(19) Евразийское патентное ведомство

(21) 202193231 (13) A1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОЙ ЗАЯВКЕ

(43) Дата публикации заявки
2023.04.05

(22) Дата подачи заявки
2021.12.22

(51) Int. Cl. ***G06F 16/20*** (2019.01)
G06F 16/33 (2019.01)
G06F 40/253 (2020.01)
G06N 3/02 (2006.01)
G06N 20/00 (2019.01)

(54) РЕКОМЕНДАЦИИ ВАКАНСИЙ НА МАШИННОМ ОБУЧЕНИИ С ПРИМЕНЕНИЕМ RNN DSSM НАД SUBWORD'АМИ ДЛЯ ОЦЕНКИ ВЕРОЯТНОСТИ ОТКЛИКА

(96) **2021000141 (RU) 2021.12.22**

(71) Заявитель:
ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "ХЭДХАНТЕР" (RU)

(72) Изобретатель:
Сидоров Александр Алексеевич, Реушкин Виктор Викторович, Яркин Станислав Викторович, Даньщин Георгий Андреевич (RU)

(74) Представитель:
Киселев А.Е. (RU)

(57) Изобретение характеризуется тем, что способ автоматизированного поиска персонала путем сравнения описаний резюме и вакансий заключается в том, что из каждого описания выделяют три блока признаков, производят раздельное сравнение соответствующих блоков сравниваемых пар описаний, соответствие описаний пар определяют по результатам сравнения всех блоков, где в объеме первого блока выделяют исчисляемые параметры, сочетаниям которых для каждого из документов присваивают промежуточные векторные представления исчисляемых параметров, а в качестве результата сравнения, определяющего степень соответствия, используют скалярные произведения промежуточных векторных представлений исчисляемых параметров; в объеме второго блока выделяют семантические составляющие описаний, определяют поисковые вектора для семантических составляющих и производят определение соответствия описаний путем сравнения векторов и вычисления вероятности соответствия семантических составляющих; в объеме третьего блока для каждого из описаний выделяют семантические структуры описаний, в которых существенным является порядок следования грамматических форм, и производят определение вероятности соответствия указанных структур сравниваемых описаний, при этом при сравнении вторых блоков не учитывают семантические составляющие описаний, для которых установлено несоответствие семантических структур описаний, соответствующих соответствующим семантическим составляющим, а для выделения семантических структур, в которых существенным является порядок следования грамматических форм, используют рекуррентную нейронную сеть, при этом производят обучение рекуррентной нейронной сети по парам резюме-вакансия, для которых определено соответствие между описаниями и парам резюме-вакансия, для которых установлено несоответствие между описаниями.

202193231 A1

202193231 A1

РЕКОМЕНДАЦИИ ВАКАНСИЙ НА МАШИННОМ ОБУЧЕНИИ С ПРИМЕНЕНИЕМ RNN DSSM НАД SUBWORD'АМИ ДЛЯ ОЦЕНКИ ВЕРОЯТНОСТИ ОТКЛИКА

ОБЛАСТЬ ТЕХНИКИ

Предложенное изобретение относится к поисковым системам и может быть использовано для поиска вакансий и резюме в рекомендательных системах подбора персонала с большим объемом записей в базах данных и вакансиях, где присутствует большое количество сложных смысловых конструкций, состоящих из множества слов и предложений, меняющих или теряющих смысл при изменении порядка слов. Также способ может быть использован для сортировки вакансий и резюме по заданным критериям сходства. Изобретение относится к поисковым и рекомендательным системам и может быть использовано для поиска и рекомендаций вакансий и резюме в системах подбора персонала, в резюме и вакансиях которых большое количество сложных смысловых конструкций, состоящих из множества слов и предложений, и меняющих или теряющих смысл при изменении порядка слов.

УРОВЕНЬ ТЕХНИКИ

Одной из проблем, возникающей при поиске вакансий, соответствующих резюме, является большое количество документов, выдаваемых по поисковым запросам, что обусловлено, например, поиском сотрудников одновременно большим количеством работодателей, некоторые из которых, крупные компании, размещают множество таких объявлений в различных географических регионах, а также размещением объявлений о поиске работы трудоустроенными соискателями, не остро нуждающимися в работе, но рассматривающими возможность смены работы на приемлемых условиях. Большое количество однотипных результатов не позволяет соискателям принять решение о выборе вакансии с минимальными затратами времени, в связи с чем, имеется потребность в экспертных системах, снижающих время поиска работы соискателями, и, как следствие, время простоя рабочих мест у работодателей при выборе сотрудников, а также обеспечивающих соискателей полной и достоверной информациях об открытых вакансиях, при большом количестве предложений.

При найме и поиске работы, без применения средств автоматизации, работодатели и соискатели используют собственные аналитические способности для определения соответствия найденных документов собственным потребностям, при этом часть

признаков, которые используются для определения соответствия, изложены в нечеткой форме, например, в форме расстановки слов в предложении.

В настоящее время отсутствуют автоматизированные средства и методы, обеспечивающие возможность анализа текстов в полном соответствии с интеллектом человека, используемые системы подбора резюме и вакансий в рекомендательных и поисковых системах в области найма применяют машинное обучение, чтобы прогнозировать поведение соискателей и работодателей для определения вероятности отклика и приглашения, и на основании этого отбирать какие резюме и вакансии показывать работодателям и соискателям. Т.е. модели в машинном обучении должны учитывать те же признаки, которые учитывает человек при поиске требуемых документов.

Для использования машинного обучения, тексты преобразуются в вектора чисел. В настоящее время наиболее часто применяется стемминг, совместно с BOW и TF/IDF векторизацией.

Стемминг и BOW по n-граммам, а также TF/IDF векторизация не может использовать учет порядка слов в словосочетаниях длиннее 2-3 слов из-за того, что количество возможных перестановок с увеличением размерности растёт комбинаторно и требует такого количества оперативной памяти, закупка и установка которого не окупается экономическим эффектом от получаемого прироста качества выдачи.

Поэтому в большинстве современных систем считают, например, что вектора резюме «без необходимости холодных продаж» и вакансии «необходимо иметь опыт в холодных продажах» будут близки, в результате чего качество сравнения учёта этой особенности работы менеджера по продажам, а также многих других особенностей и многих других профессий не учитывается должным образом при проведении поиска.

Одной из экспертных систем, наиболее близкой к предложенному изобретению, является система, описанная в патентной заявке США US2018173803 (A1), опубликованной 21.06.2018. В известном решении используется машинное обучение для сжатия векторов описаний резюме и вакансий на основании семантического сходства векторов. При проведении машинного обучения используется определение коэффициентов сходства векторов между различными областями, что подразумевает смысловой анализ данных, представленных для обработки. Недостатком известного решения является необходимость учета семантики поисковых запросов и индексируемых данных, что предполагает необходимость предварительной сортировки данных по отраслям. В связи с тем, что часть вакансий не может быть объективно отнесена к определенным отраслям, для проведения машинного обучения требуется обработка данных оператором. Таким образом, эффективность использования известного технического решения зависит от субъективных

качеств оператора или операторов, подготавливающих исходные данные для настройки системы, процесс поиска не является полностью формализованным.

Обработка вакансий и правильное их распределение по отраслям требует большого количества трудозатрат, преимущественно, практикующих специалистов, отлично знающих предметную область.

СУЩНОСТЬ ИЗОБРЕТЕНИЯ

Предложенное решение решает задачу создания автоматизированной экспертной системы подбора персонала. Техническим результатом, достигаемым при использовании изобретения, является сокращение трудозатрат на подготовку материалов, необходимых для машинного обучения, сокращение объемов машинных вычислений, ускорение получения результатов, сокращение времени на анализ результатов поисков и, главным образом, повышение точности при определении соответствия между записями резюме и вакансий.

Одной из предпосылок создания изобретения является необходимость экспертной системы с машинным обучением, минимально использующей вмешательство операторов при машинном обучении системы и ее настройках.

Базы данных рекомендательной системы могут разделяться логически и физически на отдельные структуры, в соответствии с назначением данных, хранящихся в отдельных структурах.

Для управления, создания и использования баз данных могут использоваться различные системы управления базами данных (СУБД). В частных случаях, для хранения документов и перечисленных ниже результатов их обработки, могут использоваться постоянные хранилища данных, а для хранения данных, необходимых для обработки одного поискового запроса, могут использоваться оперативные запоминающие устройства. Специфика обработки больших объемов информации требует, чтобы для достижения быстродействия, при котором использования предложенной и аналогичных систем становилось целесообразным, в процессе работы применялись только оперативные запоминающие устройства или ОЗУ (RAM), с асинхронным сохранением и, при необходимости, повторным считыванием из постоянного запоминающего устройства или ПЗУ, выполненного, например, в виде твердотельного диска, сетевого диска или других аналогичных блоков хранения данных, имеющих большую емкость, но относительно низкую скорость обмена данными. Предложенное изобретение, в данной части, позволяет обрабатывать заметно большие объемы данных с приемлемой скоростью, чем известные

решения, при том же объеме оперативных запоминающих устройств, чем известные решения.

Сравнение производительности используемых ведущими поисковыми системами технологий с предложенной показывает, что применение предложенной технологии дает преимущество по производительности серверов более, чем в 50 раз на поисковый запрос. При этом поисковые системы ведущих систем остаются более эффективными при обработке не специализированных поисковых запросов. Основной предпосылкой достижения технического результата является то, что формируется общий индекс для всего массива документов, поступивших в базу данных в то время, как предложенное изобретение формирует независимые индексы для резюме и вакансий, и при реализации изобретения не требуется определять сходство, например документов, характеризующих резюме, между собой. При этом в предложенном изобретении ускоряется ранжирование документов, соответствующих результатам поискового запроса, а при формировании запроса, соответствующего документу, уже поступившему в базу даны, не требуется предварительная обработка информации, соответствующей запросу.

