. US 2016/0035078 A1, 04.02.2016. US 20160125274 A1, 05.05.2016. US 2016/0034788 A1, 04.02.2016. RU 2635259 C1, 09.11.2017.**

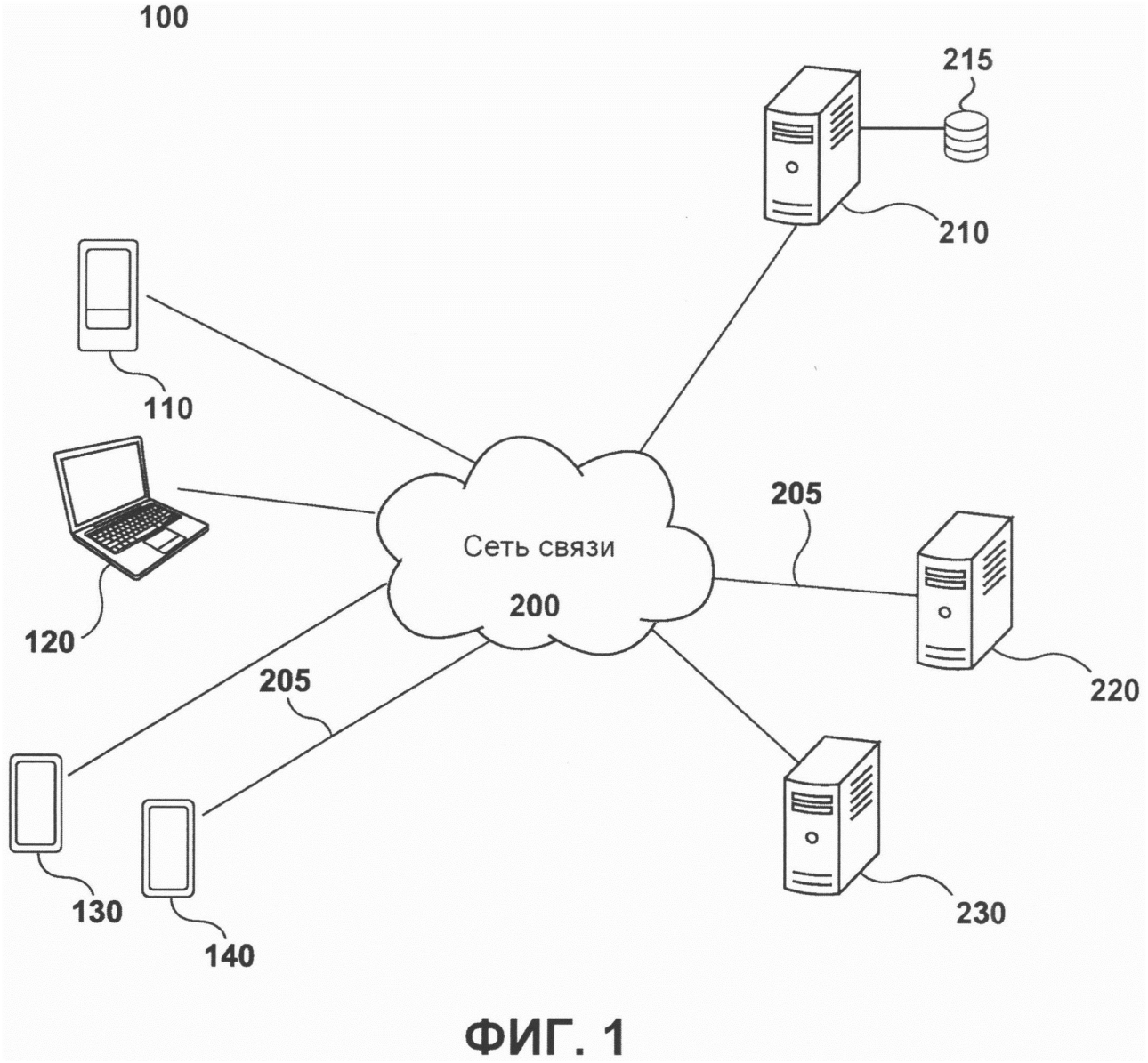
(54) СИСТЕМА И СПОСОБ ФОРМИРОВАНИЯ ОБУЧАЮЩЕГО НАБОРА ДЛЯ АЛГОРИТМА МАШИННОГО ОБУЧЕНИЯ

(57) Реферат:

Изобретение относится к вычислительной технике. Технический результат - расширение арсенала технических средств для формирования набора обучающих объектов для алгоритма машинного обучения и обучения алгоритма машинного обучения с использованием сформированного набора. Способ и система формирования набора обучающих объектов для алгоритма машинного обучения (MLA) включают в себя: получение данных поисковых запросов, каждый из которых связывается с первым набором результатов поиска изображений; формирование вектора запроса для каждого из поисковых запросов; распределение векторов запросов между множеством кластеров векторов запросов; связывание с каждым кластером векторов запросов второго набора результатов поиска изображений, содержащего по меньшей мере часть каждого из первых наборов результатов поиска изображений, связанных с векторами запросов, входящими в состав каждого соответствующего кластера векторов запросов; сохранение для каждого кластера векторов запросов каждого результата поиска изображений из второго набора результатов поиска изображений в виде обучающего объекта в наборе обучающих объектов, при этом каждый результат поиска изображений связан с меткой кластера. 3 н. и 17 з.п. ф-лы, 5 ил.

RU 2711125 C2

RU 2711125 C2



ФИГ. 1

RU 2711125 C2

RU 2711125 C2

RUSSIAN FEDERATION

FEDERAL SERVICE
FOR INTELLECTUAL PROPERTY

(19) **RU** (11) **2 711 125**(13) **C2**

(51) Int. Cl.
G06K 9/62 (2006.01)
G06N 20/00 (2019.01)
G06N 3/02 (2006.01)

(12) ABSTRACT OF INVENTION

(52) CPC
G06K 9/62 (2019.02); G06N 20/00 (2019.02); G06N 3/02 (2019.02); G06F 17/40 (2019.02); G06F 15/00 (2019.02); G06T 1/20 (2019.02)

(21)(22) Application: **2017142709, 07.12.2017**

(24) Effective date for property rights:
07.12.2017

Registration date:
15.01.2020

Priority:
(22) Date of filing: **07.12.2017**

(43) Application published: **10.06.2019** Bull. № **16**

(45) Date of publication: **15.01.2020** Bull. № **2**

Mail address:
119021, Moskva, ul. Lva Tolstogo, 16, OOO "Yandeks", otdel pravovoj okhrany tekhnologij

(72) Inventor(s):
Lakhman Konstantin Viktorovich (RU),
Chigorin Aleksandr Aleksandrovich (RU),
Yurchenko Viktor Sergeevich (RU)

(73) Proprietor(s):
OBSHCHESTVO S OGRANICHENNOJ OTVETSTVENNOSTYU "YANDEKS" (RU)

RU 2711125 C2

(54) SYSTEM AND METHOD OF FORMING TRAINING SET FOR MACHINE LEARNING ALGORITHM

(57) Abstract:

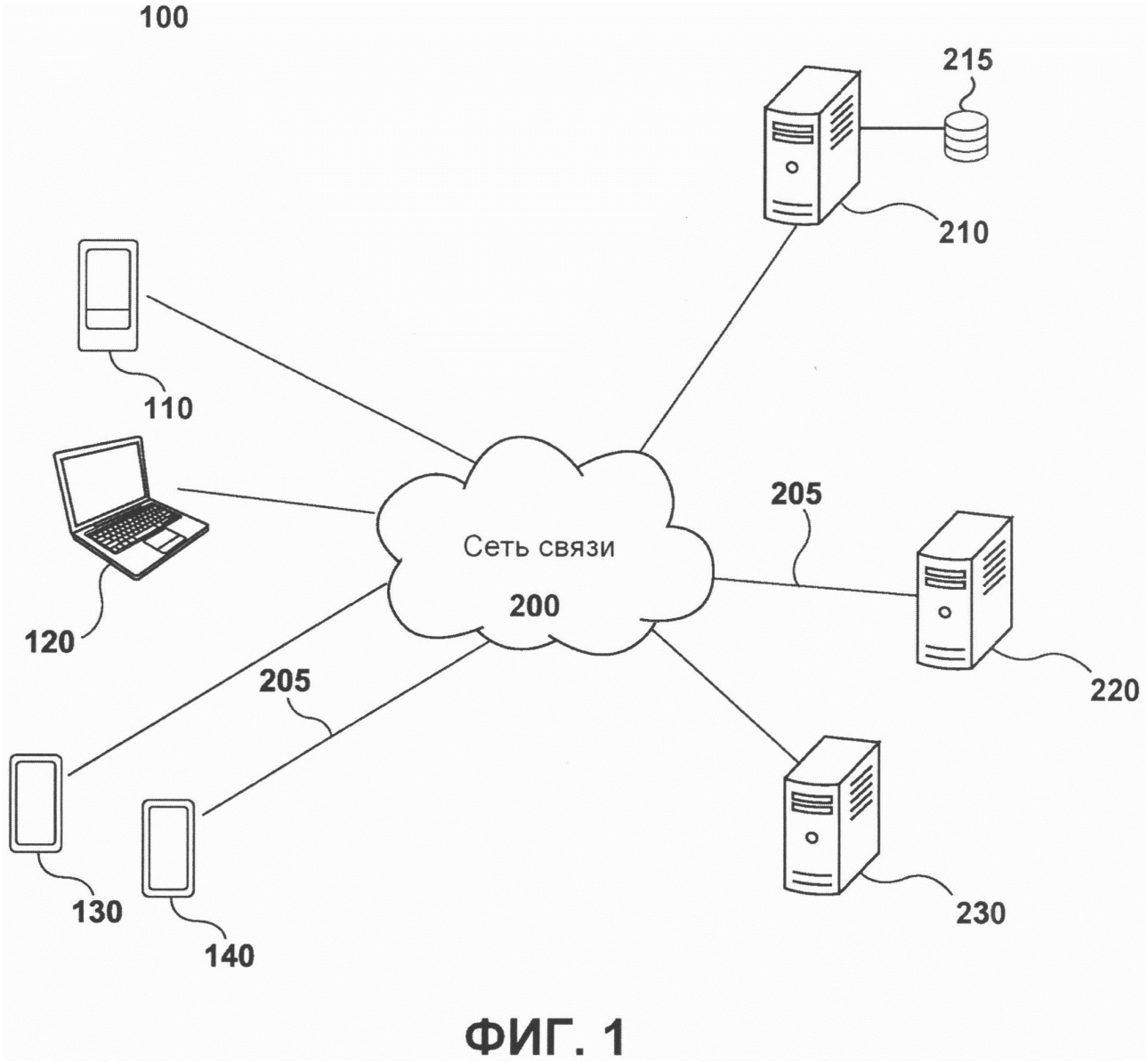
FIELD: physics.

SUBSTANCE: invention relates to the computer equipment. Method and system for generating a set of training objects for machine learning algorithm (MLA) include: obtaining search request data, each of which is associated with a first set of image search results; generating a query vector for each of the search requests; distributing query vectors between multiple clusters of query vectors; associating with each cluster a query vector of a second set of image search results comprising at least a portion of each of the first sets of image search results associated with the query vectors included in each corresponding cluster of query vectors; storing for each cluster search vectors of each image search result from a second set of image search results in the form of a training object in a set of training objects, wherein each image search result is associated with a cluster mark.

EFFECT: technical result is wider range of technical means for forming a set of training objects for machine learning algorithm and learning machine learning algorithm using formed set.

20 cl, 5 dwg

RU 2711125 C2



ФИГ. 1

RU 2711125 C2

RU 2711125 C2

ОБЛАСТЬ ТЕХНИКИ, К КОТОРОЙ ОТНОСИТСЯ ИЗОБРЕТЕНИЕ

[1] Настоящая технология в целом относится к алгоритмам машинного обучения и, в частности, к способу и системе для формирования обучающего набора для обучения алгоритма машинного обучения.

УРОВЕНЬ ТЕХНИКИ

[2] Усовершенствование компьютерной техники и технологии в сочетании с увеличением количества подключенных электронных устройств привело к росту заинтересованности в разработке систем искусственного интеллекта и решений для автоматизации выполнения задач, предсказания итогов, классификации информации и обучения на опыте, что привело к возникновению машинного обучения. Машинное обучение, тесно связанное с интеллектуальным анализом данных, вычислительной статистикой и оптимизацией, имеет дело с изучением и созданием алгоритмов, способных обучаться и выполнять прогнозирование на основе данных.

[3] В течение последнего десятилетия область машинного обучения значительно расширилась, что обеспечило значительные успехи в веб-поиске, распознавании образов и речи, создании самоуправляемых автомобилей, персонализации, понимании генома человека и т.д.

[4] Компьютерное зрение, также известное как машинное зрение, представляет собой область машинного обучения, связанную с автоматическим извлечением, анализом и пониманием полезной информации, содержащейся в отдельном изображении или в последовательности изображений. Одна из распространенных задач систем компьютерного зрения заключается в классификации изображений по категориям на основе признаков, извлеченных из изображения. Например, система компьютерного зрения может классифицировать изображения, как содержащие или не содержащие обнаженное тело для цензуры (например, как часть приложений для родительского контроля).

[5] Доказано, что нейронные сети (NN) и глубинное обучение являются методами обучения, применимыми в компьютерном зрении, распознавании речи, образов и последовательностей, интеллектуальном анализе данных, переводе, извлечении информации и т.д. В целом, нейронные сети обычно имеют слои, состоящие из соединенных друг с другом узлов с функциями активации. Образы могут загружаться в сеть через входной слой, соединенный со скрытыми слоями, а обработка может выполняться посредством взвешенных соединений узлов. Ответ выводится посредством выходного слоя, соединенного со скрытыми слоями.

[6] Алгоритмы машинного обучения (MLA) можно разделить на широкие категории, такие как обучение с учителем, обучение без учителя и обучение с подкреплением. В случае обучения с учителем обучающие данные, состоящие из входной и выходной информации, размеченной экспертами, анализируются алгоритмом машинного обучения, при этом цель обучения заключается в определении алгоритмом машинного обучения общего правила для определения соответствия между входной и выходной информацией. В случае обучения без учителя анализируются неразмеченные данные с применением алгоритма машинного обучения, при этом цель заключается в поиске алгоритмом машинного обучения структуры или скрытых закономерностей в данных. В случае обучения с подкреплением алгоритм развивается в меняющихся условиях без использования размеченных данных или исправления ошибок.

[7] Важный аспект обучения с учителем заключается в подготовке для алгоритма машинного обучения большого количества качественных обучающих наборов данных для улучшения прогнозирующей способности MLA. Обычно обучающие наборы данных

размечаются экспертами, которые присваивают документам метки релевантности с использованием оценки человеком. Эксперты могут помечать пары запрос-документ, изображения, видеоматериалы и т.д. как релевантные или нерелевантные с использованием числовых оценок или любым другим способом.

[8] Разработаны различные способы обучения алгоритмов MLA с применением нейронных сетей и методов глубинного обучения.