Векторное представление резюме и вакансии может формироваться с использованием разбиения текста на subword'ы или семантически связанные наборы слов. В другом возможном варианте реализации последовательности subword'ов в тексте векторизуются с помощью нейросетей с архитектурой RNN. Вероятность отклика для пары документов «резюме-вакансия» может быть оценена с помощью нейросети DSSM, использующей на входе данные от нейросети RNN. Кроме того, общее сочетание DSSM и RNN обучается на примерах из обучающей выборки одновременно, с помощью одной back propagation. В известном решении также используется LM-слои нейронов, которые обучаются отдельно на задаче языкового моделирования по корпусу текстов резюме и вакансий. RNN-слои получают данные от LM-слоёв.

Для достижения технического результата предлагается способ автоматизированного поиска персонала путем сравнения описаний резюме и вакансий, заключающийся в том, что из каждого описания выделяют три блока признаков, производят раздельное сравнение соответствующих блоков сравниваемых пар описаний, соответствие описаний пар определяют по результатам сравнения всех блоков, где:

в объеме первого блока выделяют исчисляемые параметры, сочетаниям которых, для каждого из документов, присваивают промежуточные векторные представления исчисляемых параметров, а в качестве результата сравнения, определяющего степень соответствия, используют скалярные произведения промежуточных векторных представлений исчисляемых параметров;

в объеме второго блока выделяют семантические составляющие описаний, определяют поисковые вектора для семантических составляющих и производят определение соответствия описаний путем сравнения векторов и вычисления вероятности соответствия семантических составляющих;

в объеме третьего блока для каждого из описаний выделяют семантические структуры описаний, в которых существенным является порядок следования грамматических форм, и производят определение вероятности соответствия указанных структур сравниваемых описаний, при этом:

при сравнении вторых блоков не учитывают семантические составляющие описаний, для которых установлено несоответствие семантических структур описаний, соответствующих соответствующим семантическим составляющим, а:

для выделения семантических структур, в которых существенным является порядок следования грамматических форм, используют рекуррентную нейронную сеть, при этом производят обучение рекуррентной нейронной сети по парам резюме-вакансия, для которых определено соответствие между описаниями и парам резюме-вакансия, для которых установлено несоответствие между описаниями.

В частном случае реализации, для обучения рекуррентной нейронной сети используют пары резюме-вакансия, максимально соответствующие друг другу, где степень соответствия определяется с использованием действий соискателей и работодателей, разместивших резюме и вакансии, в том числе, откликов и приглашений на собеседования, где первый и второй блоки сгруппированы в виде слоев нейронов предварительно обученной нейронной сети, при этом, для обучения рекуррентной нейронной сети, в частном случае, используют 4-6% пар резюме-вакансия, где пары резюме-вакансия, равномерно распределенные по датам составления за календарный год. Векторные представления исчисляемых параметров могут быть сформированы путем анализа текста документов с использованием слоев нейронной сети.

Между парами документов может фиксироваться взаимодействие, по результатам использования документов пользователями, например, отклик, просмотр контактов вакансии, приглашение после отклика. В результате обучения формируются такие параметры нейронов в нейросети, что прогноз приглашения соискателя с определённым резюме на определённую вакансию считается эффективным при применении к векторам, полученным из значений на последних слоях нейронов для заданных резюме и вакансии, при выполнении операции вычисления скалярного произведения (dot product). В случае если количество документов недостаточно для проведения машинного обучения, временной период может быть увеличен, при этом для обучения могут быть выбраны и

использованы пары резюме-вакансия, равномерно распределенные по датам составления за календарный год или другой заранее выбранный период.

Документы, представляющие собой описание резюме или вакансии имеют уникальный идентификационный признак документа, а каждый из документов может содержать по крайней мере одно индексируемое поле, содержащее сведения, характеризующее документ с использованием терминов естественного языка, при этом каждое из индексируемых полей имеет уникальный идентификационный признак поля;

Ветки нейросети, обрабатывающие категориальные и числовые признаки резюме и вакансий имеют архитектуру многослойных перцептронов. В них веса у параметров резюме и вакансии есть только на первом слое нейронов. Во втором слое нейронов – упрощённо, веса уже не параметров, а нейронов из первого слоя, причём частично выученные, а частично искусственно занулённые (такое зануление, которое называется dropout, нужно, чтобы предотвратить переобучение, overfitting).

Основное преимущество, которое дает использование DSSM заключается в том, что, что прогноз отклика и/или приглашения соискателя с определённым резюме на определённую вакансию можно определить, применив к результирующим векторам операцию вычисления dot product (скалярного произведения).

Рабочая база данных документов обновляется путем удаления устаревших документов и добавления новых документов; а обучающая база данных документов, как показано выше, формируется из документов, для которых получены отклики соискателей на вакансии и/или приглашения соискателей на вакансии работодателями, для которых может быть определена релевантность документов поисковым пользовательским запросам. Далее поисковая база данных может обновляться путем добавления документов, для которых получены новые отклики соискателей на вакансии и/или приглашения соискателей на вакансии работодателями, которые могут быть использованы для определения степени соответствия документов пользовательским запросам.

В одном из частных вариантов реализации производят предварительную векторизацию семантических структур естественного языка с помощью слоёв нейронов, обученных отдельно, на задаче языкового моделирования, по корпусу текстов резюме и вакансий (LM-слоёв). При этом, слои LM могут представлять собой дополнительно выделенную RNN.

В частном случае, построение нейросети осуществляется, последовательным обучением: сначала её части обучаются на задаче языкового моделирования, а потом всё вместе обучается одновременно на задачах прогнозирования вероятности приглашения, зарплаты в резюме и в вакансии.

КРАТКОЕ ОПИСАНИЕ ГРАФИЧЕСКИХ МАТЕРИАЛОВ

На фиг. 1 показан пример компьютерной системы общего назначения, которая может быть использована при создании и конфигурировании отдельных элементов программно-аппаратного комплекса, используемого в системе предназначенной для использования изобретения, например, серверов обработки и хранения данных, а также модулей системы.

На фиг. 2 показана упрощенная последовательность шагов по формированию поискового вектора документа, в качестве которого может использоваться описание резюме или вакансии.

На фиг. 3 представлена последовательность операций по предварительному анализу документов, а также для формирования и настройки программно-аппаратного комплекса, непосредственного предназначенного для поиска соответствия между резюме и вакансиями.

На фиг. 4, обобщенно показана структура программно-аппаратного комплекса, реализующего нейронную сеть, предназначенную для индексации вакансий, поступивших в базу данных.

На фиг. 5, иллюстративно показан пример использования предварительно настроенной системы для поиска вакансий, соответствующих резюме пользователя, разместившего запрос.

На фиг. 6 показана структура базы данных, хранящей индексы резюме и вакансий.

На фиг. 7, иллюстративно показано использование функции хэширования для преобразования совокупности векторов документов и другой значащей информации.

На фиг. 8 представлена схематически последовательность операций по обработке пользовательского запроса.

Как показано на Фиг. 2, при формировании поискового вектора 232 документа, где в качестве документа может использоваться описание резюме или вакансии общее описание 203, содержащее заголовок и текст вакансии, разбивается на разделы 201 и 204, подлежащие независимой обработке. Раздел 204 анализируется как текстовая последовательность, представленная в машиночитаемом формате, а раздел 201 анализируется в части полей, имеющих самостоятельное смысловое значение с формированием 212 блоков данных 230. Как показано на фиг. 2, из текста документа могут быть выделены текстовые описания, характеризующие отрасль экономики 213, профессиональный опыт 220, специализацию 214, и профессиональную область 215. При этом данные, характеризующие профессиональный опыт могут быть разбиты по категориям должность (221), опыт работы (222) и режим работы (223). Для сокращения

объема вычислений, указанные описания упорядочиваются заранее заданным способом, с формированием совокупности последовательных блоков 230. Соответствующие данные объединяются и форматируются 209, с формированием одного из поисковых векторов 231, после чего данные уплотняются или хэшируются 210 таким образом, чтобы одинаковые наборы данных имели идентичные значения хэшей. Исчисляемые параметры 250 используются в виде представлений, соответствующих значениям исчисляемых параметров для добавления в поисковый вектор, формируемый слиянием двух потоков данных 206 и 211. Площади каждого из прямоугольников, характеризующих данные 230 и вектора 231, 232, примерно соответствуют объему данных или количеству ячеек для хранения данных, представляющих каждый из указанных объектов.

Заранее заданные содержательные части описания вакансии, в том числе, название вакансии или резюме, а также отдельные предложения или значащие фразы из описания вакансии анализируются 205 в виде последовательного набора символов без разделения на разделы с использованием рекуррентных нейронных сетей 204 (RNN) с формированием поисковых векторов. Наборы слов, используемых в описаниях также используются для построения дополнительных поисковых векторов, с использованием алгоритмов BOW, а также совместно с векторизацией, например, совместно с TF/IDF векторизацией, где под BOW (мешок слов) понимается простой анализ слов в тексте описания, а TF и IDF обозначают анализ частоты использования слов в документе и инверсии частот, с которыми эти слова встречается в документах из корпуса текстов, соответственно. Таким образом для документа формируется BOW (bag of words или множество уни- и биграмм из лемм слов и коротких словосочетаний в документе), после чего формируются вектора путем векторизациии с помощью TF/IDF.

BOW может обрабатываться не только с использованием TF/IDF-векторизации, но и PPMI-векторизации, в которой числа, получаемые от наличия уни- и биграмм в BOW зависят не от частоты, с которой они встречаются в корпусе текстов, а от того, насколько часто были взаимодействия между такими текстами. Например, если соискателя со словом «ЛОР» в резюме приглашали на вакансию со словом «отоларинголог», то вес такой униграммы при сопоставлении конкретных текстов резюме и вакансии больше нуля, если не приглашают, то ноль.