[9] Например, первый способ предусматривает обучение MLA на обучающих примерах, включающих в себя изображения, предварительно размеченные экспертами в соответствии с поставленной задачей (например, для классификации изображений по породам собак). Затем ML A получает ранее неизвестные данные (то есть изображения, на которых представлены собаки), с целью классификации алгоритмом ML A изображений по породе собаки. В этом случае, если MLA требуется использовать для новой задачи (например, для классификации изображений на основе присутствия или отсутствия обнаженного тела), то MLA должен быть обучен на обучающих примерах, относящихся к новой задаче.

[10] Второй способ, известный как перенос навыков, предусматривает предварительное обучение MLA на большом наборе данных обучающих примеров, которые могут не относится к какой-либо конкретной задаче, и последующее обучение MLA на более конкретном меньшем наборе данных для конкретной задачи. Такой способ позволяет экономить время и ресурсы за счет предварительного обучения MLA.

[11] В патентной заявке США №2016/140438 A1, опубликованной 19 мая 2016 г., (Hyper-Class Augmented And Regularized Deep Learning For Fine-Grained Image Classification, Nec Laboratories America Inc.) описаны системы и способы обучения обучающейся машины, которые предусматривают дополнение данных, полученных в результате мелкоструктурного распознавания изображения, размеченными данными с указанием на один или несколько гиперклассов; выполнение многозадачного глубинного обучения; мелкоструктурную классификацию и классификацию на уровне гиперклассов для совместного использования и обучения одних и тех же функциональных слоев; применение регуляризации в многозадачном глубинном обучении для использования одной или нескольких взаимосвязей между мелкоструктурными классами и гиперклассами.

[12] В патентной заявке США №2011/258149 A1, опубликованной 19 апреля 2011 г., (Ranking Search Results Using Click-Based Data, Microsoft Corp.) описаны способы и компьютерный носитель информации, содержащий выполняемые компьютером команды, которые упрощают формирование модели с машинным обучением для ранжирования результатов поиска с использованием данных, основанных на выборе пользователей. Данные основаны на запросах пользователей, которые могут включать в себя результаты поиска, сформированные стандартными поисковыми системами и вертикальными поисковыми системами. Из результатов поиска формируется обучающий набор, в котором основанные на выборе пользователей оценки связываются с результатами поиска. Исходя из основанных на выборе пользователей оценок, определяются опознаваемые признаки из результатов поиска в обучающем наборе. На основе определения опознаваемых признаков в обучающем наборе формируется набор правил для ранжирования последующих результатов поиска.

[13] В патентной заявке США №2016/0125274 A1, опубликованной 5 мая 2016 г., (Discovering visual concepts from weakly labeled image collections, PayPal Inc.) указано, что изображения, загружаемые на веб-сайты фотохостинга, часто включают в себя некоторые теги или фразовые описания. В примере осуществления эти теги или описания,

которые могут относится к содержимому изображения, используются как слабые метки этих изображений. Слабые метки могут применяться для идентификации концептов изображений с использованием итеративного жесткого алгоритма обучения на примерах для выявления визуальных концептов из представлений меток и визуальных признаков в изображениях со слабыми метками. Средства обнаружения визуальных концептов могут непосредственно применяться для распознавания и обнаружения концептов.

РАСКРЫТИЕ ИЗОБРЕТЕНИЯ

[14] Разработчики настоящей технологии исходят из по меньшей мере одной технической проблемы, связанной с использованными ранее подходами к формированию обучающих наборов для алгоритмов машинного обучения. Техническая проблема, решаемая настоящей технологией, заключается в расширении арсенала технических средств определенного назначения, а именно, технических средств для формирования набора обучающих объектов для алгоритма машинного обучения и обучения алгоритма машинного обучения с использованием сформированного набора. Технический результатом является реализация указанного назначения.

[15] Разработчики настоящей технологии исходят из того, что для MLA, реализующего нейронные сети и алгоритмы глубинного обучения, требуется большое количество документов на этапе обучения. Несмотря на очевидность подхода, заключающегося в использовании размеченных экспертами документов, вследствие огромного количества необходимых документов это оказывается утомительной, трудоемкой и дорогостоящей задачей. На оценки экспертов также может влиять их необъективность, особенно, когда для разметки требуется субъективное решение (например, с точки зрения соответствия изображения определенному поисковому запросу и т.д.).

[16] В частности, разработчики настоящей технологии исходят из того, что несмотря на существование огромных общедоступных наборов данных, таких как ImageNet™, которые могут быть полезными при формировании обучающих наборов данных для обучения и предварительного обучения MLA, такие наборы данных преимущественно содержат изображения определенных категорий, не всегда содержат достаточно классов изображений и не всегда соответствуют обычным запросам пользователей при вертикальном поиске изображений.

[17] Кроме того, наборы данных, содержащие сформированные пользователями теги и текст, не всегда соответствуют поставленной задаче (и могут рассматриваться как недостаточно качественные для целей обучения).

[18] Разработчики настоящей технологии исходят из того, что операторам поисковых систем, таких как Google™, Yandex™, Bing™, Yahoo™ и т.д., доступно большое количество данных о действиях пользователей после получения результатов поиска. В частности, поисковые системы обычно выполняют «вертикальные поиски», которые относятся к вертикали изображений. Иными словами, когда некоторый пользователь ищет изображения, типовая поисковая система представляет ему результаты из вертикали изображений. Затем данный пользователь может «взаимодействовать» с такими результатами вертикального поиска изображений. Эти действия включают в себя предварительный просмотр, пропуск, выбор и т.д.

[19] Варианты осуществления настоящей технологии относятся к способу и системе формирования обучающего набора для алгоритма машинного обучения на основе данных о действиях пользователей, полученных из журнала поисковой системы.

[20] В соответствии с первым аспектом настоящей технологии разработан способ формирования набора обучающих объектов для алгоритма машинного обучения (MLA), предназначенного для классификации изображений, при этом способ выполняется на

сервере, осуществляющем ML A, и включает в себя: получение из журнала поиска данных поисковых запросов, выполненных во время вертикального поиска изображений, каждый из которых связывается с первым набором результатов поиска изображений; формирование вектора запроса для каждого из поисковых запросов; распределение векторов запросов между множеством кластеров векторов запросов; связывание с каждым кластером векторов запросов второго набора результатов поиска изображений, содержащего по меньшей мере часть каждого из первых наборов результатов поиска изображений, связанных с векторами запросов, входящими в состав каждого соответствующего кластера векторов запросов; формирование набора обучающих объектов путем сохранения для каждого кластера векторов запросов каждого результата поиска изображений из второго набора результатов поиска изображений в виде обучающего объекта в наборе обучающих объектов, при этом каждый результат поиска изображений связывается с меткой кластера, указывающей на кластер векторов запросов, с которым связан результат поиска изображений.

[21] В некоторых вариантах осуществления формирование вектора запроса включает в себя применение алгоритма векторизации слов (word embedding) для каждого поискового запроса.

[22] В некоторых вариантах осуществления перед связыванием второго набора результатов поиска изображений с каждым кластером векторов запросов способ дополнительно включает в себя: получение для каждого первого набора результатов поиска изображений соответствующего набора метрик, каждая из которых указывает на действия пользователя с соответствующим результатом поиска изображений из первого набора результатов поиска изображений, при этом связывание второго набора результатов поиска изображений с каждым кластером векторов запросов включает в себя: выбор по меньшей мере части каждого первого набора результатов поиска изображений, включенных в состав второго набора результатов поиска изображений, на основе превышающих заранее заданный порог соответствующих метрик результатов поиска изображений из первого набора результатов поиска изображений.

[23] В некоторых вариантах осуществления кластеры векторов запросов формируются на основе степени близости векторов запросов в N-мерном пространстве.

[24] В некоторых вариантах осуществления используется один из следующих алгоритмов векторизации слов: word2vec, GloVe (глобальные векторы для представления слов), LDA2Vec, sense2vec и wang2vec.

[25] В некоторых вариантах осуществления кластеризация осуществляется с использованием одного из следующих алгоритмов: кластеризация методом k-средних, кластеризация методом максимизации ожиданий, кластеризация методом максимальной удаленности (farthest first clustering), иерархическая кластеризация, кластеризация методом cobweb и кластеризация на основе плотности.

[26] В некоторых вариантах осуществления каждый результат поиска изображений из первого набора результатов поиска изображений связан с соответствующей метрикой, указывающей на действия пользователя с результатом поиска изображений, а формирование вектора запроса включает в себя: формирование вектора признаков для каждого результата поиска изображений из выбранного подмножества результатов поиска изображений, связанных с поисковым запросом; взвешивание каждого вектора признаков с использованием соответствующей метрики; объединение векторов признаков, взвешенных с использованием соответствующих метрик.

[27] В некоторых вариантах осуществления перед формированием вектора признаков для каждого результата поиска изображений из выбранного подмножества результатов

поиска изображений способ дополнительно включает в себя: выбор по меньшей мере части каждого первого набора результатов поиска изображений, включенных в состав выбранного подмножества результатов поиска изображений, на основе превышающих заранее заданный порог соответствующих метрик результатов поиска изображений из первого набора результатов поиска изображений.

[28] В некоторых вариантах осуществления второй набор результатов поиска изображений включает в себя все результаты поиска изображений из первого набора результатов поиска изображений, связанных с векторами запросов, входящими в состав каждого соответствующего кластера.

[29] В некоторых вариантах осуществления соответствующая метрика представляет собой коэффициент переходов (CTR) или количество переходов.

[30] В некоторых вариантах осуществления кластеризация осуществляется с использованием одного из следующих алгоритмов: кластеризация методом k-средних, кластеризация методом максимизации ожиданий, кластеризация методом максимальной удаленности, иерархическая кластеризация, кластеризация методом cobweb и кластеризация на основе плотности.

[31] В соответствии со вторым аспектом настоящей технологии разработан способ обучения алгоритма машинного обучения (MLA), предназначенного для классификации изображений, при этом способ выполняется на сервере, осуществляющем MLA, и включает в себя: получение из журнала поиска данных поисковых запросов, выполненных во время вертикального поиска изображений, каждый из которых связывается с первым набором результатов поиска изображений, при этом каждый результат поиска изображений связан с соответствующей метрикой, указывающей на действия пользователя с результатом поиска изображений; выбор для каждого поискового запроса результатов поиска изображений из первого набора результатов поиска изображений, имеющих соответствующую метрику, превышающую заранее заданный порог, для добавления в соответствующее выбранное подмножество результатов поиска изображений; формирование вектора признаков для каждого результата поиска изображений из соответствующего выбранного подмножества результатов поиска изображений, связанных с каждым поисковым запросом; формирование вектора запроса для каждого поискового запроса на основе векторов признаков и соответствующих метрик результатов поиска изображений из соответствующего выбранного подмножества результатов поиска изображений; распределение векторов запросов между множеством кластеров векторов запросов; связывание с каждым кластером векторов запросов второго набора результатов поиска изображений, включающего в себя соответствующие выбранные подмножества результатов поиска изображений, связанных с векторами запросов, входящими в состав каждого соответствующего кластера векторов запросов; формирование набора обучающих объектов путем сохранения для каждого кластера векторов запросов каждого результата поиска изображений из второго набора результатов поиска изображений в виде обучающего объекта в наборе обучающих объектов, при этом каждый результат поиска изображений связывается с меткой кластера, указывающей на кластер векторов запросов, с которым связан результат поиска изображений; и обучение MLA для классификации изображений с использованием сохраненного набора обучающих объектов.

[32] В некоторых вариантах осуществления обучение представляет собой первый этап обучения с целью грубого обучения MLA для классификации изображений.

[33] В некоторых вариантах осуществления способ дополнительно включает в себя

точное обучение MLA с использованием дополнительного набора точно настроенных обучающих объектов.

[34] В некоторых вариантах осуществления MLA представляет собой алгоритм обучения искусственной нейронной сети (ANN).

[35] В некоторых вариантах осуществления MLA представляет собой алгоритм глубинного обучения.

[36] В соответствии с третьим аспектом настоящей технологии разработана система формирования набора обучающих объектов для алгоритма машинного обучения (MLA), предназначенного для классификации изображений, содержащая процессор и машиночитаемый физический носитель информации, содержащий команды, при выполнении которых процессор осуществляет следующие действия: получение из журнала поиска данных поисковых запросов, выполненных во время вертикального поиска изображений, каждый из которых связывается с первым набором результатов поиска изображений; формирование вектора запроса для каждого поискового запроса; распределение векторов запросов между множеством кластеров векторов запросов; связывание с каждым кластером векторов запросов второго набора результатов поиска изображений, содержащего по меньшей мере часть каждого из первых наборов результатов поиска изображений, связанных с векторами запросов, входящими в состав каждого соответствующего кластера векторов запросов; формирование набора обучающих объектов путем сохранения для каждого кластера векторов запросов каждого результата поиска изображений из второго набора результатов поиска изображений в виде обучающего объекта в наборе обучающих объектов, при этом каждый результат поиска изображений связывается с меткой кластера, указывающей на кластер векторов запросов, с которым связан результат поиска изображений.

[37] В некоторых вариантах осуществления каждый результат поиска изображений из первого набора результатов поиска изображений связан с соответствующей метрикой, указывающей на действия пользователя с результатом поиска изображений, при этом для формирования вектора запроса процессор выполняет следующие действия: формирование вектора признаков для каждого результата поиска изображений из выбранного подмножества результатов поиска изображений, связанных с каждым поисковым запросом; взвешивание каждого вектора признаков с использованием соответствующей метрики; объединение векторов признаков, взвешенных с использованием соответствующих метрик.

[38] В некоторых вариантах осуществления перед формированием вектора признаков для каждого результата поиска изображений из выбранного подмножества результатов поиска изображений процессор дополнительно выполняет следующие действия: выбор по меньшей мере части каждого первого набора результатов поиска изображений, включенных в состав выбранного подмножества результатов поиска изображений, на основе превышающих заранее заданный порог соответствующих метрик результатов поиска изображений из первого набора результатов поиска изображений.

[39] В некоторых вариантах осуществления второй набор результатов поиска изображений включает в себя все результаты поиска изображений из первого набора результатов поиска изображений, связанных с векторами запросов, входящими в состав каждого соответствующего кластера.

[40] В контексте настоящего описания термин «сервер» означает компьютерную программу, выполняемую соответствующими аппаратными средствами и способную принимать запросы (например, от электронных устройств) через сеть и выполнять эти запросы или инициировать их выполнение. Аппаратные средства могут представлять

собой один физический компьютер или одну компьютерную систему, но это не критично для данной технологии. В настоящем контексте выражение «сервер» не означает, что каждая задача (например, принятая команда или запрос) или некоторая определенная задача принимается, выполняется или запускается одним и тем же сервером (т.е. одними и теми же программными и/или аппаратными средствами); это выражение означает, что любое количество программных средств или аппаратных средств может принимать, отправлять, выполнять или инициировать выполнение любой задачи или запроса либо результатов любых задач или запросов; все эти программные и аппаратные средства могут представлять собой один сервер или несколько серверов, при этом оба эти случая подразумеваются в выражении «по меньшей мере один сервер».

[41] В контексте настоящего описания термин «электронное устройство» означает любое компьютерное аппаратное средство, способное выполнять программы, подходящие для решения данной задачи. Таким образом, некоторые (не имеющие ограничительного характера) примеры электронных устройств включают в себя персональные компьютеры (настольные, ноутбуки, нетбуки и т.п.), смартфоны и планшеты, а также сетевое оборудование, такое как маршрутизаторы, коммутаторы и шлюзы. Следует отметить, что в данном контексте устройство, функционирующее как электронное устройство, также может функционировать как сервер для других электронных устройств. Использование выражения «электронное устройство» не исключает использования нескольких электронных устройств для приема, отправки, выполнения или инициирования выполнения любой задачи или запроса либо результатов любых задач или запросов, либо шагов любого описанного здесь способа.

[42] В контексте настоящего описания термин «база данных» означает любой структурированный набор данных, независимо от его конкретной структуры, программного обеспечения для управления базой данных или компьютерных аппаратных средств для хранения этих данных, их применения или обеспечения их использования иным способом. База данных может располагаться в тех же аппаратных средствах, что и процесс, обеспечивающий хранение или использование информации, хранящейся в базе данных, либо база данных может располагаться в отдельных аппаратных средствах, таких как специализированный сервер или множество серверов.

[43] В контексте настоящего описания выражение «информация» включает в себя информацию любого вида, допускающую хранение в базе данных. Таким образом, информация включает в себя аудиовизуальные произведения (изображения, фильмы, звукозаписи, презентации и т.д.), данные (данные о местоположении, числовые данные и т.д.), текст (мнения, комментарии, вопросы, сообщения и т.д.), документы, электронные таблицы и т.д., но не ограничивается ими.

[44] В контексте настоящего описания выражение «пригодный для использования в компьютере носитель информации» означает носители любого типа и вида, такие как ОЗУ, ПЗУ, диски (CD-ROM, DVD, гибкие диски, жесткие диски и т.д.), USB-накопители, твердотельные накопители, накопители на магнитных лентах и т.д.

[45] В контексте настоящего описания, если явно не указано другое, в качестве данных об информационном элементе может выступать сам информационный элемент, а также указатель, ссылка, гиперссылка или другое косвенное средство, с помощью которого получатель данных может найти место в сети, в памяти, в базе данных или на другом машиночитаемом носителе информации, откуда можно извлечь этот информационный элемент. Например, данные о документе могут содержать сам документ (т.е. его содержимое) или они могут представлять собой уникальный дескриптор документа, указывающий файл в определенной файловой системе, или

какие-либо другие средства для указания получателю этих данных места в сети, адреса памяти, таблицы в базе данных или другого места, где можно получить доступ к файлу. Специалистам в данной области очевидно, что степень точности, требуемая для таких данных, зависит от объема предварительных пояснений относительно интерпретации информации, которой обмениваются отправитель и получатель данных. Например, если перед началом обмена данными между отправителем и получателем известно, что данные информационного элемента представляют собой ключ базы данных для элемента в определенной таблице заранее заданной базы данных, содержащей информационный элемент, то для эффективной передачи этого информационного элемента получателю достаточно оправить ключ базы данных, даже если сам информационный элемент не передается между отправителем и получателем данных.

[46] В контексте настоящего описания числительные «первый» «второй», «третий» и т.д. используются только для указания различия между существительными, к которым они относятся, но не для описания каких-либо определенных взаимосвязей между этими существительными. Например, должно быть понятно, что использование терминов «первый сервер» и «третий сервер» не подразумевает какого-либо определенного порядка, типа, хронологии, иерархии или классификации, в данном случае, серверов, а также что их использование (само по себе) не подразумевает наличие «второго сервера» в любой ситуации. Кроме того, как встречается в настоящем описании в другом контексте, ссылка на «первый» элемент и «второй» элемент не исключает того, что эти два элемента могут быть одним и тем же реальным элементом. Таким образом, например, в некоторых случаях «первый» сервер и «второй» сервер могут представлять собой одно и то же программное и/или аппаратное средство, а в других случаях - различные программные и/или аппаратные средства.

[47] Каждый вариант осуществления настоящей технологии имеет отношение к по меньшей мере одной из вышеупомянутых целей и/или аспектов, но не обязательно ко всем ним. Должно быть понятно, что некоторые аспекты настоящей технологии, связанные с попыткой достижения вышеупомянутой цели, могут не соответствовать этой цели и/или могут соответствовать другим целям, явным образом здесь не упомянутым.

[48] Дополнительные и/или альтернативные признаки, аспекты и преимущества вариантов реализации настоящей технологии очевидны из дальнейшего описания, приложенных чертежей и формулы изобретения.

КРАТКОЕ ОПИСАНИЕ ЧЕРТЕЖЕЙ

[49] Дальнейшее описание приведено для лучшего понимания настоящей технологии, а также других аспектов и их признаков, и должно использоваться совместно с приложенными чертежами.

[50] На фиг. 1 представлена схема системы, реализованной согласно вариантам осуществления настоящей технологии, не имеющим ограничительного характера.

[51] На фиг. 2 приведено схематическое представление первого генератора обучающей выборки согласно вариантам осуществления настоящей технологии, не имеющим ограничительного характера.

[52] На фиг. 3 приведено схематическое представление второго генератора обучающей выборки согласно вариантам осуществления настоящей технологии, не имеющим ограничительного характера.

[53] На фиг. 4 представлена блок-схема способа, реализующего первый генератор обучающей выборки и выполняемого в системе по фиг. 1.

[54] На фиг. 5 представлена блок-схема способа, реализующего второй генератор

обучающей выборки и выполняемого в системе по фиг. 1.

Осуществление изобретения

[55] Представленные в данном описании примеры и условный язык предназначены для лучшего понимания принципов настоящей технологии, а не для ограничения ее объема до таких конкретных примеров и условий. Очевидно, что специалисты в данной области техники способны разработать различные способы и устройства, которые явно не описаны и не показаны, но осуществляют принципы настоящей технологии в пределах ее существа и объема.

[56] Кроме того, чтобы способствовать лучшему пониманию, последующее описание может содержать упрощенные варианты реализации настоящей технологии. Специалистам в данной области очевидно, что различные варианты реализации настоящей технологии могут быть значительно сложнее.

[57] В некоторых случаях также приводятся полезные примеры модификаций настоящей технологии. Они способствуют пониманию, но также не определяют объем или границы настоящей технологии. Представленный перечень модификаций не является исчерпывающим и специалист в данной области может разработать другие модификации в пределах объема настоящей технологии. Кроме того, если в некоторых случаях модификации не описаны, это не означает, что они невозможны и/или что описание содержит единственный вариант реализации того или иного элемента настоящей технологии.

[58] Более того, описание принципов, аспектов и вариантов осуществления настоящей технологии, а также их конкретные примеры, предназначены для охвата их структурных и функциональных эквивалентов, независимо от того, известны они в настоящее время или будут разработаны в будущем. Например, специалистам в данной области техники должно быть очевидно, что любые представленные здесь структурные схемы соответствуют концептуальным представлениям иллюстративных схем, осуществляющих принципы настоящей технологии. Аналогично, должно быть очевидно, что любые блок-схемы, схемы процессов, диаграммы изменения состояния, псевдокоды и т.п. соответствуют различным процессам, которые могут быть представлены на машиночитаемом носителе информации и могут выполняться компьютером или процессором, независимо от того, показан ли явно такой компьютер или процессор или нет.

[59] Функции различных элементов, показанных на фигурах, включая любой функциональный блок, обозначенный как «процессор» или «графический процессор», могут быть реализованы с использованием специализированных аппаратных средств, а также с использованием аппаратных средств, способных выполнять соответствующее программное обеспечение. Если используется процессор, эти функции могут выполняться одним выделенным процессором, одним совместно используемым процессором или множеством отдельных процессоров, некоторые из которых могут использоваться совместно. В некоторых вариантах осуществления настоящей технологии процессор может представлять собой процессор общего назначения, такой как центральный процессор (CPU), или специализированный процессор, такой как графический процессор (GPU). Кроме того, явное использование термина «процессор» или «контроллер» не должно трактоваться как указание исключительно на аппаратные средства, способные выполнять программное обеспечение, и может подразумевать, помимо прочего, аппаратные средства цифрового сигнального процессора (DSP), сетевой процессор, специализированную интегральную схему (ASIC), программируемую вентильную матрицу (FPGA), ПЗУ для хранения программного обеспечения, ОЗУ и

энергонезависимое ЗУ. Также могут подразумеваться другие аппаратные средства, обычные и/или заказные.

[60] Программные модули или просто модули, реализуемые в виде программных средств, могут быть представлены в данном документе как любое сочетание элементов блок-схемы или других элементов, указывающих на выполнение шагов процесса и/или содержащих текстовое описание. Такие модули могут выполняться аппаратными средствами, которые показаны явно или подразумеваются.

[61] Учитывая вышеизложенные принципы, далее рассмотрены некоторые не имеющие ограничительного характера примеры, иллюстрирующие различные варианты реализации аспектов настоящей технологии.

[62] На фиг. 1 представлена система 100, реализованная согласно вариантам осуществления настоящей технологии. Система 100 содержит первое клиентское устройство 110, второе клиентское устройство 120, третье клиентское устройство 130 и четвертое клиентское устройство 140, соединенные с сетью 200 связи соответствующими линиями 205 связи. Система 100 содержит сервер 210 поисковой системы, сервер 220 анализа и обучающий сервер 230, соединенные с сетью 200 связи соответствующими линиями 205 связи.

[63] В качестве примера, первое клиентское устройство 110 может быть реализовано как смартфон, второе клиентское устройство 120 может быть реализовано как ноутбук, третье клиентское устройство 130 может быть реализовано как смартфон, а четвертое клиентское устройство 140 может быть реализовано как планшет. В некоторых не имеющих ограничительного характера вариантах осуществления настоящей технологии сеть 200 связи может представлять собой сеть Интернет. В других вариантах осуществления настоящей технологии сеть 200 связи может быть реализована иначе, например, в виде произвольной глобальной сети связи, локальной сети связи, личной сети связи и т.д.

[64] На реализацию линии 205 связи не накладывается каких-либо особых ограничений, она зависит от реализации первого клиентского устройства 110, второго клиентского устройства 120, третьего клиентского устройства 130 и четвертого клиентского устройства 140. В качестве примера, не имеющего ограничительного характера, в тех вариантах осуществления настоящей технологии, в которых по меньшей мере одно из клиентских устройств, таких как первое клиентское устройство 110, второе клиентское устройство 120, третье клиентское устройство 130 и четвертое клиентское устройство 140, реализовано как беспроводное устройство связи (такое как смартфон), линия 205 связи может быть реализована как беспроводная линия связи (такая как канал сети связи 3G, канал сети связи 4G, Wireless Fidelity или сокращенно WiFi®, Bluetooth® и т.п.). В тех примерах, где по меньшей мере одно из клиентских устройств, таких как первое клиентское устройство 110, второе клиентское устройство 120, третье клиентское устройство 130 и четвертое клиентское устройство 140, реализованы как ноутбук, смартфон или планшетный компьютер, линия 205 связи может быть как беспроводной (такой как Wireless Fidelity или кратко WiFi®, Bluetooth® и т.п.), так и проводной (такой как соединение на основе Ethernet).

[65] Очевидно, что варианты реализации первого клиентского устройства 110, второго клиентского устройства 120, третьего клиентского устройства 130, четвертого клиентского устройства 140, линии 205 связи и сети 200 связи приведены только для иллюстрации. Специалистам в данной области очевидны и другие конкретные детали реализации первого клиентского устройства 110, второго клиентского устройства 120,

третьего клиентского устройства 130, четвертого клиентского устройства 140, линии 205 связи и сети 200 связи. Представленные выше примеры никоим образом не ограничивают объем настоящей технологии.

[66] Несмотря на то, что на фиг.1 показаны лишь четыре клиентских устройства 110, 120, 130 и 140, предполагается, что к системе 100 может быть подключено любое количество клиентских устройств 110, 120, 130 и 140. Также предполагается, что в некоторых вариантах реализации к системе 100 могут быть подключены десятки или сотни тысяч клиентских устройств 110, 120,130 и 140.

[67] К сети 200 связи также подключен вышеупомянутый сервер 210 поисковой системы. Сервер 210 поисковой системы может быть реализован как традиционный компьютерный сервер. В примере осуществления настоящей технологии сервер 210 поисковой системы может быть реализован как сервер Dell™ PowerEdge™, работающий под управлением операционной системы Microsoft™ Windows Server™. Сервер 210 поисковой системы может быть реализован с применением любых других подходящих аппаратных средств и/или программного обеспечения и/или встроенного программного обеспечения или их сочетания. В представленном не имеющем ограничительного характера варианте осуществления настоящей технологии сервер 210 поисковой системы представляет собой один сервер. В других не имеющих ограничительного характера вариантах осуществления настоящей технологии функции сервера 210 поисковой системы могут быть распределены между несколькими серверами. В некоторых вариантах осуществления настоящей технологии сервер 210 поисковой системы управляется и/или администрируется оператором поисковой системы. В качестве альтернативы, сервер 210 поисковой системы может управляться и/или администрироваться поставщиком услуг.

[68] В целом, сервер 210 поисковой системы (i) осуществляет поиск (подробное описание приведено ниже); (ii) анализирует и ранжирует результаты поиска; (iii) группирует результаты и формирует страницу результатов поиска (SERP) для оправки в электронное устройство (такое как первое клиентское устройство 110, второе клиентское устройство 120, третье клиентское устройство 130 и четвертое клиентское устройство 140).

[69] На сервер 210 поисковой системы, предназначенный для выполнения поиска, не накладывается каких-либо особых ограничений. Специалистам в данной области известен ряд способов и средств выполнения поиска с использованием сервера 210 поисковой системы, поэтому различные структурные компоненты сервера 210 поисковой системы описаны в общем виде. Сервер 210 поисковой системы может поддерживать базу 215 данных журнала поиска.

[70] В некоторых вариантах осуществления настоящей технологии сервер 210 поисковой системы может выполнять несколько поисков, включая общий поиск и вертикальный поиск, но не ограничиваясь ими. Как известно специалистам в данной области, сервер 210 поисковой системы может выполнять общие веб-поиски. Сервер 210 поисковой системы также может выполнять один или несколько вертикальных поисков, таких как вертикальный поиск изображений, вертикальный поиск музыкальных произведений, вертикальный поиск видеоматериалов, вертикальный поиск новостей, вертикальный поиск карт и т.д. Как известно специалистам в данной области, сервер 210 поисковой системы также выполнен с возможностью выполнять алгоритм обходчика, согласно которому сервер 210 поисковой системы выполняет обход сети Интернет и индексирует посещенные веб-сайты в одной или нескольких индексных базах данных, таких как база 215 данных журнала поиска.