Далее сформированные по каждому из методов поисковые вектора объединяются 206 и 211 между собой и дополняются исчислимыми данными 250, после чего уплотняются 207 и 208 таким образом, чтобы вектор RNN в дальнейшем имел преимущество, в результате чего формируется итоговый поисковый вектор 232, в котором учитываются результаты обработки ранее сформированных векторов всех потоков.

Благодаря свойствам RNN, в частности, применению механизма маркировки attention (внимание или существенный признак), нейросеть учитывает последовательности слов в словосочетаниях и целых предложений, например, до 300 слов. Благодаря тому, что обе части обучаются на одних примерах, вектора для документа со словами «без необходимости холодных продаж» и для документа со словами «необходимо иметь опыт в холодных продажах» в RNN-части оказываются далёкими, и DSSM-слои верно прогнозируют для них низкую вероятность отклика.

Нейросеть RNN DSSM (с улучшенной или углубленной моделью семантического сходства) обучается заранее, при машинном обучении. Вектора-полуфабрикаты, состоящие из значений на выходе предпоследнего слоя нейронов RNN DSSM нейросети для вакансий и резюме вычисляются для последующей индексации и складываются в индекс и в кеш соответствующих записей БД соответственно. При вычислении рекомендаций в реальном времени, они извлекаются и над ними производится всего одно вычисление, а именно, скалярное произведение или dot product, что позволяет экономить вычислительные ресурсы и при этом в 95% случаев укладываться в 50 мс по времени ответа.

Эти свойства позволили улучшить качества рекомендаций вакансий так, что позволили при той же аудитории соискателей получить дополнительно около 250 000 откликов на вакансии в день с 36000 приглашений на собеседование после откликов в день, что составляет более 1,3 млн. руб. в день по экономическому эффекту и увеличивает среднее количество откликов в день на соискателя примерно на 18%, что составляет больше полумиллиона рублей в день по экономическому эффекту.

Итоговые вектора сводятся в базу данных, в которой хранятся поисковые вектора и описания активных вакансий. Аналогичным образом, с использованием разбиения текста документа и векторизации, формируется база данных активных резюме.

При добавлении в базу данных нового описания резюме или вакансии, для соответствующего документа также формируется уникальный поисковый хэш, который может быть использован для поиска, непосредственно после составления и обработки документа. При отсутствии результатов, пользователь может исключить из описания отдельные параметры, фразы и термины для проведения уточняющего поиска.

Таким образом, в объеме первого блока выделяют исчисляемые параметры, каждому из которых присваивают весовой коэффициент, а результаты сравнения формируют в виде степени соответствия блоков;

в объеме второго блока выделяют семантические составляющие описаний, определяют поисковые вектора для семантических составляющих и производят

определение соответствия описаний путем сравнения векторов и вычисления вероятности соответствия семантических составляющих;

в объеме третьего блока для каждого из описаний выделяют семантические структуры описаний, в которых существенным является порядок следования грамматических форм, и производят определение вероятности соответствия указанных структур сравниваемых описаний, при этом:

при сравнении вторых блоков не учитывают семантические составляющие описаний, для которых установлено несоответствие семантических структур описаний, соответствующих соответствующим семантическим составляющим, а:

для выделения семантических структур, в которых существенным является порядок следования грамматических форм, используют рекуррентную нейронную сеть, при этом производят обучение рекуррентной нейронной сети по парам резюме-вакансия, для которых определено соответствие между описаниями и парам резюме-вакансия, для которых установлено несоответствие между описаниями. Соответствие или несоответствие определяется наличием или отсутствием в обучающем множестве пар резюме-вакансия, между которыми были отклики соискателей с резюме на вакансию и/или приглашения соискателей с такими резюме работодателями на собеседование на замещение соответствующей вакансии.

Векторизация текста для дальнейшего использования векторов нейросетью происходит следующим образом:

К тексту применяется словарь subword'ов, который строится по всему корпусу текстов резюме и вакансий алгоритмом BPE .

Дальше берутся первые 300 subword'ов, которые содержатся в тексте и в названии резюме и вакансии, и подаются на вход нейросети. Кроме этого, в перспективе, могут быть использованы эвристические подходы, основанные на машинном обучении, или гибридные подходы для выделения отдельных предложений, абзацев, смысловых блоков (таких как обязанности, требования, условия).

Поскольку пользователи иногда ошибаются с выбором разделов, предназначенных для определенной информации, или формируют описания (тело) резюме и вакансий без использования разделов, отдельные фразы, предложения или пункты списков относятся к разделам формализованных структур не только на основе эвристик, но и на основе отдельных или выделенных моделей машинного обучения, которые классифицируют фразы из документов.

В описываемом изобретении использованы программно-аппаратные логические структуры, которые позволяют выполнять операции по формированию рекомендаций

максимально эффективно. В качестве программно-аппаратных структур, при реализации изобретения, могут использоваться отдельные сервера с одним и более процессорами или совокупности серверов, используемые для хранения и обработки данных как по отдельности, так и в режим совместной работы, например, в виде кластера. Предложенное изобретение обладает масштабируемостью и возможностью управления конфигурацией используемого оборудования, в связи с чем, при реализации изобретения могут быть использованы несколько кластеров с балансировкой и eventual consistency между ними и внутри них, в трёх разных датацентрах, с несколькими десятками серверов в каждом из центров. При этом, отдельные сервисы внутри системы могут быть выполнены с использованием технологии виртуальных машин или контейнеров, где несколько контейнеров устанавливаются на одном физическом сервере.

Такая конфигурация гарантирует одновременно необходимую производительность, а также доступность, отказоустойчивость, управляемость, масштабируемость и простоту конфигурирования. Таким образом, в системе, на которой осуществлено изобретение, для реализации каждого из перечисленных ниже блоков и модулей используется от одного до нескольких десятков серверов, объединенных каналами передачи данных. Для реализации части блоков использовались виртуальные сервера и виртуальные машины, по функциональности аналогичные отдельным аппаратным серверам. В частности, в системе используется блок распределенного хранения данных, обеспечивающий распределение данных и согласование работы на удаленных друг от друга серверах.

В описываемом изобретении осуществляется формирование рабочей базы данных для документов, каждый из которых описывает резюме или вакансию, где каждый документ имеет уникальный идентификационный признак документа, и содержит по крайней мере одно индексируемое поле, содержащее сведения, характеризующее документ с использованием терминов естественного языка, при этом каждое из индексируемых полей имеет уникальный идентификационный признак поля.

В качестве идентификационного признака документа может использоваться уникальный идентификатор резюме или вакансии, при этом, идентификатор признака документа может содержать указание на пользователя, составившего документ. В частном случае, идентификатор может содержать адрес электронной почты или номер мобильного телефона пользователя или другой уникальный идентификатор, а также порядковый номер резюме или вакансии.

Каждый документ, включенный в рабочую базу данных, в общем случае, содержит одно поле, но в частных случаях может содержать несколько полей, каждое из которых содержит информацию, по крайней мере частично пригодную для определения

соответствия документа поисковому запросу. Под полем, в объеме изобретения, подразумевается совокупность поисковых признаков, по крайней мере часть из которых, учитывается при обработке поисковых запросов, как описано в рамках настоящего изобретения. Каждое поле также имеет уникальный идентификатор, который может быть образован идентификатором документа и номером поля в документе или назначением поля. Например, может быть указано, что поле является полем «образование» или «опыт работы».

Описываемыми в настоящем изобретении поисковыми запросами могут являться как поисковые строки, сформированные в режиме реального времени, так и сами документы. В частных случаях реализации изобретения, описание резюме или вакансии полностью или частично может использоваться в качестве совокупности данных, характеризующей поисковый запрос. Например, при поиске резюме, описание вакансии может быть использовано как поисковый запрос, и наоборот.

Изначально, пользователи заполняют электронные формы бланков резюме, которые, в общем случае, имеют по крайней мере одно поле, позволяющее присвоить резюме или вакансии уникальный идентификатор (ID – от англ. data name, identifier — опознаватель). Другие данные могут вводиться в произвольном или формализованном виде в одно или несколько дополнительных полей. Другими полями могут быть «опыт работы», «образование», «личные качества» и прочее. Поля с указанными названиями могут использоваться как для описания резюме, так и для описания вакансий. В связи с этим, если не указано иное, в системе производится обработка всех данных резюме и вакансий совместно.

При обработке поисковых запросов могут быть использованы описания или резюме, или вакансий, которые соответствуют тематике запроса. Тематика запроса, относящаяся к поиску резюме или к поиску вакансий может быть указана явно в названии или специальном поле, предназначенного для описания тематики, либо может следовать из совокупности сведений, указанных в поисковом выражении.

В рамках реализации настоящего изобретения может осуществляется обновление рабочей базы данных для документов (рабочей базы данных документов) путем удаления устаревших документов и добавления новых документов. Обновление рабочей базы данных может осуществляться на основании прямых указаний пользователей, на удаление и добавление вакансий и резюме в базу данных. Также могут быть удалены из базы данных документы, для которых истек срок действия. При этом удаленный из базы данных документы, для которых были получены отклики, могут продолжать использоваться для обучения системы или отдельных ее элементов.

В рамках реализации настоящего изобретения осуществляется формирование обучающей базы данных, содержащей документы, для которых получены отзывы пользователей, а именно, отклики соискателей с определёнными резюме на вакансии и/или приглашения соискателей с резюме на собеседование. Отзывы пользователей характеризуют релевантность документов поисковым пользовательским запросам. Для хранения обучающей базы данных используется блок хранения данных с высокой надежностью, поскольку утраченные сведения, относящиеся к документам, удаленным из рабочей базы данных, не могут быть восстановлены. Блок хранения данных с высокой надежностью обеспечивает отказоустойчивое хранение наиболее важных данных в реляционной структуре, поддержку их целостности и оперативный доступ к данным. Предпочтительно использовать блок хранения данных с высокой надежностью для хранения документов в исходном формате, например, в представлении на естественном языке.