[71] Параллельно или последовательно с общим веб-поиском, сервер 210 поисковой системы может выполнять один или несколько вертикальных поисков в соответствующих вертикальных базах данных, которые могут входить в состав базы 215 данных журнала поиска. Для целей настоящего описания термин «вертикальный» (например, в выражении «вертикальный поиск») предназначен для обозначения поиска, выполняемого на подмножестве большего набора данных, сгруппированном в соответствии с некоторым атрибутом данных. Например, если один из вертикальных поисков выполняется сервером 210 поисковой системы в пределах сервиса изображений, то можно сказать, что сервер 210 поисковой системы выполняет поиск на подмножестве (изображения) набора данных (все потенциально доступные для поиска данные), при этом такое подмножество данных хранится в базе 215 данных журнала поиска, связанной с сервером 210 поисковой системы.

[72] Сервер 210 поисковой системы выполнен с возможностью формировать ранжированный список результатов поиска, включающий в себя результаты общего веб-поиска и результаты вертикального веб-поиска. Известно множество алгоритмов ранжирования результатов поиска, которые могут быть реализованы на сервере 210 поисковой системы.

[73] В качестве примера, не имеющего ограничительного характера, некоторые известные способы ранжирования результатов поиска по степени соответствия сделанному пользователем поисковому запросу основываются на некоторых или всех следующих критериях: (i) популярность данного поискового запроса или соответствующего ответа при выполнении поисков; (ii) количество результатов; (iii) наличие определяющих терминов (таких как «изображения», «фильмы», «погода» и т.п.) в запросе; (iv) частота использования другими пользователями данного поискового запроса с определяющими терминами; (v) частота выбора другими пользователями, выполняющими аналогичный поиск, определенного ресурса или определенных результатов вертикального поиска, когда результаты были представлены с использованием SERP. Сервер 210 поисковой системы может рассчитывать и назначать коэффициент релевантности (основанный на различных представленных выше критериях) для каждого результата поиска, полученного по сделанному пользователем поисковому запросу, а также формировать SERP, где результаты поиска ранжированы согласно их коэффициентам релевантности.

[74] К сети 200 связи также подключен вышеупомянутый сервер 220 анализа. Сервер 220 анализа может быть реализован как традиционный компьютерный сервер. В примере осуществления настоящей технологии сервер 220 анализа может быть реализован как сервер Dell™ PowerEdge™, работающий под управлением операционной системы Microsoft™ Windows Server™. Очевидно, что сервер 220 анализа может быть реализован с использованием любых других подходящих аппаратных средств и/или программного обеспечения и/или встроенного программного обеспечения или их сочетания. В представленном не имеющем ограничительного характера варианте осуществления настоящей технологии сервер 220 анализа представляет собой один сервер. В других не имеющих ограничительного характера вариантах осуществления настоящей технологии функции сервера 220 анализа могут быть распределены между несколькими серверами. В других вариантах осуществления функции сервера 220 анализа могут полностью или частично выполняться сервером 210 поисковой системы. В некоторых вариантах осуществления настоящей технологии сервер 220 анализа управляется и/или администрируется оператором поисковой системы. В качестве альтернативы, сервер 220 анализа может управляться и/или администрироваться другим поставщиком услуг.

[75] Сервер 220 анализа предназначен для отслеживания действий пользователей с результатами поиска, предоставленными сервером 210 поисковой системы по запросам пользователей (например, посредством первого клиентского устройства 110, второго клиентского устройства 120, третьего клиентского устройства 130 и четвертого клиентского устройства 140). Сервер 220 анализа может отслеживать действия пользователей или относительные данные о переходах пользователей при выполнении пользователями общих веб-поисков и вертикальных веб-поисков на сервере 210 поисковой системы. Действия пользователей могут отслеживаться сервером 220 анализа в форме метрик.

[76] Не имеющие ограничительного характера примеры отслеживаемых сервером 220 метрик включают в себя:

[77] - переходы (clicks): количество переходов, выполненных пользователем;

[78] - коэффициент переходов (CTR, click-through rate): количество случаев выбора элемента, деленное на количество показов элемента;

[79] - средний коэффициент перехода (CTR) для запроса: CTR для запроса равен 1, если выполняется один или несколько переходов, в противном случае он равен 0.

[80] Разумеется, представленный выше список не является исчерпывающим и он может включать в себя метрики других видов без выхода за пределы объема настоящей технологии.

[81] В некоторых вариантах осуществления на сервере 220 анализа могут храниться метрики и соответствующие результаты поиска. В других вариантах осуществления сервер 220 анализа может передавать метрики и соответствующие результаты поиска в базу 215 данных журнала поиска сервера 210 поисковой системы. В других не имеющих ограничительного характера вариантах осуществления настоящей технологии функции сервера 220 анализа и сервера 210 поисковой системы могут быть реализованы в одном сервере.

[82] К сети 200 связи также подключен вышеупомянутый сервер 230 обучения. Сервер 230 обучения может быть реализован как традиционный компьютерный сервер. В примере осуществления настоящей технологии сервер 230 обучения может быть реализован как сервер Dell™ PowerEdge™, работающий под управлением операционной системы Microsoft™ Windows Server™. Очевидно, что сервер 230 обучения может быть реализован с использованием любых других подходящих аппаратных средств и/или программного обеспечения и/или встроенного программного обеспечения или их сочетания. В представленном не имеющем ограничительного характера варианте осуществления настоящей технологии сервер 230 обучения представляет собой один сервер. В других не имеющих ограничительного характера вариантах осуществления настоящей технологии функции сервера 230 обучения могут быть распределены между несколькими серверами. В контексте настоящей технологии описанные здесь способы и система могут быть частично реализованы на сервере 230 обучения. В некоторых вариантах осуществления настоящей технологии сервер 230 обучения управляется и/или администрируется оператором поисковой системы. В качестве альтернативы, сервер 230 обучения может управляться и/или администрироваться другим поставщиком услуг.

[83] Сервер 230 обучения предназначен для обучения одного или нескольких алгоритмов машинного обучения (MLA), используемых сервером 210 поисковой системы, сервером 220 анализа и/или другими серверами (не показаны), связанными с оператором поисковой системы. Сервер 230 обучения может, например, обучать один или несколько алгоритмов машинного обучения, связанных с оператором поисковой системы, предназначенных для оптимизации общих и вертикальных веб-поисков,

предоставления рекомендаций, прогнозирования итогов и для других сфер применения. Обучение и оптимизация алгоритмов машинного обучения могут выполняться в течение заранее заданного периода времени или когда оператор поисковой системы сочтет это необходимым.

[84] В представленных вариантах осуществления сервер 230 обучения может быть выполнен с возможностью формировать обучающие выборки для MLA с использованием первого генератора 300 обучающей выборки и/или второго генератора 400 обучающей выборки (показаны на фиг.2 и фиг.3, соответственно) и соответствующих способов, которые более подробно описаны в последующих абзацах. Несмотря на то, что это описание относится к вертикальным поискам изображений и к результатам поиска изображений, настоящая технология также может применяться для общих веб-поисков и/или для других видов вертикальных поисков в определенной предметной области. Не ограничивая общего характера вышеизложенного, варианты осуществления настоящей технологии, не имеющие ограничительного характера, могут применяться для документов других типов, таких как результаты веб-поиска, видеоматериалы, музыка, новости, а также для поисков других видов.

[85] На фиг. 2 представлен первый генератор 300 обучающей выборки, соответствующий не имеющим ограничительного характера вариантам осуществления настоящей технологии. Первый генератор 300 обучающей выборки может быть реализован на сервере 230 обучения.

[86] Первый генератор 300 обучающей выборки включает в себя агрегатор 310 поисковых запросов, генератор 320 векторов запросов, генератор 330 кластеров и генератор 340 меток. Согласно не имеющим ограничительного характера вариантам осуществления настоящей технологии, агрегатор 310 поисковых запросов, генератор 320 векторов запросов, генератор 330 кластеров и генератор 340 меток могут быть реализованы в виде программных процедур или модулей, одного или нескольких специально запрограммированных вычислительных устройств, встроенного программного обеспечения или их сочетания.

[87] Агрегатор 310 поисковых запросов может быть выполнен с возможностью получать, объединять, фильтровать и связывать вместе запросы, результаты поиска изображений и метрики изображений. Агрегатор 310 поисковых запросов может получать из базы 215 данных журнала поиска сервера 210 поисковой системы данные 301 поисковых запросов, выполненных пользователями (например, посредством первого клиентского устройства 110, второго клиентского устройства 120, третьего клиентского устройства 130 и четвертого клиентского устройства 140) во время вертикального поиска изображений на сервере 210 поисковой системы. Данные 301 поисковых запросов могут включать в себя (1) поисковые запросы, (2) соответствующие результаты поиска изображений и (3) соответствующие метрики действий пользователя. Поисковые запросы, соответствующие результаты поиска изображений и соответствующие метрики действий пользователя могут быть получены из одной базы данных, например, из базы 215 данных журнала поиска (где они были предварительно обработаны и сохранены вместе), или из различных баз данных, например, из базы 215 данных журнала поиска и базы данных журнала анализа (не показана) сервера 220 анализа, и объединены агрегатором 310 поисковых запросов. В некоторых вариантах осуществления возможно получение только пар запрос-документ $<q_n; d_n>$, а метрики m_n, связанные с каждым документом d_n, могут быть получены из базы 215 данных журнала поиска впоследствии.

[88] В представленном варианте осуществления данные 301 поисковых запросов

включают в себя множество кортежей 304 запрос-документ-метрика в форме <q_n; d_n; m_n>, где q_n - запрос, d_n - документ или результат поиска изображений, полученный по запросу q_n во время вертикального поиска изображений на сервере 210 поисковой системы, m_n - метрика, связанная с результатом поиска изображений и указывающая на действия пользователя с результатом d_n поиска изображений, например, CTR или количество переходов.

[89] На способ выбора поисковых запросов из множества кортежей 304 запрос-документ-метрика в составе данных 301 поисковых запросов не накладывается каких-либо ограничений. Агрегатор 310 поисковых запросов может, например, получать заранее определенное количество наиболее популярных запросов, введенных пользователями сервера 210 поисковой системы при вертикальном поиске в течение заранее заданного периода времени, например, возможно получение 5000 наиболее популярных запросов q_1, …, q_{5000} (и соответствующих результатов поиска изображений), введенных в сервер 210 поисковой системы в течение последних 90 суток. В других вариантах осуществления возможно получение поисковых запросов на основе заранее заданных тем поиска, таких как люди, животные, машины, природа и т.д. В некоторых вариантах осуществления поисковые запросы q_n могут быть случайным образом выбраны из базы 215 данных журнала поиска сервера 210 поисковой системы. В некоторых вариантах осуществления поисковые запросы в данных 301 поисковых запросов могут быть выбраны в соответствии с различными критериями и могут зависеть от задачи, которая должна быть выполнена с использованием MLA.

[90] В целом, агрегатор 310 поисковых запросов может получать ограниченное или заранее заданное количество кортежей 304 запрос-документ-метрика, содержащих данный запрос q_n. В других вариантах осуществления для данного запроса q_n агрегатор 310 поисковых запросов может получать кортежи 304 запрос-документ-метрика на основе коэффициента $R(d_n)$ релевантности документа d_n на данной странице SERP в базе 215 данных журнала поиска сервера 210 поисковой системы. В не имеющем ограничительного характера примере возможно получение только кортежей 304 запрос-документ-метрика с документами, имеющими коэффициент $R(d_n)$ релевантности, превышающий заранее заданное пороговое значение. В другом не имеющем ограничительного характера примере для данного запроса q_n возможно получение только заранее заданного количества документов с наибольшими рангами (например, первые 100 ранжированных результатов поиска изображений <q_1; d_1; m_1>, …, <q_1; d_{100}; m_{100}>), выданных поисковой системой по запросу q_1. В других вариантах осуществления для данного запроса q_n возможно получение кортежей 304 запрос-документ-метрика с метриками, превышающими заранее заданный порог, например, возможно получение только кортежей 304 запрос-документ-метрика с коэффициентом CTR, превышающим 0,6.

[91] Затем агрегатор 310 поисковых запросов может связывать каждый запрос 317 с первым набором 319 результатов поиска изображений, который содержит все полученные по запросу 317 результаты поиска изображений и соответствующие метрики из данных 301 поисковых запросов. Агрегатор 310 поисковых запросов может выдавать набор 315 запросов и результатов поиска изображений.

[92] Генератор 320 векторов запросов может быть выполнен с возможностью получать в качестве входных данных набор 315 запросов и результатов поиска

изображений и выдавать набор 325 векторов запросов, при этом каждый вектор 327 запроса из набора 325 векторов запросов связан с соответствующим запросом 317 из набора 315 запросов и результатов поиска изображений. Генератор 320 векторов запросов может выполнять алгоритм векторизации слов и применять этот алгоритм векторизации слов к каждому запросу 317 из набора 315 запросов и результатов поиска изображений и формировать соответствующий вектор 327 запроса. В целом, генератор 320 векторов запросов может преобразовывать текст из запросов 317, сделанных пользователями, в числовое представление в форме вектора 327 запроса из непрерывных значений. Генератор 320 векторов запросов может представлять запросы 317 в виде векторов низкой размерности путем сохранения контекстного подобия слов. В не имеющем ограничительного характера примере генератор 320 векторов запросов может выполнять один из следующих алгоритмов векторизации слов: word2vec, GloVe (глобальные векторы для представления слов), LDA2Vec, sense2vec и wang2vec. В некоторых вариантах осуществления каждый вектор 327 запроса из набора 325 векторов запросов может также включать в себя результаты поиска изображений и соответствующие связанные с ними метрики. В некоторых вариантах осуществления набор 325 векторов запросов может, по меньшей мере частично, формироваться на основе соответствующих метрик результатов поиска изображений из первого набора 319 результатов поиска изображений из набора 315 запросов и результатов поиска изображений.

[93] Затем генератор 320 векторов запросов может выдавать набор 325 векторов запросов.

[94] Генератор 330 кластеров может быть выполнен с возможностью принимать в качестве входных данных набор 325 векторов запросов и выдавать набор 335 кластеров векторов запросов. Генератор 330 кластеров может переносить набор 325 векторов запросов в N-мерное пространство признаков, где каждый вектор 327 запроса из набора 325 векторов запросов может представлять собой точку в N-мерном пространстве признаков. В некоторых вариантах осуществления размерность N-мерного пространства может быть меньше размерности векторов 327 запросов из набора 325 векторов запросов. В других вариантах осуществления, в зависимости от способа кластеризации, генератор 330 кластеров может выполнять кластеризацию векторов 327 запросов в N-мерном пространстве признаков для получения к кластеров или подмножеств на основе функции близости или подобия. В некоторых вариантах осуществления количество кластеров может быть задано заранее. В целом, векторы 327 запросов из одного кластера 337 векторов запросов могут быть более похожими друг на друга, чем векторы 327 запросов из других кластеров. В не имеющем ограничительного характера примере векторы 327 запросов из одного кластера могут быть семантически тесно связаны друг с другом.

[95] Способы кластеризации известны в данной области техники и кластеризация может выполняться с использованием одного из следующих: алгоритма кластеризации методом k-средних, алгоритма нечеткой кластеризации методом C-средних, алгоритмов иерархической кластеризации, алгоритмов гауссовой кластеризации, алгоритмов кластеризации методом пороговых значений и т.д.

[96] Затем генератор 330 кластеров может связывать соответствующий второй набор 338 результатов поиска изображений с каждым кластером 337 векторов запросов из набора 335 кластеров векторов запросов. Соответствующий второй набор 338 результатов поиска изображений может содержать по меньшей мере часть каждого первого набора 319 результатов поиска изображений, связанных с частью векторов

327 запросов из данного кластера 337 векторов запросов. В настоящем варианте осуществления соответствующий второй набор 338 результатов поиска изображений целиком содержит первый набор 319 результатов поиска изображений. В других вариантах осуществления настоящей технологии результаты поиска изображений из первого набора 319 результатов поиска изображений, которые представляют собой часть соответствующего второго набора 338 результатов поиска изображений, также могут выбираться или отфильтровываться, если соответствующие метрики, связанные с каждым результатом поиска изображений, превышают заранее заданный порог, например, для добавления во второй набор 338 результатов поиска изображений может выбираться каждый результат поиска изображений из каждого первого набора 319 результатов поиска изображений с коэффициентом CTR, превышающим 0,6. В других вариантах осуществления генератор 330 кластеров может рассматривать только заранее заданное количество результатов поиска изображений независимо от порога, например, для добавления во второй набор 338 результатов поиска изображений могут выбираться результаты поиска изображений, связанные со 100 наиболее высокими коэффициентами CTR.

[97] Затем генератор 330 кластеров может выдать набор 335 кластеров векторов запросов, при этом каждый кластер 337 векторов запросов связан с соответствующим вторым набором 338 результатов поиска изображений.

[98] Затем генератор 340 меток может принимать в качестве входных данных набор 335 кластеров векторов запросов, при этом каждый кластер 337 векторов запросов связан с соответствующим вторым набором 338 результатов поиска изображений. Далее каждый результат поиска изображений из второго набора 338 результатов поиска изображений, связанных с каждым кластером 337 векторов запросов, может быть размечен генератором 340 меток с применением идентификатора кластера, который может использоваться в качестве метки для обучения MLA на сервере 230 обучения. Каждый кластер 337 векторов запросов может представлять собой коллекцию семантически связанных запросов, каждый из которых связан с результатами поиска изображений, наилучшим образом представляющими запрос с точки зрения пользователей сервера 210 поисковой системы. Часть результатов поиска изображений из тех же кластеров запросов может быть размечена одной меткой (вследствие их принадлежности к одному и тому же кластеру) и может использоваться для обучения MLA. Таким образом, варианты осуществления настоящей технологии обеспечивают кластеризацию результатов поиска изображений для данного поискового запроса с назначением им метки кластера (вследствие их принадлежности к одному и тому же кластеру). Кластеры 337 векторов запросов могут быть понятны или непонятны человеку, т.е. часть изображений из одного и того же кластера может иметь или не иметь смысл для человека, но, тем не менее, они могут быть полезными для предварительного обучения алгоритма машинного обучения, реализующего нейронные сети или алгоритмы глубинного обучения.

[99] Затем сервер 230 обучения может сохранить каждый результат поиска изображений из второго набора 338 результатов поиска изображений с соответствующей меткой кластера в виде обучающего объекта 347 для формирования набора 345 обучающих объектов.

[100] Затем набор 345 обучающих объектов может использоваться для обучения MLA на сервере 230 обучения, где алгоритм MLA должен относить предлагаемый результат поиска изображений к данному кластеру после просмотра примеров обучающих объектов 347. В других вариантах осуществления набор 345 обучающих

объектов может быть сделан общедоступным для обучения алгоритмов MLA.

[101] Набор 345 обучающих объектов может использоваться для грубого обучения MLA на первом этапе обучения для классификации изображений. Далее MLA может обучаться на втором этапе обучения на наборе точно настроенных обучающих объектов (не показан) для конкретной задачи классификации изображений.

[102] На фиг. 3 представлен второй генератор 400 обучающей выборки, соответствующий не имеющим ограничительного характера вариантам осуществления настоящей технологии. Второй генератор 400 обучающей выборки может быть реализован с применением сервера 230 обучения.

[103] Второй генератор 400 обучающей выборки содержит выделитель 430 признаков, агрегатор 420 поисковых запросов, генератор 440 векторов запросов, генератор 450 кластеров и генератор 460 меток. Согласно различным не имеющим ограничительного характера вариантам осуществления настоящей технологии, выделитель 430 признаков, агрегатор 420 поисковых запросов, генератор 440 векторов запросов, генератор 450 кластеров и генератор 460 меток могут быть реализованы в виде программных процедур или модулей, одного или нескольких специально запрограммированных вычислительных устройств, встроенного программного обеспечения или их сочетания.

[104] Агрегатор 420 поисковых запросов может быть выполнен с возможностью получать, объединять, фильтровать и связывать вместе запросы, результаты поиска изображений и метрики изображений. Агрегатор 420 поисковых запросов может получать из базы 215 данных журнала поиска на сервере 210 поисковой системы данные 401 поисковых запросов, выполненных пользователями (например, посредством первого клиентского устройства 110, второго клиентского устройства 120, третьего клиентского устройства 130 и четвертого клиентского устройства 140) во время вертикального поиска изображений на сервере 210 поисковой системы. Данные 401 поисковых запросов могут включать в себя: (1) поисковые запросы, (2) соответствующие результаты поиска изображений и (3) соответствующие метрики действий пользователя. Поисковые запросы, соответствующие результаты поиска изображений и соответствующие метрики действий пользователя возможно получать из одной базы данных, например, из базы 215 данных журнала поиска (где они предварительно обработаны и сохранены вместе), или из различных баз данных, например, из базы 215 данных журнала поиска и базы данных журнала анализа (не показана) на сервере 220 анализа, и объединять их посредством агрегатора 310 поисковых запросов.

[105] В представленном варианте осуществления данные 401 поисковых запросов включают в себя множество кортежей 404 запрос-документ-метрика в форме $<q_n; d_n; m_n>$, где q_n - запрос, d_n - документ или результат поиска изображений, полученный по запросу q_n во время вертикального поиска изображений на сервере 210 поисковой системы, m_n - метрика, связанная с результатом d_n поиска изображений и указывающая на действия пользователя с результатом d_n поиска изображений, например, CTR или количество переходов.

[106] На способ выбора поисковых запросов, водящих в состав множества кортежей 404 запрос-документ-метрика в данных 401 поисковых запросов, не накладывается каких-либо ограничений. Агрегатор 420 поисковых запросов может, например, получать заранее заданное количество наиболее популярных запросов, введенных пользователями сервера 210 поисковой системы при вертикальном поиске в течение заранее заданного периода времени, например, возможно получение 5000 наиболее популярных запросов q_n (и соответствующих результатов поиска изображений), направленных серверу 210

поисковой системы в течение последних 90 суток. В других вариантах осуществления поисковые запросы возможно получать на основе предварительно определенных тем поиска, таких как люди, животные, машины, природа и т.д. В некоторых вариантах осуществления поисковые запросы q_n могут быть выбраны случайным образом из базы 215 данных журнала поиска сервера 210 поисковой системы. В некоторых вариантах осуществления поисковые запросы в данных 401 поисковых запросов могут быть выбраны в соответствии с различными критериями и могут зависеть от задачи, которая должна быть выполнена с использованием MLA.

[107] Агрегатор 420 поисковых запросов может получать ограниченное или заранее заданное количество кортежей 404 запрос-документ-метрика, содержащих данный запрос q_n. В некоторых вариантах осуществления для данного запроса q_n агрегатор 420 поисковых запросов может получать кортежи 404 запрос-документ-метрика на основе коэффициента $R(d_n)$ релевантности документа d_n на данной странице SERP в базе 215 данных журнала поиска сервера 210 поисковой системы. В не имеющем ограничительного характера примере возможно получение только документов с коэффициентом $R(d_n)$ релевантности, превышающим заранее заданное пороговое значение. В другом не имеющем ограничительного характера примере для данного запроса q_n возможно получение только заранее заданного количества документов с наибольшими рангами (например, первые 100 ранжированных результатов поиска изображений $<q_1; d_1; m_1>$, …, $<q_1; d_{100}; m_{100}>$ с наибольшими рангами, выданных поисковой системой по запросу q_n). В других вариантах осуществления для данного запроса q_n возможно получение кортежей 404 запрос-документ-метрика с метриками, превышающими заранее заданный порог, например, возможно получение кортежей 404 запрос-документ-метрика с коэффициентом CTR, превышающим 0,6.

[108] Затем агрегатор 420 поисковых запросов может связывать каждый запрос 424 с первым набором результатов поиска изображений, который содержит все полученные по запросу 424 результаты поиска изображений и соответствующие метрики из данных 401 поисковых запросов. В вариантах осуществления, которые предусматривают фильтрацию кортежей 404 запрос-документ-метрика на основе метрик, превышающих заранее заданный порог, кортежи 404 запрос-документ-метрика могут добавляться в выбранное подмножество 426 результатов поиска изображений. Агрегатор 420 поисковых запросов может выдавать набор 422 запросов и результатов поиска изображений, в котором каждый запрос 424 связан с соответствующим подмножеством 426 результатов поиска изображений.

[109] Выделитель 430 признаков может быть выполнен с возможностью принимать в качестве входных данных набор 406 изображений и выдавать набор 432 векторов признаков. Выделитель 430 признаков может связываться с агрегатором 420 поисковых запросов с целью получения информации об изображениях из результатов поиска изображений для получения и извлечения из них признаков. В не имеющем ограничительного характера примере выделитель 430 признаков может получать идентификаторы результатов поиска изображений, отфильтрованных агрегатором 420 поисковых запросов, а также получать набор 406 изображений посредством сервера 210 поисковой системы для извлечения признаков. Изображения в наборе 406 изображений могут соответствовать всем изображениям в выбранных подмножествах 426 результатов поиска изображений из набора 422 запросов и результатов поиска изображений. В других вариантах осуществления функции выделителя 430 признаков

могут объединяться с функциями агрегатора 420 поисковых запросов.

[110] На способ извлечения признаков выделителем 430 признаков из набора 406 изображений для получения набора 432 векторов признаков не накладывается каких-либо ограничений. В некоторых не имеющих ограничительного характера вариантах осуществления настоящей технологии выделитель 430 признаков может быть реализован в виде предварительно обученной нейронной сети (выполненной с возможностью анализировать изображения и извлекать признаки изображений из проанализированных изображений). В другом не имеющем ограничительного характера примере выделитель 430 признаков может извлекать признаки с использованием одного из следующих алгоритмов выделения признаков: масштабно-инвариантной трансформации признаков (SIFT), гистограммы направленных градиентов (HOG), ускоренных робастных признаков (SURF), локальных бинарных шаблонов (LBP), вейвлетов Хаара, гистограмм цветов и т.д. Выделитель 430 признаков может выдавать набор 432 векторов признаков, в котором каждый вектор 417 признаков соответствует числовому представлению изображения, полученного по запросу из набора 402 поисковых запросов.

[111] Генератор 440 векторов запросов может быть выполнен с возможностью получать в качестве входных данных набор 432 векторов признаков и набор 422 запросов и результатов поиска изображений и выдавать набор 445 векторов запросов, при этом каждый вектор 447 запроса из набора 445 векторов запросов связан с соответствующим запросом из набора 422 запросов и результатов поиска изображений. В целом, каждый вектор 447 запроса из набора 445 векторов запросов может представлять собой низкоразмерное векторное представление признаков наиболее популярных результатов поиска изображений, полученных по данному запросу и выбранных пользователями сервера 210 поисковой системы. В одном возможном варианте осуществления для конкретного запроса вектор 447 запроса может представлять собой линейную комбинацию всех векторов 417 признаков из набора 432 векторов признаков, взвешенных с использованием константы, умноженной на соответствующую метрику. Иными словами, каждый вектор 447 запроса из набора 445 векторов запросов может представлять собой взвешенное среднее векторов признаков результатов поиска изображений из выбранного подмножества 426 результатов поиска изображений, которые наилучшим образом представляют запрос согласно выбору пользователей, взаимодействующих с сервером 210 поисковой системы. В других вариантах осуществления вектор 447 запроса может представлять собой нелинейную комбинацию соответствующих метрик и векторов признаков.

[112] Генератор 450 кластеров может быть выполнен с возможностью принимать в качестве входных данных набор 445 векторов запросов и выдавать набор 455 кластеров векторов запросов. Генератор 450 кластеров может переносить набор 445 векторов запросов в N-мерное пространство признаков, где каждый вектор 447 запроса из набора 445 векторов запросов может представлять собой точку в N-мерном пространстве признаков. Затем генератор 450 кластеров может выполнять кластеризацию векторов 447 запросов в N-мерном пространстве признаков для получения k кластеров или подмножеств на основе функции близости или подобия (например, манхэттенского расстояния, квадрата евклидова расстояния, косинусного расстояния и расстояния Брэгмана для алгоритма кластеризации методом k-средних), где векторы 447 запросов в каждом кластере считаются похожими друг на друга в соответствии с функцией близости или подобия. В не имеющем ограничительного характера примере с использованием алгоритма кластеризации методом k-средних могут определяться k центроидов в N-мерном пространстве, а векторы 447 запросов могут рассматриваться

как находящиеся в определенном кластере, если они ближе к данному центроиду, чем к любому другому центроиду. В целом, векторы 447 запросов в одном кластере 337 могут быть более похожими друг на друга, чем векторы 447 запросов в других кластерах. В зависимости от способа кластеризации, кластеры 457 векторов запросов могут быть непонятны человеку, т.е. кластеры могут не иметь смысла для человека, но тем не менее, они могут быть полезны для предварительного обучения алгоритма машинного обучения, реализующего нейронные сети или алгоритмы глубинного обучения, поскольку они содержат изображения со сходными признаками.

[113] Способы кластеризации общеизвестны. Например, кластеризация может выполняться с использованием одного из следующих: алгоритма кластеризации методом k-средних, алгоритма нечеткой кластеризации методом C-средних, алгоритмов иерархической кластеризации, алгоритмов гауссовой кластеризации, алгоритмов кластеризации методом пороговых значений и других известных в данной области техники алгоритмов.

[114] Затем генератор 450 кластеров может связывать соответствующий второй набор 458 результатов поиска изображений с каждым кластером 457 векторов запросов из набора 455 кластеров векторов запросов. Генератор 450 кластеров может анализировать каждый кластер в наборе 445 кластеров векторов запросов и получать ссылку на все изображения, связанные с векторами 447 запросов, включенными в состав каждого кластера 457 векторов запросов, в форме второго набора 458 результатов поиска изображений.

[115] Затем генератор 450 кластеров может выдавать набор 455 кластеров векторов запросов, при этом каждый кластер 457 векторов запросов из набора 455 кластеров векторов запросов включает в себя множество векторов 447 запросов из набора 445 кластеров векторов запросов и связан с соответствующим вторым набором 458 результатов поиска изображений.