Система, реализующая настоящее изобретение, может содержать модуль хранения обработанных данных и модуль хранения обрабатываемых данных. Обработанные данные делятся по категориям и распределяются для хранения в соответствии с требуемой категорией надежности хранения. Основное требование к модулю хранения обработанных данных – надежность хранения и доступность данных для локальных служб. Для этого может использоваться распределённое хранение в соответствии с признаком «расположение», соответствующему расположению вакансии или желаемому месту работы, указанному в резюме. В другом примере реализации, используется шардирование по хешу. Это увеличивает производительность и простоту эксплуатации, т.к. шарды получаются очень похожего объёма, в отличие от шардов, организованных по географическому признаку.

Также в системе может использоваться модуль интерфейса, который позволяет конечным пользователям и персоналу, обслуживающему функциональность системы (операторам), взаимодействовать с системой с помощью конечных устройств доступа, с использованием веб-браузеров, мобильных приложений и других сторонних систем, используемых для обмена данными, визуализации результатов обмена данными и для обслуживания устройство ввода информации.

Дополнительно может использоваться модуль формирования подсказок в поисковой строке, используемый в случае, если поисковое выражение для поисковой строки формируется в режиме реального времени, а также модуль исправления опечаток, который может использоваться, как в процессе ввода данных пользователем, так и для

предварительной обработки документов перед нормализацией теста документов и индексацией.

Обновление поисковой базы данных осуществляется путем добавления в базу данных документов, для которых получены новые отзывы, например, отклики и приглашения пользователей, указывающие на соответствие документов пользовательским запросам. Документы из обучающей базы данных, как правило, не удаляются, но могут быть удалены, например, в связи с потерей профессией актуальности на рынке.

Для обучения предложенной системы. а также для последующего проведения поиска, может применяться индексирование путем преобразования сведений индексируемого поля в индексную табличную строку. Каждая из ячеек индексной табличной строки соответствует наличию или отсутствию в индексируемом поле заранее заданного признака и его значению, используемому при индексации. То есть, все признаки, используемые для индексации распределены по номерам ячеек, а значения, занесенные в ячейки табличных строк, соответствуют наличию признаков, например, дате или координатам используемых при индексации. В частном случае, сведения индексируемого поля могут быть преобразованы в индексную табличную строку с использованием латентно-семантического индексирования, а отбор документов, релевантных к поисковому выражению, и формирование списка с упорядочиванием по степени релевантности, при проведении поиска, может осуществляться с использованием латентно-семантического анализа.

Для других полей формализация и нормализация признаков может производиться, например, методами латентно-семантической индексации, в результате чего, для каждого поля формируется индексное выражение, представленное в виде строки.

В частном случае, производят нормализацию семантических структур естественного языка по заранее заданному алгоритму, а латентно-семантическое индексирование производят с использованием нормализованных структур.

Перед проведением латентно-семантического индексирования может быть произведено исправление опечаток и очевидных ошибок, связанных, например, с ошибками при автозамене слов на компьютере пользователя, но преимущественным является применение нейронных сетей на DSSM на символьных триграммах, что позволяет игнорировать 1-2 ошибки в слове .

Хорошие модели ранжирования результатов, в том числе латентно-семантический анализ, позволяют достичь высокого качества поиска релевантных документов, но имеют существенные ограничения на объем обрабатываемых данных. В частности, затраты на ранжирование увеличиваются пропорционально квадрату количества записей. При реализации изобретения используются процедуры формальной обработки данных,

обеспечивающие сокращение количества одновременно обрабатываемых записей при проведении латентно-семантического анализа. За счет этого, при реализации изобретения, достигается линейная зависимость времени ранжирования от количества используемых документов, а максимально возможное количество документов превышает максимальные объёмы, достижимые без использования изобретения, например, 50 миллионов документов.

При использовании машинного обучения может использоваться стандартная или специальная функция потерь, значение которой зависит и от количества ошибок, и от серьезности ошибок.

Группировка и слияние полей может осуществляться осуществляется путем вычисления хеша многомерной координаты вектора поля таким образом, чтобы вектора ближайшие, друг к другу имели тождественный хэш. В частном случае, может использоваться полей с применением алгоритма Locality-sensitive hashing (LSH) (хэширования, чувствительного к местоположению) – вероятностный метод понижения размерности многомерных данных, основной принцип которого состоит в таком подборе хеш-функций для некоторых измерений, чтобы похожие объекты с высокой степенью вероятности имели одинаковый хэш. Преимущественной технологией для использования в настоящем изобретении является supervised hashing function или хеширующая модель, обученная с помощью ML. Такое хеширование обеспечивает тождественность хэшей, если вакансия с одним набором слов сравнивается с резюме с другим набором слов, например, такая функция хэширования сформирует одинаковый хэш для фраз «менеджер по продажам» и «специалист по работе с клиентами».

Длина хэш функции может быть задана заранее, например, после вычисления значений хэш функций для строк матрицы с пониженной размерностью и проверки адекватности применения используемой модели хэширования к матрице с пониженной размерностью. В частном случае реализации, хэш функция может быть задана в табличном виде.

Снижение размерности матрицы путем уменьшения длины векторов с использованием, например, хэширования векторов признаков уменьшает объём данных о термах (или признаках) в документах настолько, что их становится возможно хранить в индексе поисковой и группирующих систем, быстро вызывать из памяти индексные данные и применять их при формировании и групп и списков рекомендаций. Кроме того, снижение размерности матрицы улучшает качество фильтрующих и ранжирующих моделей, например, используемых в библиотеке XGBoost или LightGBM, по сравнению с тем, как если бы использовались разреженные вектора, получающиеся в результате, к примеру,

TF/IDF-векторизации bag of words с использованием словаря в несколько десятков тысяч униграммм и биграмм.

В качестве обучающих данных для каждого из полей используется реакция пользователей на предоставленные им результаты работы поисковой системы, например, отклики, приглашения, просмотры контактов, просто просмотры резюме, вакансий, добавление в избранное, распечатка, попытки позвонить, перейти в почтовый клиент, в мессенджер и т.д. В частном случае, для проведения машинного обучения, запрашивают у пользователя реальную степень соответствия запросу элементов списка. В различных частных случаях реализации изобретения оценка может быть дана пользователем непосредственно, например, по шкале градаций соответствия. В другом частном случае, степень соответствия определяется опосредованно, например, для резюме, степень соответствия повышается от игнорирования пользователем представленной ему ссылки на резюме, до трудоустройства кандидата, разместившего резюме, с промежуточными градациями «резюме просмотрено» и кандидат приглашен на собеседование. То есть, используется т.н. graded target (многоуровневая степень соответствия), где, например, например у отклика вес больше, чем у просмотра, а у приглашения на собеседование после отклика – больше, чем у простого отклика. Таким образом, для формирования обучающих данных, отслеживаются действия пользователя, заключающиеся в просмотре резюме или вакансий и направлении откликов или приглашений. При этом целесообразно рассматривать отклики и приглашения, как неравноценные события.

При проведении обучения поисковой системы, каждая обучающая группа или выборка может формировать свое решающее дерево, предпочтительно, несколько сотен решающих деревьев, сформированных путем градиентного бустинга. Сформированные деревья могут использоваться совместно при выборе вакансий, подходящих к резюме, даже при анализе группы вакансий, не связанной с группой вакансий, соответствующей обучающей выборке.

Поля рабочей базы данных также могут быть преобразованы в матричный или табличный вид с последующим сжатием матрицы и группировкой полей сжатой матрицы по признаку формального сходства.

При формировании пользователем запроса на поиск, то есть описания желаемой вакансии или резюме, пользователь может заполнить поля, предназначенные для запроса или может использоваться описание вакансии (резюме) в качестве поискового выражения.

Для поискового выражения, первоначально определяется одно или несколько значений хэш функции MLH, релевантных к поисковому выражению, а затем определяются релевантные к поисковому выражению поля.

Например, для поискового выражения, производится обработка полей по алгоритму сжатия матрицы и вычисление хэш функции, чувствительной к местоположению, что позволяет выявить группы полей из рабочей базы данных, сходных с поисковым выражением.

Дополнительно, из сведений, предназначенных для поиска могут исключаться документы, заведомо не соответствующие целям поиска, например, из поиска могут исключаться резюме соискателей не имеющих высшего образования, не готовых к переезду к месту работы и другие.

После этого. для проведения точного поиска и сортировки выявленных данных используются индексные строки, соответствующие полям, сходных с поисковым выражением. Поисковое выражение также индексируется с использованием латентно-семантической индексации. Для проведения поиска используется латентно-семантический анализ полей поискового выражения и полей из рабочей базы данных.

Методом латентно-семантического анализа определяются поля наиболее релевантные поисковому выражению. В частном случае реализации изобретения, список релевантных документов сортируется по степени релевантности и предоставляется пользователю для использования.

По результатам обработки представленных пользователю рекомендаций, производится обновление и уточнение обучающей базы данных, используемой в системе.

На ФИГ. 1 показан пример компьютерной системы общего назначения, которая включает в себя многоцелевое вычислительное устройство в виде компьютера 20 или сервера, или мобильного (вычислительного) устройства, или модуля описываемой в настоящем изобретении системы, которые, в частном случае, могут являться оконечными (вычислительными) устройствами (например, пользователя, оператора и т.д.), включающего в себя процессор 21, системную память 22 и системную шину 23, которая связывает различные системные компоненты, включая системную память с процессором 21.

Системная шина 23 может быть любого из различных типов структур шин, включающих шину памяти или контроллер памяти, периферийную шину и локальную шину, использующую любую из множества архитектур шин. Системная память 22 включает постоянное запоминающее устройство (ПЗУ) 24 и оперативное запоминающее устройство (ОЗУ) 25. В ПЗУ 24 хранится базовая система ввода/вывода 26 (БИОС), состоящая из основных подпрограмм, которые помогают обмениваться информацией между элементами внутри компьютера 20, например, в момент запуска.