[116] Генератор 460 меток может быть выполнен с возможностью принимать в качестве входных данных набор 455 кластеров векторов запросов и выдавать набор обучающих объектов 465, при этом каждый кластер 457 векторов запросов связан с соответствующим вторым набором 465 результатов поиска изображений. Затем генератор 460 меток может размечать каждый результат поиска изображений из соответствующего второго набора 458 результатов поиска изображений с использованием идентификатора кластера для получения обучающих объектов 467. На способ реализации идентификатора кластера не накладывается каких-либо ограничений. В не имеющем ограничительного характера примере каждому результату поиска изображений из второго набора 458 результатов поиска изображений может присваиваться числовой идентификатор. Генератор 460 меток может непосредственно получать и размечать изображения, а также сохранять каждый второй набор 458 результатов поиска изображений в виде набора обучающих объектов 465 на сервере 230 обучения. В других вариантах осуществления генератор 460 меток может связывать идентификаторы кластеров с каждым изображением в базе данных (не показана) сервера 230 обучения.

[117] Затем набор 465 обучающих объектов может использоваться для обучения MLA на сервере 230 обучения. В других вариантах осуществления набор 465 обучающих объектов может быть сделан общедоступным в репозитории для обучения алгоритмов MLA.

[118] Набор 465 обучающих объектов может использоваться для грубого обучения MLA на первом этапе обучения для классификации изображений. Затем MLA может

обучаться на втором этапе обучения на наборе точно настроенных обучающих объектов (не показан) для конкретной задачи классификации изображений.

[119] На фиг. 4 представлена блок-схема способа 500 формирования набора обучающих объектов для алгоритма машинного обучения. Способ 500 выполняется с использованием первого генератора 300 обучающей выборки 300 на сервере 230 обучения.

[120] Способ 500 может начинаться с шага 502.

[121] ШАГ 502: получение из журнала поиска данных о поисковых запросах, выполненных во время вертикального поиска изображений, каждый из которых связывается с первым набором результатов поиска изображений.

[122] На шаге 502 агрегатор 310 поисковых запросов сервера 230 обучения может получать из базы 215 данных журнала поиска сервера 210 поисковой системы данные 301 поисковых запросов, выполненных во время вертикального поиска изображений, при этом данные 301 поисковых запросов содержат множество кортежей 304 запрос-документ-метрика, каждый из которых включает в себя запрос, результат поиска изображений, полученный по запросу, и метрику, указывающую на действия пользователя с результатом поиска изображений. Затем агрегатор 310 поисковых запросов может выдавать набор 315 запросов и результатов поиска изображений, в котором каждый запрос 317 связан с первым набором 319 результатов поиска изображений. В некоторых вариантах осуществления каждый результат поиска изображений из первого набора 319 результатов поиска изображений связан с соответствующей метрикой, указывающей на действия пользователя с соответствующим результатом поиска изображений.

[123] Затем способ 500 может продолжаться на шаге 504.

[124] ШАГ 504: формирование вектора признаков для каждого поискового запроса с применением алгоритма векторизации слов для каждого запроса.

[125] На шаге 504 генератор 320 векторов запросов сервера 230 обучения может формировать набор 325 векторов запросов, включающий в себя вектор 327 запроса для каждого запроса из набора 315 запросов и результатов поиска изображений. Каждый вектор 327 запроса может формироваться с применением алгоритма векторизации слов для каждого запроса из набора 315 запросов и результатов поиска изображений. Может использоваться один из следующих алгоритмов векторизации слов: word2vec, GloVe (глобальные векторы для представления слов), LDA2Vec, sense2vec и wang2vec. В некоторых вариантах осуществления, в зависимости от способа кластеризации, каждый вектор 327 запроса из набора 325 векторов запросов может представлять собой точку в N-мерном пространстве признаков.

[126] Затем способ 500 может продолжаться на шаге 506.

[127] ШАГ 506: распределение векторов запросов между множеством кластеров векторов запросов.

[128] На шаге 506 генератор 330 кластеров сервера 230 обучения может выполнять кластеризацию векторов 327 запросов из набора 325 векторов запросов для получения k кластеров или подмножеств на основе функции близости или подобия. В некоторых вариантах осуществления кластеризация может выполняться на основе степени близости векторов запросов в N-мерном пространстве признаков. Генератор 330 кластеров может применять алгоритм кластеризации методом k-средних, алгоритм нечеткой кластеризации методом C-средних, алгоритмы иерархической кластеризации, алгоритмы гауссовой кластеризации и алгоритмы кластеризации методом пороговых значений.

[129] Затем способ 500 может продолжаться на шаге 508.

[130] ШАГ 508: получение для каждого результата из первого набора результатов поиска изображений соответствующего набора метрик, каждая из которых указывает на действия пользователя с соответствующим результатом поиска изображений из первого набора результатов поиска изображений.

[131] На шаге 508 агрегатор 310 поисковых запросов и/или генератор 340 меток сервера 230 обучения может получать из базы 215 данных журнала поиска соответствующий набор метрик для каждого результата поиска изображений из каждого первого набора 319 результатов поиска изображений, каждая из которых указывает на действия пользователя с соответствующим результатом поиска изображений из первого набора 319 результатов поиска изображений. В некоторых вариантах осуществления возможно получение соответствующих метрик для каждого результата поиска изображений из каждого первого набора 319 результатов поиска изображений на шаге 502 в составе данных 301 поисковых запросов.

[132] Затем способ 500 может продолжаться на шаге 510.

[133] ШАГ 510: связывание с каждым кластером векторов запросов второго набора результатов поиска изображений путем выбора результатов поиска изображений из первого набора результатов поиска изображений, которые должны быть включены во второй набор результатов поиска изображений, на основе превышающих заранее заданный порог соответствующих метрик результатов поиска изображений из первого набора результатов поиска изображений.

[134] На шаге 510 генератор 330 кластеров сервера 230 обучения может связывать с каждым кластером 337 векторов запросов из набора 335 кластеров векторов запросов второй набор 338 результатов поиска изображений путем выбора по меньшей мере части результатов поиска изображений из первого набора 319 результатов поиска изображений, которые должны быть включены во второй набор 338 результатов поиска изображений, на основе превышающих заранее заданный порог соответствующих метрик результатов поиска изображений из первого набора 319 результатов поиска изображений.

[135] Затем способ 500 может продолжаться на шаге 512.

[136] ШАГ 512: формирование набора обучающих объектов путем сохранения для каждого кластера векторов запросов каждого результата поиска изображений из второго набора результатов поиска изображений в виде обучающего объекта в наборе обучающих объектов, при этом каждый результат поиска изображений связывается с меткой кластера, указывающей на кластер векторов запросов, с которым связан результат поиска изображений.

[137] На шаге 512 генератор 340 меток сервера 230 обучения может формировать набор обучающих объектов 345 путем сохранения для каждого кластера 337 векторов запросов каждого результата поиска изображений из второго набора 338 результатов поиска изображений в виде обучающего объекта 347 в наборе 345 обучающих объектов, при этом каждый результат поиска изображений связывается с меткой кластера, указывающей на кластер 337 векторов запросов, с которым связан результат поиска изображений. Метка кластера может представлять собой слово, число или сочетание символов для уникальной идентификации кластера векторов запросов.

[138] Затем способ 500 может опционально продолжаться на шаге 514 или может завершаться на шаге 512.

[139] ШАГ 514: обучение MLA для классификации изображений с использованием сохраненного набора обучающих объектов.

[140] На шаге 514 алгоритм MLA сервера 230 обучения может обучаться с

использованием набора 345 обучающих объектов. Алгоритм MLA может получать примеры результатов поиска изображений и связанные с ними метки кластеров, а затем обучаться с целью распределения изображений в различные кластеры на основе векторов признаков, извлеченных из этих изображений.

[141] Затем способ 500 может быть завершен.

[142] В целом, первый генератор 300 обучающей выборки и способ 500 позволяют формировать кластеры семантически связанных запросов и для каждой части запросов из кластеров запросов связывать наиболее представительные результаты поиска изображений с кластерами запросов в соответствии с выбором пользователей сервера 210 поисковой системы. Таким образом, обучающие объекты могут формироваться путем назначения данной метки для части результатов поиска изображений из одного кластера.

[143] На фиг.5 представлена блок-схема способа 600 формирования набора обучающих объектов для алгоритма машинного обучения. Способ 600 выполняется с использованием второго генератора 400 обучающей выборки 300 на сервере 230 обучения.

[144] Способ 600 может начинаться с шага 602.

[145] ШАГ 602: получение из журнала поиска данных поисковых запросов, выполненных во время вертикального поиска изображений, каждый из которых связывается с первым набором результатов поиска изображений, при этом каждый результат поиска изображений связан с соответствующей метрикой, указывающей на действия пользователя с результатом поиска изображений.

[146] На шаге 602 агрегатор 420 поисковых запросов сервера 230 обучения может получать из базы 215 данных журнала поиска сервера 210 поисковой системы данные 401 поисковых запросов, выполненных во время вертикального поиска изображений на сервере 210 поисковой системы, при этом данные 401 поисковых запросов содержат множество кортежей 404 запрос-документ-метрика, каждый из которых включает в себя запрос, результат поиска изображений, полученный по этому запросу, и метрику, указывающую на действия пользователя с результатом поиска изображений.

[147] Затем способ 600 может продолжаться на шаге 604.

[148] ШАГ 604: выбор для каждого поискового запроса результатов поиска изображений из первого набора результатов поиска изображений, имеющих соответствующую метрику, превышающую заранее заданный порог, для добавления в соответствующее выбранное подмножество результатов поиска изображений.

[149] На шаге 604 агрегатор 420 поисковых запросов сервера 230 обучения может фильтровать каждый кортеж 404 запрос-документ-метрика путем выбора кортежа 404 запрос-документ-метрика с соответствующей метрикой, превышающей заранее заданный порог. Затем агрегатор 420 поисковых запросов может связывать каждый запрос 424 с выбранным подмножеством 426 результатов поиска изображений для выдачи набора 422 запросов и результатов поиска изображений.

[150] Затем способ 600 может продолжаться на шаге 606.

[151] ШАГ 606: формирование вектора признаков для каждого результата поиска изображений из соответствующего выбранного подмножества результатов поиска изображений, связанных с каждым поисковым запросом.

[152] На этапе 606 выделитель 430 признаков сервера 230 обучения может получать информацию о выбранном подмножестве 426 результатов поиска изображений из агрегатора 420 поисковых запросов, а также получать набор 406 изображений, содержащий изображения из каждого выбранного подмножества 426 результатов

поиска изображений. Затем выделитель 430 признаков может формировать вектор 434 признаков для каждого изображения из выбранного подмножества 426 результатов поиска изображений и выдавать набор 432 векторов признаков.

[153] Затем способ 600 может продолжаться на шаге 608.

[154] ШАГ 608: формирование вектора запроса для каждого поискового запроса на основе векторов признаков и соответствующих метрик результатов поиска изображений из соответствующего выбранного подмножества результатов поиска изображений.

[155] На шаге 608 генератор 440 векторов запросов сервера 230 обучения может получать набор 432 векторов признаков и набор 422 запросов и результатов поиска изображений, а затем для каждого запроса 424 из набора 422 запросов и результатов поиска изображений он может формировать вектор 447 запроса. Каждый вектор 447 запроса из набора 445 векторов запросов может формироваться для данного запроса путем взвешивания каждого вектора 434 признаков из набора 432 векторов признаков с использованием соответствующей метрики и объединения векторов 434 признаков, взвешенных с использованием соответствующих метрик. В некоторых вариантах осуществления каждый вектор 447 запроса может представлять собой линейную комбинацию векторов признаков наиболее часто выбираемых результатов поиска изображений, взвешенных с использованием соответствующих метрик.

[156] Затем способ 600 может продолжаться на шаге 610.

[157] ШАГ 610: распределение векторов запросов между множеством кластеров векторов запросов.

[158] На шаге 610 генератор 450 кластеров сервера 230 обучения может выполнять кластеризацию векторов 447 запросов из набора 445 векторов запросов для получения k кластеров или подмножеств на основе функции близости или подобия в N-мерном пространстве. Затем генератор 450 кластеров может выдавать набор 455 кластеров векторов запросов, при этом каждый кластер 457 векторов запросов из набора 455 кластеров векторов запросов содержит множество векторов 447 запросов.

[159] Затем способ 600 может продолжаться на шаге 610.

[160] ШАГ 612: связывание с каждым кластером векторов запросов второго набора результатов поиска изображений, включающего в себя соответствующие выбранные подмножества результатов поиска изображений, связанных с векторами запросов, входящими в состав каждого соответствующего кластера векторов запросов.

[161] На шаге 612 генератор 460 меток сервера 230 обучения может связывать с каждым кластером 457 векторов запросов из набора 455 кластеров векторов запросов второй набор 458 результатов поиска изображений, содержащий выбранное подмножество 426 результатов поиска изображений, связанных с каждым вектором 447 запроса, входящим в состав каждого соответствующего кластера 457 векторов запросов.

[162] Затем способ 600 может продолжаться на шаге 614.

[163] ШАГ 614: формирование набора обучающих объектов путем сохранения для каждого кластера векторов запросов каждого результата поиска изображений из второго набора результатов поиска изображений в виде обучающего объекта в наборе обучающих объектов, при этом каждый результат поиска изображений связывается с меткой кластера, указывающей на кластер векторов запросов, с которым связан результат поиска изображений.

[164] На шаге 614 генератор 460 меток сервера 230 обучения может формировать набор обучающих объектов 465 путем сохранения для каждого кластера 457 векторов запросов каждого результата поиска изображений из второго набора 458 результатов

поиска изображений в виде обучающего объекта 467 в наборе 465 обучающих объектов, при этом каждый результат поиска изображений связывается с меткой кластера, указывающей на кластер 457 векторов запросов, с которым связан результат поиска изображений.

[165] Затем способ 600 может опционально продолжаться на шаге 616 или может завершаться.

[166] ШАГ 616: обучение MLA для классификации изображений с использованием сохраненного набора обучающих объектов.

[167] На шаге 616 алгоритм MLA сервера 230 обучения может обучаться с использованием набора 465 обучающих объектов. Алгоритм MLA может получать примеры результатов поиска изображений и связанные с ними метки кластеров, а затем обучаться с целью распределения изображений в различные кластеры на основе векторов признаков, извлеченных из изображений.

[168] Затем способ 600 может быть завершен.

[169] В целом, второй генератор 400 обучающей выборки и способ 500 позволяют формировать кластеры из комплексно взвешенных признаков наиболее популярных (или всех) результатов поиска изображений, связанных с запросом, при этом каждый кластер может содержать наиболее схожие изображения с точки зрения их векторов признаков. Таким образом, обучающие объекты могут формироваться путем назначения определенной метки для части результатов поиска изображений из одного кластера.

(57) Формула изобретения

1. Способ формирования набора обучающих объектов для алгоритма машинного обучения (MLA), предназначенного для классификации изображений, выполняемый на сервере, осуществляющем MLA, и включающий в себя:

получение из журнала поиска данных о поисковых запросах, выполненных во время вертикального поиска изображений, каждый из которых связан с первым набором результатов поиска изображений;

формирование вектора запроса для каждого поискового запроса;

распределение векторов запросов между множеством кластеров векторов запросов;

связывание с каждым кластером векторов запросов второго набора результатов поиска изображений, содержащего по меньшей мере часть каждого первого набора результатов поиска изображений, связанных с векторами запросов, входящими в состав каждого соответствующего кластера векторов запросов; и

формирование набора обучающих объектов путем сохранения для каждого кластера векторов запросов каждого результата поиска изображений из второго набора результатов поиска изображений в виде обучающего объекта в наборе обучающих объектов, при этом каждый результат поиска изображений связан с меткой кластера, указывающей на кластер векторов запросов, с которым связан результат поиска изображений.

2. Способ по п. 1, отличающийся тем, что формирование вектора запроса включает в себя применение алгоритма векторизации слов для каждого поискового запроса.

3. Способ по п. 2, отличающийся тем, что перед связыванием второго набора результатов поиска изображений с каждым кластером векторов запросов способ дополнительно включает в себя получение для каждого первого набора результатов поиска изображений соответствующего набора метрик, каждая из которых указывает на действия пользователя с соответствующим результатом поиска изображений из первого набора результатов поиска изображений;

при этом связывание с каждым кластером векторов запросов второго набора результатов поиска изображений включает в себя выбор по меньшей мере части каждого первого набора результатов поиска изображений, входящих в состав второго набора результатов поиска изображений, на основе превышающих заранее заданный порог соответствующих метрик результатов поиска изображений из первого набора результатов поиска изображений.

4. Способ по п. 3, отличающийся тем, что кластеры векторов запросов формируются на основе степени близости векторов запросов в N-мерном пространстве.

5. Способ по п. 2, отличающийся тем, что используется один из следующих алгоритмов векторизации слов: word2vec, GloVe (глобальные векторы для представления слов), LDA2Vec, sense2vec и wang2vec.

6. Способ по п. 1, отличающийся тем, что кластеризация осуществляется с использованием одного из следующих алгоритмов: кластеризация методом k-средних, кластеризация методом максимизации ожиданий, кластеризация методом максимальной удаленности, иерархическая кластеризация, кластеризация методом cobweb и кластеризация на основе плотности.

7. Способ по п. 1, отличающийся тем, что каждый результат поиска изображений из первого набора результатов поиска изображений связан с соответствующей метрикой, указывающей на действия пользователя с результатом поиска изображений, а формирование вектора запроса включает в себя:

формирование вектора признаков для каждого результата поиска изображений из выбранного подмножества результатов поиска изображений, связанных с поисковым запросом;

взвешивание каждого вектора признаков с использованием соответствующей метрики; и

объединение векторов признаков, взвешенных с использованием соответствующих метрик.

8. Способ по п. 7, отличающийся тем, что перед формированием вектора признаков для каждого результата поиска изображений из выбранного подмножества результатов поиска изображений способ дополнительно включает в себя выбор по меньшей мере части каждого первого набора результатов поиска изображений, входящих в состав выбранного подмножества результатов поиска изображений, на основе превышающих заранее заданный порог соответствующих метрик результатов поиска изображений из первого набора результатов поиска изображений.

9. Способ по п. 8, отличающийся тем, что второй набор результатов поиска изображений включает в себя все результаты поиска изображений из первого набора результатов поиска изображений, связанных с векторами запросов, входящими в состав каждого соответствующего кластера.

10. Способ по п. 7, отличающийся тем, что соответствующая метрика представляет собой коэффициент переходов (CTR) или количество переходов.

11. Способ по п. 9, отличающийся тем, что кластеризация осуществляется с использованием одного из следующих алгоритмов: кластеризация методом k-средних, кластеризация методом максимизации ожиданий, кластеризация методом максимальной удаленности, иерархическая кластеризация, кластеризация методом cobweb и кластеризация на основе плотности.

12. Способ обучения алгоритма машинного обучения (MLA), предназначенного для классификации изображений, выполняемый на сервере, осуществляющем MLA, и включающий в себя:

получение из журнала поиска данных о поисковых запросах, выполненных во время вертикального поиска изображений, каждый из которых связан с первым набором результатов поиска изображений, при этом каждый результат поиска изображений связан с соответствующей метрикой, указывающей на действия пользователя с результатом поиска изображений;

выбор для каждого поискового запроса результатов поиска изображений из первого набора результатов поиска изображений, имеющих соответствующую метрику, превышающую заранее заданный порог, для добавления в соответствующее выбранное подмножество результатов поиска изображений;

формирование вектора признаков для каждого результата поиска изображений из соответствующего выбранного подмножества результатов поиска изображений, связанных с каждым поисковым запросом;

формирование вектора запроса для каждого поискового запроса на основе векторов признаков и соответствующих метрик результатов поиска изображений из соответствующего выбранного подмножества результатов поиска изображений;

распределение векторов запросов между множеством кластеров векторов запросов;

связывание с каждым кластером векторов запросов второго набора результатов поиска изображений, включающего в себя соответствующие выбранные подмножества результатов поиска изображений, связанных с векторами запросов, входящими в состав каждого соответствующего кластера векторов запросов;

формирование набора обучающих объектов путем сохранения для каждого кластера векторов запросов каждого результата поиска изображений из второго набора результатов поиска изображений в виде обучающего объекта в наборе обучающих объектов, при этом каждый результат поиска изображений связан с меткой кластера, указывающей на кластер векторов запросов, с которым связан результат поиска изображений; и

обучение MLA для классификации изображений с использованием сохраненного набора обучающих объектов.

13. Способ по п. 12, отличающийся тем, что обучение представляет собой первый этап обучения с целью грубого обучения MLA для классификации изображений.

14. Способ по п. 13, отличающийся тем, что дополнительно включает в себя точное обучение MLA с использованием дополнительного набора точно настроенных обучающих объектов.

15. Способ по п. 12, отличающийся тем, что MLA представляет собой алгоритм обучения искусственной нейронной сети (ANN).

16. Способ по п. 15, отличающийся тем, что MLA представляет собой алгоритм глубинного обучения.

17. Система формирования набора обучающих объектов для алгоритма машинного обучения (MLA), предназначенного для классификации изображений, содержащая физический машиночитаемый носитель информации, содержащий команды, и процессор, выполняющий эти команды и выполненный с возможностью:

получать из журнала поиска данные поисковых запросов, выполненных во время вертикального поиска изображений, каждый из которых связан с первым набором результатов поиска изображений;

формировать вектор запроса для каждого поискового запроса;

распределять векторы запросов между множеством кластеров векторов запросов;

связывать с каждым кластером векторов запросов второй набор результатов поиска изображений, содержащий по меньшей мере часть каждого первого набора результатов

поиска изображений, связанных с векторами запросов, входящими в состав каждого соответствующего кластера векторов запросов; и

формировать набор обучающих объектов путем сохранения для каждого кластера векторов запросов каждого результата поиска изображений из второго набора результатов поиска изображений в виде обучающего объекта в наборе обучающих объектов, при этом каждый результат поиска изображений связан с меткой кластера, указывающей на кластер векторов запросов, с которым связан результат поиска изображений.

18. Система по п. 17, отличающаяся тем, что каждый результат поиска изображений из первого набора результатов поиска изображений связан с соответствующей метрикой, указывающей на действия пользователя с результатом поиска изображений, а для формирования вектора запроса процессор выполнен с возможностью:

формировать вектор признаков для каждого результата поиска изображений из выбранного подмножества результатов поиска изображений, связанных с поисковым запросом;

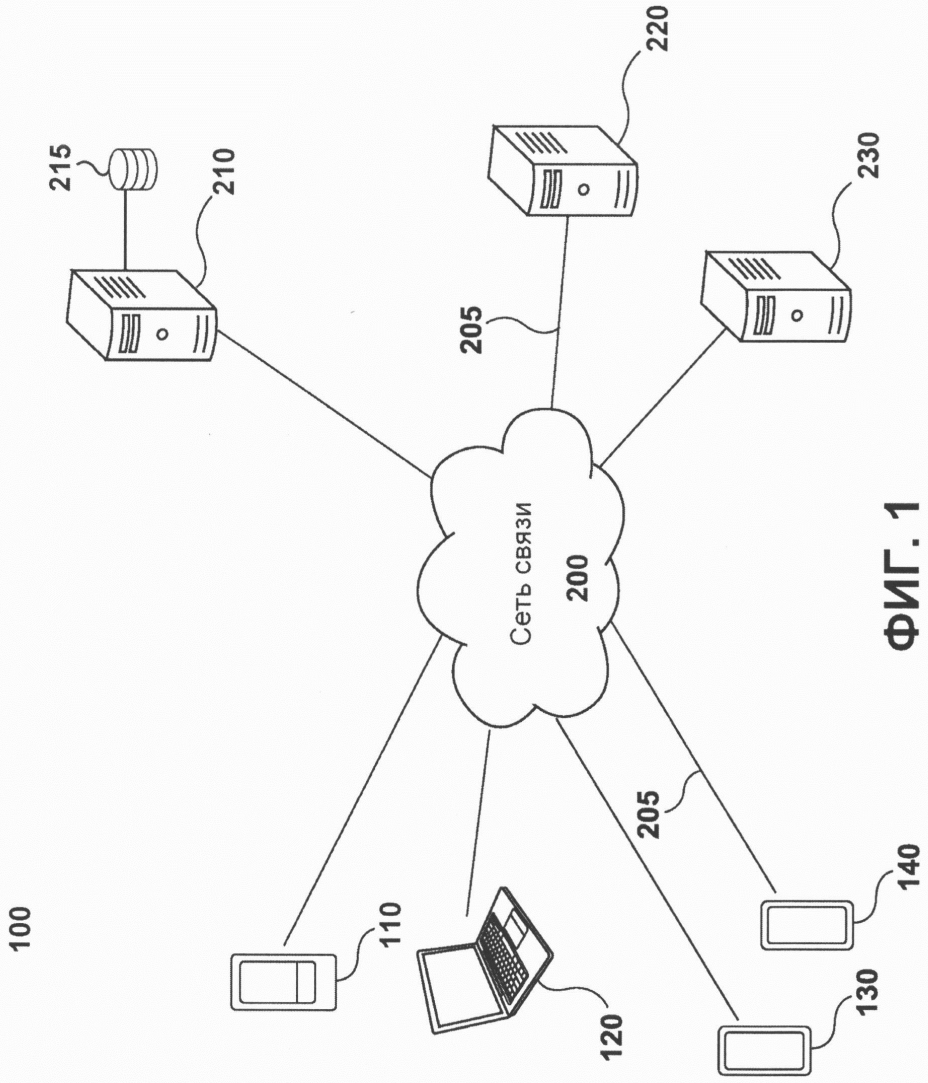
взвешивать каждый вектор признаков с использованием соответствующей метрики; и

объединять векторы признаков, взвешенные с использованием соответствующих метрик.

19. Система по п. 18, отличающаяся тем, что перед формированием вектора признаков для каждого результата поиска изображений из выбранного подмножества результатов поиска изображений процессор дополнительно выполнен с возможностью выбирать по меньшей мере часть каждого первого набора результатов поиска изображений, входящих в состав выбранного подмножества результатов поиска изображений, на основе превышающих заранее заданный порог соответствующих метрик результатов поиска изображений из первого набора результатов поиска изображений.

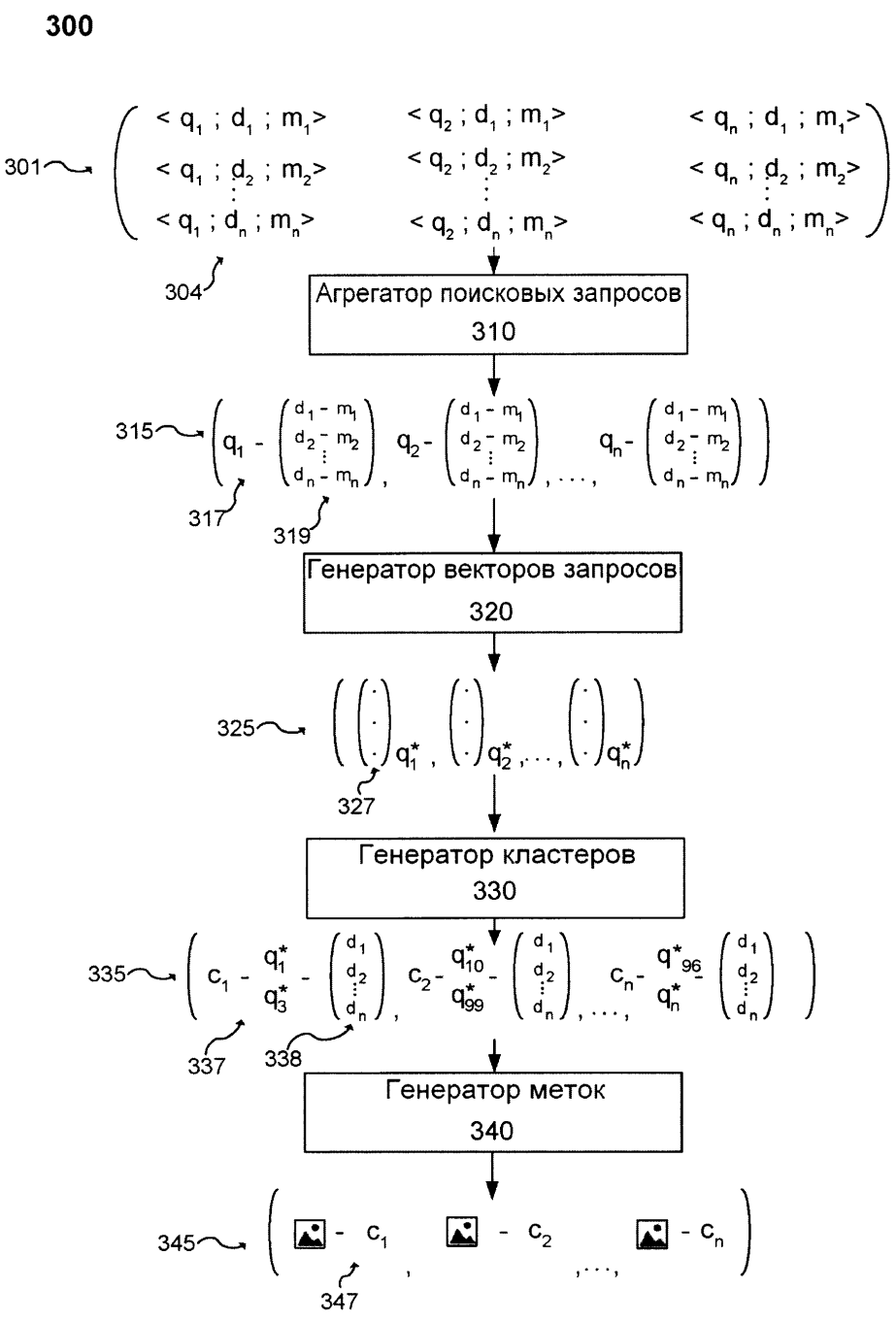
20. Система по п. 19, отличающаяся тем, что второй набор результатов поиска изображений включает в себя все результаты поиска изображений из первого набора результатов поиска изображений, связанных с векторами запросов, входящими в состав каждого соответствующего кластера.