Компьютер 20 также может включать в себя накопитель 27 на жестком диске для чтения с и записи на жесткий диск (не показан), накопитель 28 на магнитных дисках для чтения с или записи на съёмный магнитный диск 29, и накопитель 30 на оптическом диске для чтения с или записи на съёмный оптический диск 31 такой, как компакт-диск, цифровой видео-диск и другие оптические средства. Накопитель 27 на жестком диске, накопитель 28 на магнитных дисках и накопитель 30 на оптических дисках соединены с системной шиной 23 посредством, соответственно, интерфейса 32 накопителя на жестком диске, интерфейса 33 накопителя на магнитных дисках и интерфейса 34 оптического накопителя, твердотельного накопителя. Накопители и их соответствующие читаемые компьютером средства обеспечивают энергонезависимое хранение читаемых компьютером инструкций, структур данных, программных модулей и других данных для компьютера 20.

Хотя описанная здесь типичная конфигурация использует жесткий диск, съёмный магнитный диск 29 и съёмный оптический диск 31, специалист примет во внимание, что в типичной операционной среде могут также быть использованы другие типы читаемых компьютером средств, которые могут хранить данные, которые доступны с помощью компьютера, такие как магнитные кассеты, карты флеш-памяти, цифровые видеодиски, картриджи Бернулли, оперативные запоминающие устройства (ОЗУ), постоянные запоминающие устройства (ПЗУ) и т.п.

Различные программные модули, включая операционную систему 35, могут быть сохранены на жёстком диске, магнитном диске 29, оптическом диске 31, ПЗУ 24 или ОЗУ 25. Компьютер 20 включает в себя файловую систему 36, связанную с операционной системой 35 или включенную в нее, одно или более программное приложение (приложения) 37, другие программные модули 38 и программные данные 39. Пользователь может вводить команды и информацию в компьютер 20 при помощи устройств ввода, таких как клавиатура 40 и указательное устройство 42. Другие устройства ввода (не показаны) могут включать в себя микрофон, джойстик, геймпад, спутниковую антенну, сканер или любое другое.

Эти и другие устройства ввода соединены с процессором 21 через интерфейс 46, использующийся для связи с системной шиной. Монитор 47 или другой тип устройства визуального отображения также соединен с системной шиной 23 посредством интерфейса, например, видеоадаптера 48. В дополнение к монитору 47, персональные компьютеры обычно включают в себя другие периферийные устройства вывода (не показано), такие как динамики и принтеры.

Компьютер 20 может работать в сетевом окружении посредством логических соединений к одному или нескольким удаленным компьютерам 49. Удаленный компьютер (или компьютеры) 49 может представлять собой другой компьютер, сервер, роутер, сетевой

ПК, пиринговое устройство или другой узел единой сети, а также обычно включает в себя большинство или все элементы, описанные выше, в отношении компьютера 20, хотя показано только устройство хранения информации 50. Логические соединения включают в себя локальную (вычислительную) сеть (ЛВС) 51 и глобальную компьютерную сеть (ГКС) 52. Такие сетевые окружения обычно распространены в учреждениях, корпоративных компьютерных сетях, Интернете.

Компьютер 20, используемый в сетевом окружении ЛВС, соединяется с локальной сетью 51 посредством сетевого интерфейса или адаптера 53. Компьютер 20, используемый в сетевом окружении ГКС, обычно использует модем 54 или другие средства для установления связи с глобальной компьютерной сетью 52, такой как Интернет.

Модем 54, который может быть внутренним или внешним, соединен с системной шиной 23 посредством интерфейса 46 последовательного порта. В сетевом окружении программные модули или их части, описанные применительно к компьютеру 20, могут храниться на удаленном устройстве хранения информации. Надо принять во внимание, что показанные сетевые соединения являются типичными, и для установления коммуникационной связи между компьютерами могут быть использованы другие средства.

Дополнительно, вычислительном устройстве представленной выше архитектуры, могут использоваться аппаратные модули нейронной сети. Например, искусственные нейроны, аналогичны нейронам, описанным в https://singularityhub.com/2016/08/14/ibms-new-artificial-neurons-a-big-step-toward-brain-like-computers/#sm.0000j4eugecl8f5fyx02441eh1jdx, а также аппаратные синапсы, пример реализации которых показан в https://www.researchgate.net/publication/9009102_The_Artificial_Synapse_Chip_a_flexible_retinal_interface_based_on_directed_retinal_cell_growth_and_neurotransmitter_stimulation. Описания реализации, представленные в указанных документах, являются иллюстративными примерами реализации элементов заявленного изобретения и должны рассматриваться, как часть настоящего описания. Указанные выше элементы могут быть выполнены в виде одиночных микросхем, а также в виде наборов микросхем, позволяющих обеспечивать более гибкие настройки при формировании аппаратных модулей. В частном случае, аппаратная реализация обучающих систем, с использованием которых может быть реализовано изобретение, предусматривает а обучение нейросетей с помощью GPU (устройств обработки графики) и применение (inference) AVX-ядер обычных x86-64 CPU.

Для формирования автоматической аналитической системы, реализующей предложенный способ, используется несколько этапов предварительного обучения. Для обучения используются предварительно отобранные документы резюме и вакансий. Как

показано на фиг. 3, объявление о вакансии 300 или резюме 310, преобразуется 302 в текстовую последовательность, при этом, в текстовую последовательность может включаться информация, представленная в резюме или вакансии в виде изображений. Текст сегментируется 303 таким образом, чтобы последующая обработка сегментов текста могла быть осуществлена в физически приемлемые сроки. Например, обучение нейросети на 40 млн. примеров возможно в течение одних суток с использованием сервера 32 x86-64 CPU Intel Xeon 3,6 ГГц, 512 GB RAM, 2 GPU Nvidia 2080. Вычисление эмбеддингов при индексации за время не более двух секунд на документ требует использования 1 CPU обычного Intel Xeon 3,6 ГГц. Вычисление скалярного произведения (dot product) требует не более не более 1 мкс. на пару документов (резюме-вакансия) В частном случае может быть использована операция кодирования пар байтов (Byte Pair Encoding). Подробнее операции сегментирования текста с использованием кодирования пар байтов показаны в https://dyakonov.org/2019/11/29/%D1%82%D0%BE%D0%BA%D0%B5%D0%BD%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F-%D0%BD%D0%B0-%D0%BF%D0%BE%D0%B4%D1%81%D0%BB%D0%BE%D0%B2%D0%B0-subword-tokenization/ в разделе Byte Pair Encoding (BPE). Сформированная последовательность текстовых сегментов используется для обучения многослойного перцептрона, показанного, например, в https://wiki.loginom.ru/articles/multilayered-perceptron.html, который может быть использован в рекуррентной нейронной сети. RNN показана, например, на https://wiki.loginom.ru/articles/multilayered-perceptron.html, где в разделе GRU показана возможность формирования настроек рекуррентной нейронной сети со слоями (320 для вакансий и 330 для резюме), обученными 305 на задаче языковой модели, которая учитывает различия между словами в фразах. Описания реализации, представленные в указанных документах, являются иллюстративными примерами реализации элементов заявленного изобретения и должны рассматриваться, как часть настоящего описания. Дополнительно формируются промежуточные слои многослойного нейрона, используемые последующим слоем 307 для прогнозирования следующего слова справа.

Предпочтительным, на современном уровне развития средств обработки данных, является разбиение 303 текста, на фразы 304, содержащие около 250 слов естественного языка, что соответствует около 300 subword, расположенных подряд. В перспективе текст может быть разбит на более длинные или более короткие фразы, например, длиной от 100 до 400 слов естественного языка. Длина фразы, предпочтительно, определяется заранее заданным размером, а не смысловым построением текста документа, преимущественно, знаки препинания не учитываются и семантический разбор фраз не производится.

Предварительное формирование обученных слоев 307 для вакансий производится для рекуррентной нейронной сети на задаче языкового моделирования для прогнозирования следующего слова справа по предыдущим словам во всем документе, для каждого из документов, представляющего описание резюме или вакансии.

Указанная часть нейронной сети предназначена для определения соответствия порядка слов значащих фраз резюме и вакансий. Основной задачей обучения является прогнозирование следующего subword'a, для того, чтобы основная RNN получала на входе последовательность не subword'ов, а их последовательностей, то есть, смысловых конструкций из языка, который используется для описания резюме и вакансий. Обобщающая способность основной RNN, возможности которой ограничены аппаратными ресурсами, не расходуется на выделение смысловых конструкций и может быть перераспределена на задачу прогнозирования вероятности отклика в паре резюме - вакансией, а также задачу прогнозирования соответствия зарплат в резюме и вакансиях по отдельности.

Для формирования программно-аппаратного комплекса, непосредственного предназначенного для поиска соответствия между резюме и вакансиями, производится финальное обучение нейронной сети 400, как это показано на фиг. 4. В результате формируется база данных с настройками нейронной сети, которые, при необходимости, применяются для управления устройствами обработки данных. Значительная часть слоев или нейронов рекуррентной нейронной сети предусматривает использование семантических моделей с углубленным структурированием.

Для обучения используются пары вакансий 401 и резюме 402 (документы), для которых установлено 403 их взаимное соответствие, в частном случае, с заданием исчисляемой степени соответствия.

Для проведения обучения, как это было показано на фиг. 2, производится выделение из текста документов описательной части 203, характеризующей семантические свойства текста и «числовой» части 212, представляющей, но не всегда явно указывающей категориальные признаки или исчисляемые признаки, то есть формальные признаки, например, размер зарплаты, географическое положение, предпочтительная форма сотрудничества, например, удаленное и другие.