1

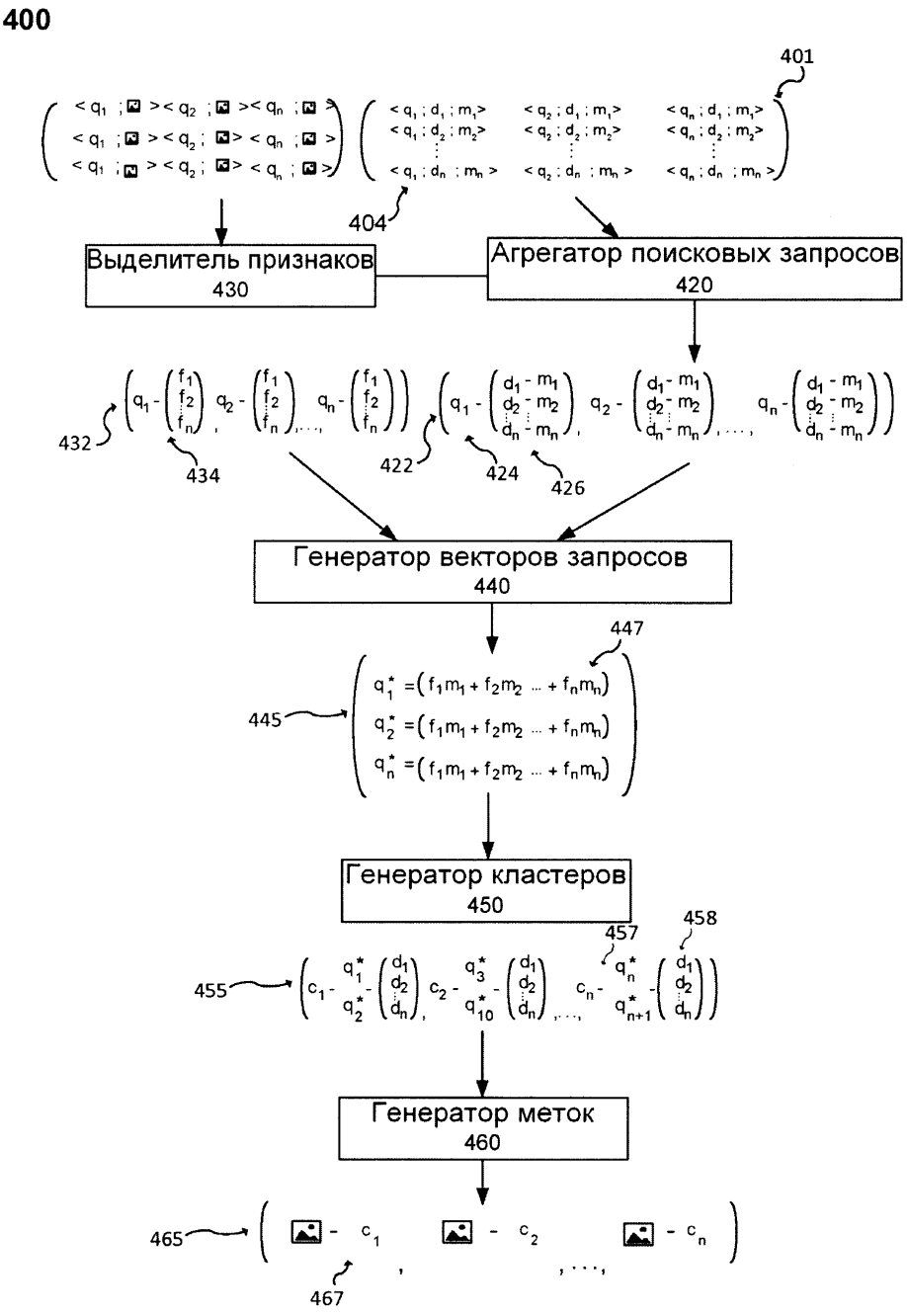


ФИГ. 1

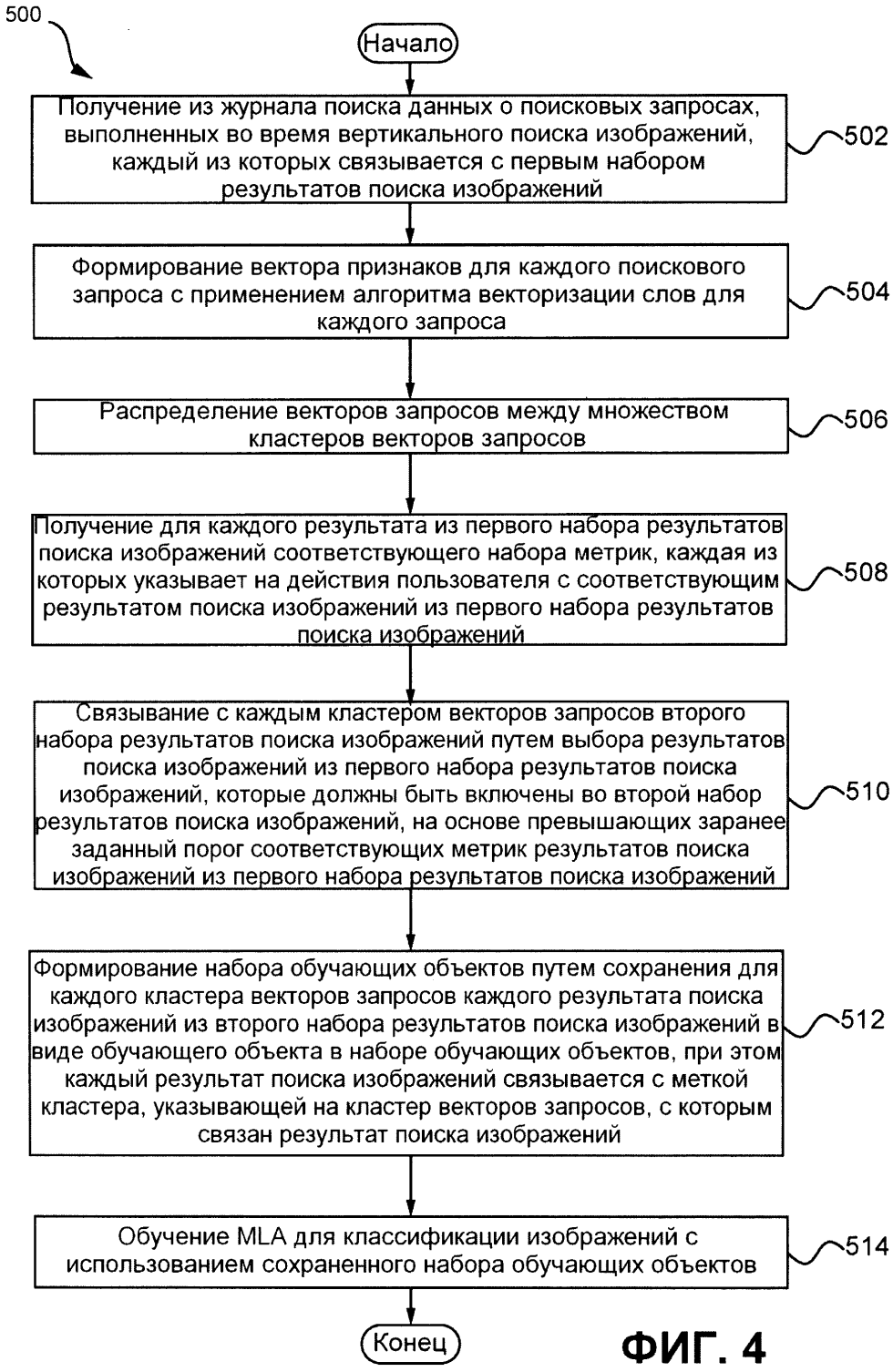
2


300
301
304
Агрегатор поисковых запросов
310
315
317
319
Генератор векторов запросов
320
325
327
Генератор кластеров
330
335
337
338
Генератор меток
340
345
347

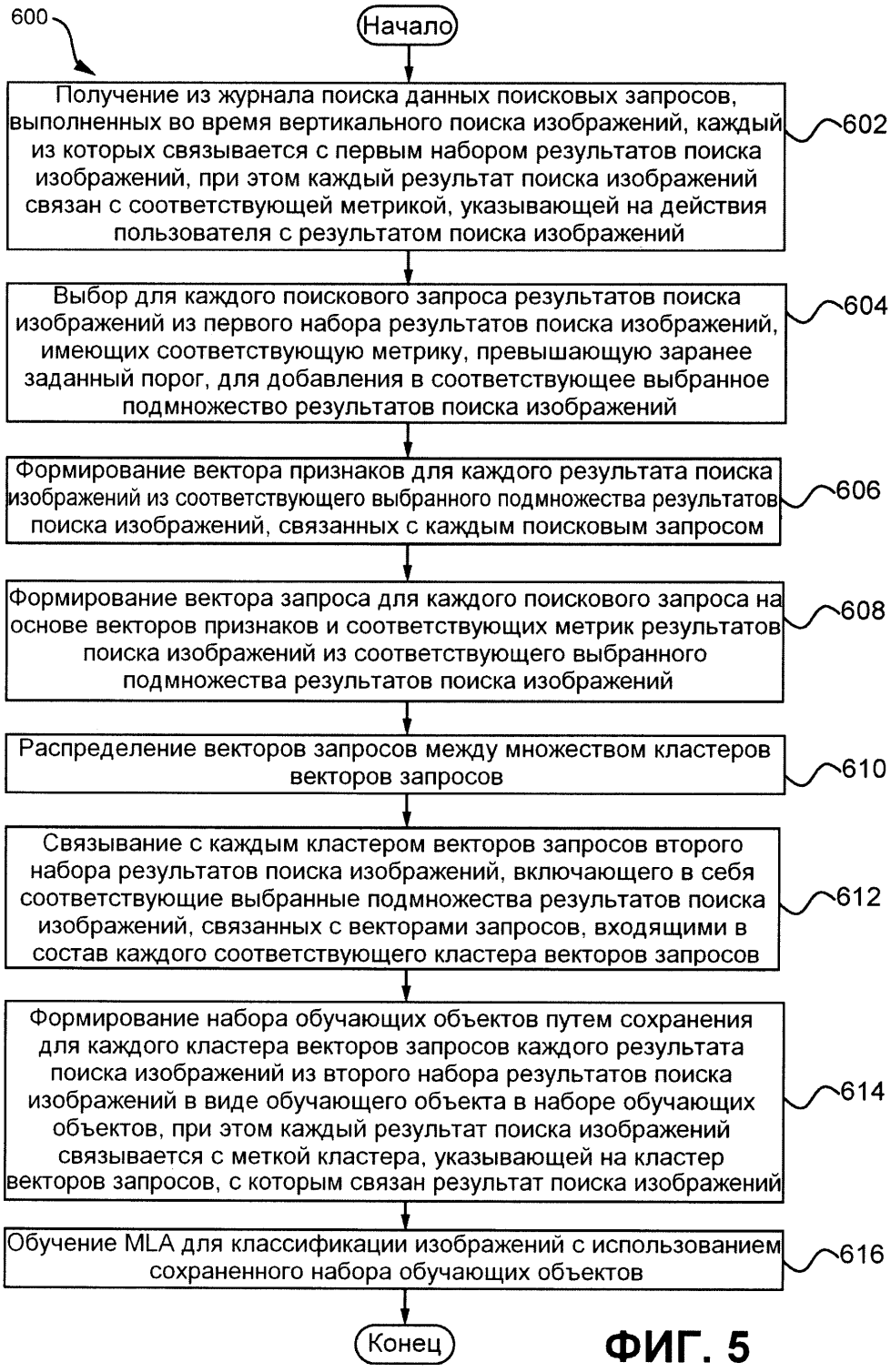
ФИГ. 2


400
401
404
Выделитель признаков
430
Агрегатор поисковых запросов
420
432
434
422
424
426
Генератор векторов запросов
440
447
445
Генератор кластеров
450
458
457
455
Генератор меток
460
465
467

ФИГ. 3


500
Начало
Получение из журнала поиска данных о поисковых запросах, выполненных во время вертикального поиска изображений, каждый из которых связывается с первым набором результатов поиска изображений
502
Формирование вектора признаков для каждого поискового запроса с применением алгоритма векторизации слов для каждого запроса
504
Распределение векторов запросов между множеством кластеров векторов запросов
506
Получение для каждого результата из первого набора результатов поиска изображений соответствующего набора метрик, каждая из которых указывает на действия пользователя с соответствующим результатом поиска изображений из первого набора результатов поиска изображений
508
Связывание с каждым кластером векторов запросов второго набора результатов поиска изображений путем выбора результатов поиска изображений из первого набора результатов поиска изображений, которые должны быть включены во второй набор результатов поиска изображений, на основе превышающих заранее заданный порог соответствующих метрик результатов поиска изображений из первого набора результатов поиска изображений
510
Формирование набора обучающих объектов путем сохранения для каждого кластера векторов запросов каждого результата поиска изображений из второго набора результатов поиска изображений в виде обучающего объекта в наборе обучающих объектов, при этом каждый результат поиска изображений связывается с меткой кластера, указывающей на кластер векторов запросов, с которым связан результат поиска изображений
512
Обучение MLA для классификации изображений с использованием сохраненного набора обучающих объектов
514
Конец

ФИГ. 4


600
Начало
Получение из журнала поиска данных поисковых запросов, выполненных во время вертикального поиска изображений, каждый из которых связывается с первым набором результатов поиска изображений, при этом каждый результат поиска изображений связан с соответствующей метрикой, указывающей на действия пользователя с результатом поиска изображений
602
Выбор для каждого поискового запроса результатов поиска изображений из первого набора результатов поиска изображений, имеющих соответствующую метрику, превышающую заранее заданный порог, для добавления в соответствующее выбранное подмножество результатов поиска изображений
604
Формирование вектора признаков для каждого результата поиска изображений из соответствующего выбранного подмножества результатов поиска изображений, связанных с каждым поисковым запросом
606
Формирование вектора запроса для каждого поискового запроса на основе векторов признаков и соответствующих метрик результатов поиска изображений из соответствующего выбранного подмножества результатов поиска изображений
608
Распределение векторов запросов между множеством кластеров векторов запросов
610
Связывание с каждым кластером векторов запросов второго набора результатов поиска изображений, включающего в себя соответствующие выбранные подмножества результатов поиска изображений, связанных с векторами запросов, входящими в состав каждого соответствующего кластера векторов запросов
612
Формирование набора обучающих объектов путем сохранения для каждого кластера векторов запросов каждого результата поиска изображений из второго набора результатов поиска изображений в виде обучающего объекта в наборе обучающих объектов, при этом каждый результат поиска изображений связывается с меткой кластера, указывающей на кластер векторов запросов, с которым связан результат поиска изображений
614
Обучение MLA для классификации изображений с использованием сохраненного набора обучающих объектов
616
Конец

ФИГ. 5