Указанные в документах числовые и категориальные признаки учитываются при проведении анализа не сами по себе, а как одна из составных частей текста документа. Такой подход позволяет выявить пары документов, для которых существует высокая вероятность отклика, даже если между документами нет прямого точного соответствия в части указанных признаков. То есть, при использовании предложенного изобретения, у

внутренних MLP-эмбеддингов (векторов) 410 числовых и категориальных признаков нет заранее определённого преимущества перед текстовыми векторами 420, сформированными по результатам анализа subwords. При использовании обучения на парах документов с минимизацией бинарной взаимной энтропии 460, при обучении, фактически уточняется допустимый разброс параметров, который не может быть определен другими способами. При отсутствии числовых и категориальных признаков, а также для уточнения указанных признаков, используется интеллектуальный анализ документов с целью выделения навыков, опыта и требований, которые являются основанием для определения соответствующих признаков. Таким образом обучаются слои многослойного персептрона (MLP), используемые для прогнозирование наиболее вероятных значений признаков, которые потом применяются для прогнозирования вероятности отклика. Итоговое значение, используемое для обучения, принимается, например, равным 0 или «ложь», если не было ни отклика соискателя с данным резюме на данную вакансию, ни приглашения кандидата с данным резюме на собеседование на данную вакансию, и, например 1 или «истина», если отклик на вакансию от соответствующего соискателя состоялся, или кандидат заинтересовался вакансией.

На этапах, которые могут быть реализованы последовательно или параллельно с описанными выше этапами, осуществляется настройка нейронов нейронной сети 470, формирующих возможность прогнозирования зарплаты для вакансии и резюме. Существенным признаком изобретения является совместное обучение участков или слоев нейросетей и для прогнозирования вероятности отклика 480 для резюме и 485 для вакансии, и для прогнозирования размера заработной платы 455 для резюме и 450 для вакансии или других категориальных признаков, например, для определения вероятности переезда кандидата или принятие предложения об удаленной работе. Важно, что предложенное изобретение неявным образом осуществляет проверку соответствия важных критериев, к которым относятся указанные или прогнозируемые ожидания по заработной плате, так как при существенных расхождениях в указанных ожиданиях, соответствие требований, например, к профессиональному опыту кандидатов, вероятность отклика невелика. Настройки 499 нейронных сетей, например, 450, 455, 480 и 485 максимально соответствуют вероятности отклика между парами документов из обучающей базы данных.

В RNN DSSM преимущественным объектом анализа является порядок слов в текстах. Дополнительно может использоваться линейное сегментирование текста по словам, с применением BOW->TF/IDF->SVD, BOW->TF/IDF->SVD->cos, BOW->PPMI->SVD, BOW->PPMI->SVD->cos, символьные 3-граммы -> 2-слойная DSSM без LM- и RNN-частей без использования многозадачности, для учёта, например, словообразования, опечаток,

выявления редких символов. Обучаемая структура, обеспечивающая анализ данных и определение вероятности отклика, состоит из последовательности, сформированной первым слоем нейронов, анализирующих результаты сегментации текстов резюме и вакансий, слоев рекуррентной нейронной сети, обученных на задаче языкового моделирования, с учетом порядка расположения слов. Предварительно обученные слои в дальнейшем не изменяются. Далее используются персептроны, анализирующие сочетания категориальных и числовых признаков, а также объединяющие сигналы от других персептронов и RNN, а также слои рекуррентной нейронной сети, обеспечивающие извлечение данных о смысловых конструкциях в тексте. Результаты используются по отдельности нейронами, прогнозирующими вероятность отклика или приглашения, а также нейронами, используемыми для прогнозирования приемлемых размеров зарплаты или других приемлемых категориальных признаков.

Обучение нейронных сетей, в результате которых формируются настройки параметров нейронов в составе персептронов и RNN может производиться по мере того как качество работы нейросетей, выраженное в количестве откликов и приглашений, снижается ниже допустимого уровня, по расписанию, например, раз в полгода или может осуществляется в соответствии другими критериями.

Настроенные нейросети могут использоваться для определения соответствия произвольно выбранных пар резюме-вакансия, однако при прямом использовании нейросетей для указанных задач объем производимых операций не позволяет осуществить анализ данных в приемлемые сроки. В связи с этим при реализации способа используется предварительный расчёт нейросетевых эмбеддингов для документов в ходе индексации описаний резюме и вакансий, а результаты индексации далее используются для отбора документов, которые в наибольшей степени соответствуют запросам кандидатов и работодателей.

Для целей индексации используется векторное представление результатов анализа резюме и вакансий нейронными сетями. Хотя вектора не имеют физической природы, нейронные сети настраиваются при обучении таким образом, чтобы векторные представления обработки данных имели степень сходства, зависящую от степени сходства резюме и вакансий, для которых пользователями подтверждено соответствие.

Подтверждение соответствия определяется по результатам действий пользователей, например, просмотр объявления является подтверждением соответствия низкого уровня, приглашение на собеседование или иная коммуникация подтверждает более высокую степень соответствия, а постоянное трудоустройство, успешное прохождение

испытательного срока, а также профессиональный или карьерный рост сотрудника на рабочем месте могут подтверждать более высокую степень соответствия требованиям.

Как показано на фиг. 5, настроенная нейронная сеть 400 с настройками 499 и пользуется для индексации вакансий, поступивших в базу данных. Для формирования индекса, поступившие от соискателей 501 описания 502 вакансий и поступившие от работодателей 503 описания резюме обрабатываются обученными или настроенными, по результатам обучения 400, рекуррентными нейронными сетями, причем, по результатам обработки, для каждого документа, описывающего резюме или вакансию, формируются вектора 505 – 508 или эмбеддинги, которые сами по себе описательными не являются, но представление векторов во внутренней структуре нейронных сетей выполнено таким образом, что сходные вектора относятся к документам, имеющим высокую степень сходства. Например, вектор 505 характеризует прогноз зарплаты по вакансии, вектор 508 характеризует прогноз зарплаты по резюме, а вектора 506 и 507 характеризуют тексты описаний вакансий и резюме, соответственно. Кроме векторов, формируемых рекуррентной нейронной сетью, для проведения поиска сходных документов могут использоваться другие векторные представления, например, векторное представление текста названия вакансии, в качестве которой также может использоваться название позиции, которая характеризует вакансию, либо названия профессии. Вместе с тем, предложенный подход не требует специального использования указанной позиции, так как проводимый, в рамках использования настоящего изобретения, анализ текста с высокой степенью достоверность определяет соответствие профессиональных качеств в парах документов, даже если пользователи допустили ошибку при их явном указании. По результатам обработки документов с использованием рекуррентных сетей в явном виде формируются прогнозы зарплаты для резюме и вакансии, которые соответствуют зарплатным ожиданиям кандидатов с соответствующими числовыми и категориальными параметрами, а также смысловыми конструкциями в резюме, которые имеют отношение к зарплатам.

Таким образом, формирование индекса документов в векторной форме производится однократно для каждого документа, а для определения степени соответствия документов друг другу используется скалярное произведение векторов, где ортогональные вектора характеризуют документы, не имеющие взаимного соответствия, то есть определяются или выявляются пары документов, для которых вероятность отклика и/или приглашения невелика, и которые не имеет смысла высоко ранжировать и/или отображать. В результате, операции непосредственного сравнения документов, требующие использования большого количества ресурсов, для анализа каждого из документов, при каждом сравнении,

заменяются операциями однократного индексирования каждого из документов. Последующая операция определения скалярного произведения полученных векторов аппаратными методами экономично использует вычислительные ресурсы, особенно при использовании AVX-ядер современных x86-64 микропроцессоров.

Для упрощения поиска сходных документов, используется процедура предварительного сравнения наиболее важных свойств документов, отраженных в векторном представлении.

С этой целью векторам документов приводят в соответствие значения функции хэширования векторов. Функция хэширования 700 supervised hashing (MLH), была разработана в соответствии с рекомендациями https://www.ee.columbia.edu/ln/dvmm/publications/12/PAMI_SSHASH.pdf, http://www.cs.toronto.edu/~norouzi/research/papers/min_loss_hashing.pdf, https://arxiv.org/pdf/1004.5370.pdf, https://arxiv.org/pdf/1509.05472.pdf, где описания реализации функций, представленные в указанных выше документах, являются иллюстративными примерами реализации элементов заявленного изобретения и должны рассматриваться, как часть настоящего описания. Функция (модель) хеширования 700 обучается и длина её хэша выбирается такой, чтобы при минимальной длине хэша, значение хэша сходных векторов, то есть, т.е. векторов резюме и вакансии между которыми велика вероятность отклика и/или приглашения, были сходны по максимальному количеству бит, а для не сходных, наоборот, отличались значениями максимального количества бит.

На этапе обучения хэш-функции 700 база данных функций делится на обучающие наборы, а также наборы тестов, построенный набор хэш-функций обучается и изучается в базе данных обучения. На этапе формального хеширования исходные функции подставляются в результаты обучения, чтобы получить соответствующий хеш-код.

В частном случае реализации изобретения, хэш код может относиться к векторному представлению описания вакансии, в котором не учитывается порядок слов в предложении, а соответствующий вектор может быть сформирован без использования нейронных сетей, например, с использованием устройств, реализующих эмбединги или вектора, полученные применением SVD к TF/IDF и PPMI векторам над bag of words из уни- и биграмм по леммам текста. Такой подход упрощает обучение хэш функции, поскольку использует вектора пониженной размерности.

Как показано на фиг. 7, с использованием функции 700 векторам RNN DSSM вакансий и резюме, 505 и 508, соответственно, а также другим возможным векторам 705 вакансий и векторам 708 резюме приводятся в соответствие хэши 715 вакансий и хэши 718 резюме.

Категориальные признаки и числовые признаки, например, значение зарплаты, уточненное по результатам анализа документов, хранятся в явном виде. Например, значение зарплаты может храниться в виде параметров функции распределения, определенной по результатам анализа реальных должностей.

По результатам обработки данных, формируется общая база данных 650, в которой представлены база данных 610 индексов резюме и база данных 620 индексов вакансий, как это показано на фиг. 6. Каждая из уникальных записей, соответствующих вакансии, содержит блок данных 611 или поле, характеризующее прогноз зарплаты, блок данных 612 хранящий значение вектора RNN DSSM вакансии, а также хэш 613 вектора RNN DSSM вакансии. Каждая из уникальных записей, соответствующих резюме, содержит блок данных 621 или поле, характеризующее прогноз зарплаты, блок данных 622 хранящий значение вектора RNN DSSM резюме и хэш 623 вектора RNN DSSM резюме. Указанные выше сведения соответствуют блокам 450, 485 715, 718, 480 и 455, значения которых определяются по мере поступления информации от пользователей. Сформированный индекс документов описывает резюме и вакансии, где каждый из документов представлен, в том числе, категориальными и числовыми признаками, например, географическим местом работы и размером зарплаты. Кроме этого, каждый из документов описывается вектором, который отображает не только текстовое преставление документа, но и порядок слов в тестовом описании. Также документ отображается упрощенным значением, которое может являться значением обучаемой хэш функции, которая может относиться к указанному вектору, либо упрощенное значение, которое используется для выявления сходства документов, может быть получено другим путем, например, представлять индексное значение, в котором не учитывается порядок слов в документе.

В иллюстративном примере реализации, при поступлении от пользователя запроса на подбор документов, как иллюстративно показано на фиг. 8, например, запроса 802 соискателя 801 на подбор описаний вакансий, соответствующих описанию резюме, выбранному или указанному пользователем, с использованием значений 821, 822 и 833, отраженных в индексе для данного резюме, на первом этапе 803 отбираются вакансии с заданной степенью соответствия значений хэш функций, например, выбираются вакансии, значение хэша 611 которых отличается от хэша 821 выбранного резюме не более чем на заданное количество бит.

На этапе 810, для векторов вакансий, отобранных на этапе 803, производится вычисление скалярных произведений каждого из векторов 612 отобранных вакансий и вектора 822 выбранного резюме. Вычисленные значения скалярных произведений далее

используются для анализа с использованием различных критериев соответствия резюме и вакансий.

На следующем шаге 804, по значениям скалярных произведений, отбираются документы, для которых значение скалярного произведения выше предварительно заданного порога. Указанная операция может быть осуществлена простой сортировкой значений и не требует существенных затрат вычислительных мощностей. При задании порогов моделей-классификаторов пороги задаются в явной форме, а количество порогов может изменяться. Из-за этого может отличаться, в некоторых пределах, время ответа системы, но за счет использования фиксированных значений порогов, сформированные модели имеют предсказуемое качество. При использовании динамически изменяющихся порогов, например, при изменении порогов при каждом запросе, качество поиска, то есть, баланс между полнотой просмотренных документов и точностью соответствия отобранных документов критериям запроса при сохранении значений, остается стабильным, а при изменении значений меняется нелинейно. В предложенной системе значение порога подбирается однократно без применения итерационных методов уточнения значений, что обеспечивает снижение требований к объемам оперативной памяти вычислительных устройств, поскольку ячейки, в которых были записаны значения векторов документов, не прошедших отбор, освобождаются сразу после проверки. В зависимости от запросов пользователей, количество документов, которые должны быть отобраны на каждом шаге, может изменяться динамически. Таким образом, на этапе 804, из индексов десериализуются признаки, которые используются в последующей фильтрующей модели, применяющей логистическую регрессию.

На следующем этапе 805, который не является обязательным, с использованием линейного фильтра, по значениям скалярных произведений и других признаков выбираются для последующего анализа вакансии, соответствующие резюме, например, по совокупности значений 10 признаков или более. Двухстадийный отбор не является обязательным, однако позволяет получить намного более высокую степень соответствия для выбранных документов при той же ресурсоёмкости, в случае, когда первая, менее ресурсоёмкая фильтрующая модель, проводит фильтрацию по примерной, вероятности отклика и/или приглашения с небольшим количеством признаков и завышенной полнотой, а вторая, более ресурсоёмкая, фильтрацию на основе более точной фильтрации на основе большего количества признаков и более сложных моделей (GBDT).

На следующем этапе 806, увеличивают количество признаков, которые используются для ранжирования между собой документов, которые представляют интерес для пользователя, отобранных на этапе 805. Например, с использованием градиентного

бустинга ансамблем решающих деревьев, отбирают вакансии, вектора которых соответствуют вектору резюме по совокупности значений 30 признаков или более. Далее, на этапе 807, с использованием ранжирования с применением градиентного бустинга с ансамблем решающих деревьев, производится сортировка документов и составление списка, в который включаются документы, вектора которых соответствуют вектору резюме, например, по совокупности значений 700 признаков. Таким образом производится ранжирование документов по степени сходства для большего количества признаков, представленных векторами документов, а также векторами интересов пользователя, связанных как с содержанием документов, просматриваемых им, а также пользователями, имеющими сходные описания резюме, так и с географическим расположением объектов, которые они описывают (например, адресом работодателя из вакансии, районом проживания соискателя из резюме).

Дополнительно, для документов, обладающим степенью сходства выше заданного порога, пользователем может быть произведена сортировка или дополнительная фильтрация в соответствии со степенью совпадения числовых и/или категориальных признаков. Отдельная сортировка по числовым и категориальным признакам может являться необходимой в некоторых случаях, в связи с тем, что у пользователя могут появляться дополнительные критерии подходящих для него вакансий, которые никак не отражены в его резюме, предыдущем поведении и местоположении. Например, соискатель может изменить место нахождения вакансии, если приемлемые вакансии отсутствуют в пределах изначально заданного региона, в этом случае система может обеспечить рекомендации вакансий, которые для похожих пользователей обычно находятся на приемлемом расстоянии, в рамках ежедневной транспортной доступности. Изменяя требования к зарплате, работодатель может определить причину отсутствия приемлемых резюме, например, соискатель может скрывать для себя вакансии, которые он не готов рассматривать на данном этапе поиска работы, в связи с чем, для таких вакансий могут отсутствовать просмотры.

В общем случае, может быть произведена оценка вакансий, на соответствие заданным критериям сходства, например с последовательным использованием порогового значения скоров моделей-классификторов, классифицирующих пары документов на такие, в которых между резюме и вакансией есть достаточная вероятность отклика и последующего приглашения, и в которых такая вероятность мала. Последовательное применение этих методов, по сути, осуществляет проверку на соответствие с постепенным ужесточением критериев.

Далее отобранные на последнем этапе документы предъявляются 808 пользователю в виде списка, где в верхней части списка располагаются наиболее релевантные документы.

Очевидно, что при обработке поисковых запросов работодателей используется аналогичная последовательность шагов, а базы данных 610 и 620 меняются ролями.

Формула изобретения

1. Способ автоматизированного поиска персонала путем сравнения описаний резюме и вакансий, заключающийся в том, что из каждого описания выделяют три блока признаков, производят раздельное сравнение соответствующих блоков сравниваемых пар описаний, соответствие описаний пар определяют по результатам сравнения всех блоков, где:

в объеме первого блока выделяют исчисляемые параметры, сочетаниям которых, для каждого из документов присваивают промежуточные векторные представления исчисляемых параметров, а в качестве результата сравнения, определяющего степень соответствия, используют скалярные произведения промежуточных векторных представлений исчисляемых параметров;

в объеме второго блока выделяют семантические составляющие описаний, определяют поисковые вектора для семантических составляющих и производят определение соответствия описаний путем сравнения векторов и вычисления вероятности соответствия семантических составляющих;

в объеме третьего блока для каждого из описаний выделяют семантические структуры описаний, в которых существенным является порядок следования грамматических форм, и производят определение вероятности соответствия указанных структур сравниваемых описаний, при этом:

при сравнении вторых блоков не учитывают семантические составляющие описаний, для которых установлено несоответствие семантических структур описаний, соответствующих соответствующим семантическим составляющим, а:

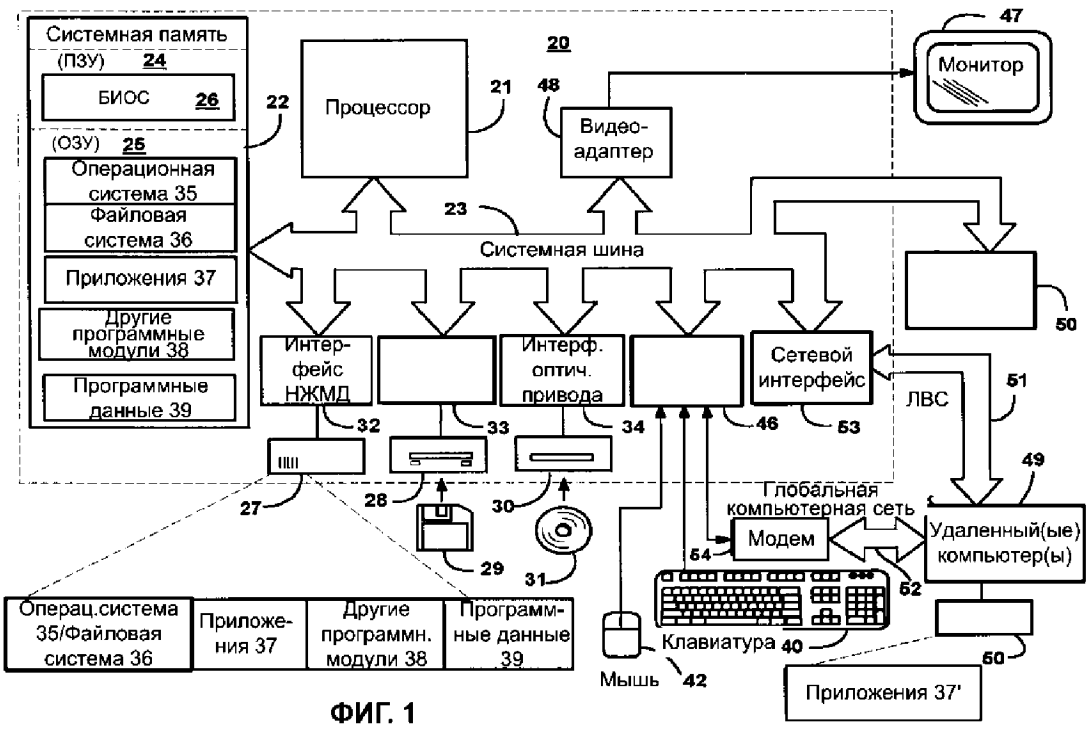
для выделения семантических структур, в которых существенным является порядок следования грамматических форм, используют рекуррентную нейронную сеть, при этом производят обучение рекуррентной нейронной сети по парам резюме-вакансия, для которых определено соответствие между описаниями и парам резюме-вакансия, для которых установлено несоответствие между описаниями.

2. Способ по пункту 1, отличающийся тем, что для обучения рекуррентной нейронной сети используют пары резюме-вакансия, максимально соответствующие друг другу, где степень соответствия определяется с использованием действий соискателей и работодателей, разместивших резюме и вакансии, в том числе, откликов и приглашений на собеседования, где первый и второй блоки сгруппированы в виде слоев нейронов предварительно обученной нейронной сети.

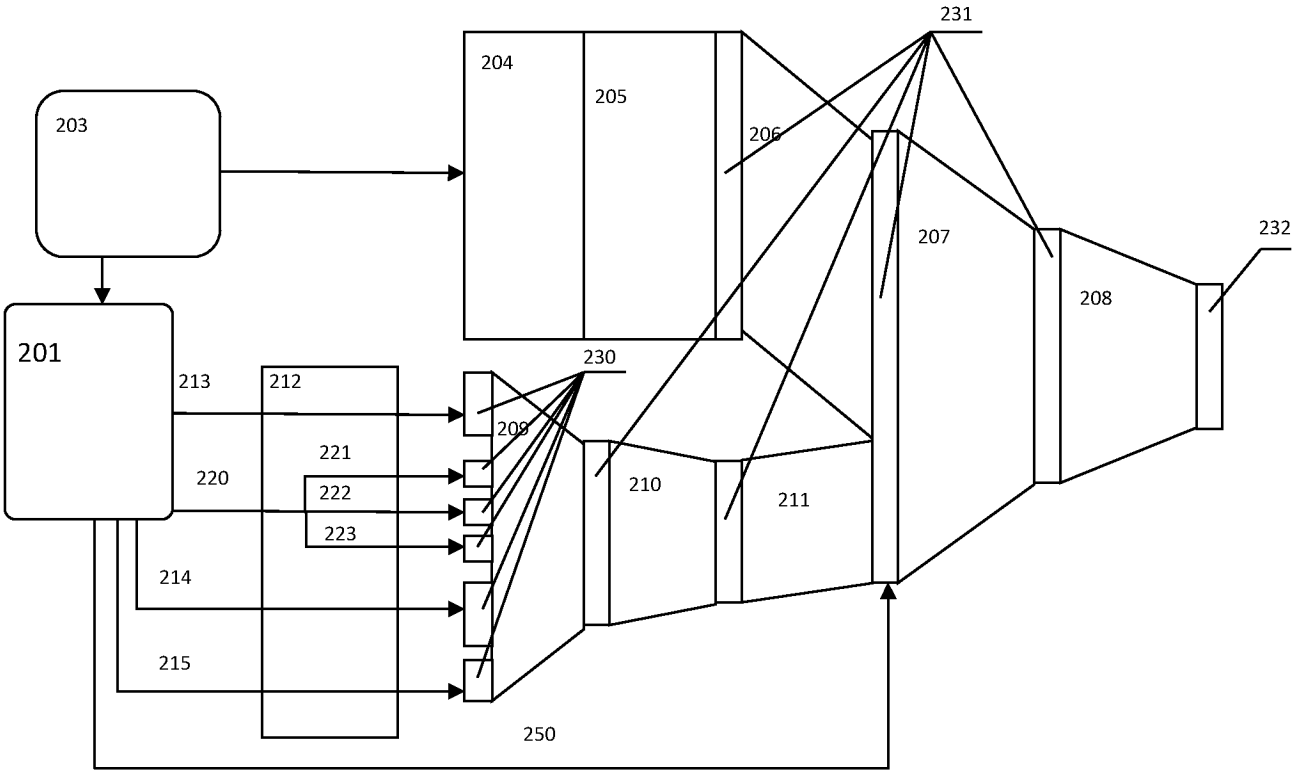
3. Способ по пункту 2, отличающийся тем, что для обучения рекуррентной нейронной сети используют 4-6% пар резюме-вакансия.

4. Способ по пункту 3, отличающийся тем что для обучения используют пары резюме-вакансия, равномерно распределенные по датам составления за календарный год.

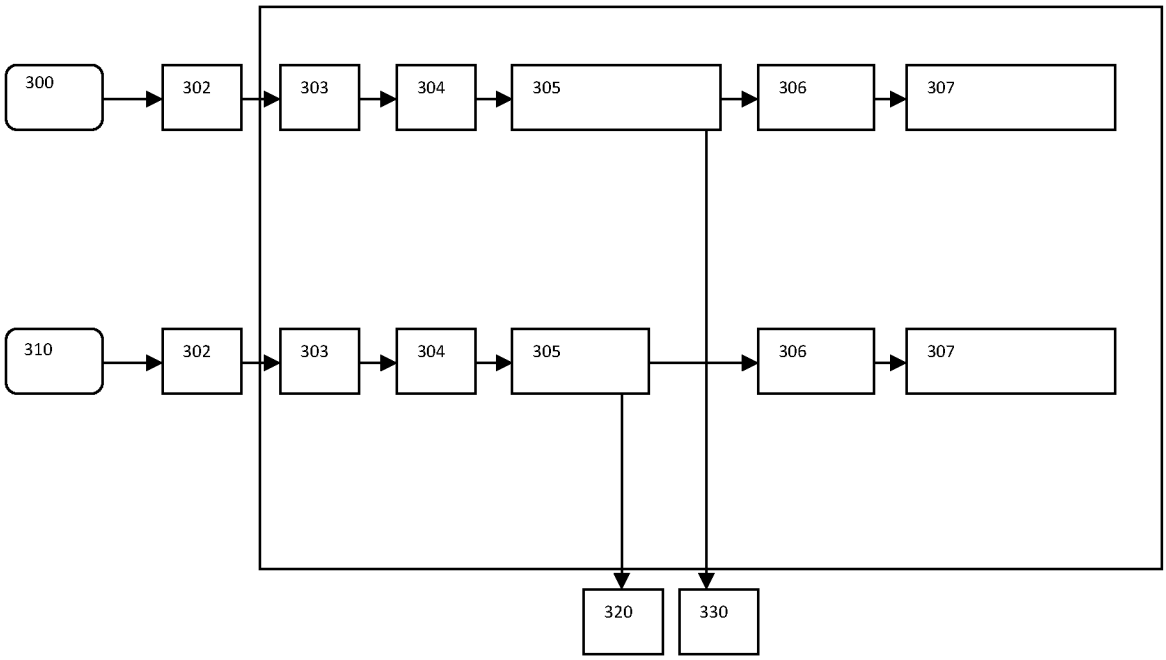
5. Способ по пункту 1, отличающийся тем, что векторные представления исчисляемых параметров формируют путем анализа текста документов с использованием слоев нейронной сети.


Системная память
(ПЗУ) 24
БИОС 26
(ОЗУ) 25
Операционная система 35
Файловая система 36
Приложения 37
Другие программные модули 38
Программные данные 39
Процессор
Видео-адаптер
Монитор
Системная шина
Интер-фейс НЖМД
Интерф. оптич. привода
Сетевой интерфейс
ЛВС
Глобальная компьютерная сеть
Модем
Удаленный(ые) компьютер(ы)
Клавиатура
Мышь
Приложения 37'
Операц.система 35/Файловая система 36
Приложе-ния 37
Другие программн. модули 38
Программ-ные данные 39

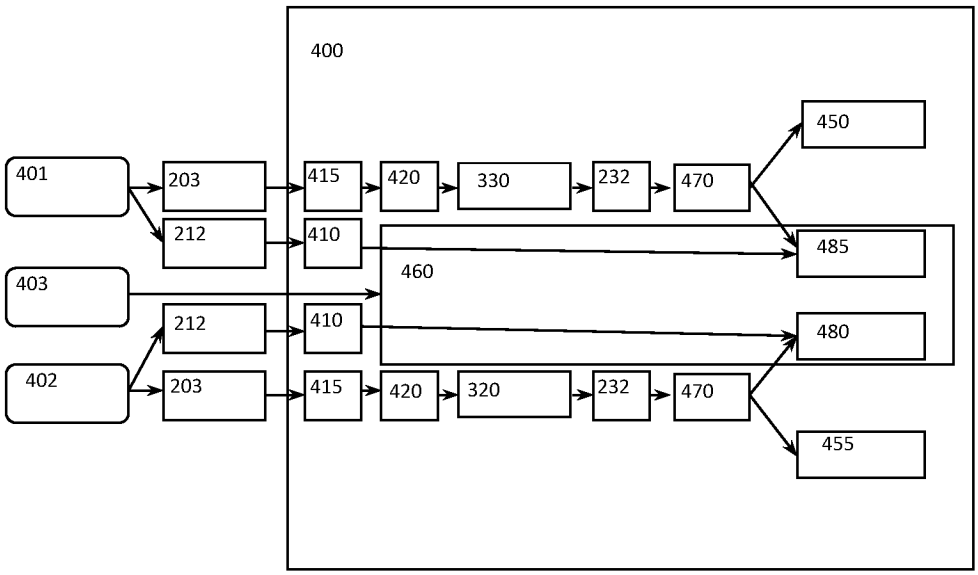
ФИГ. 1


203
201
204
205
206
207
208
231
232
230
209
210
211
212
213
220
221
222
223
214
215
250

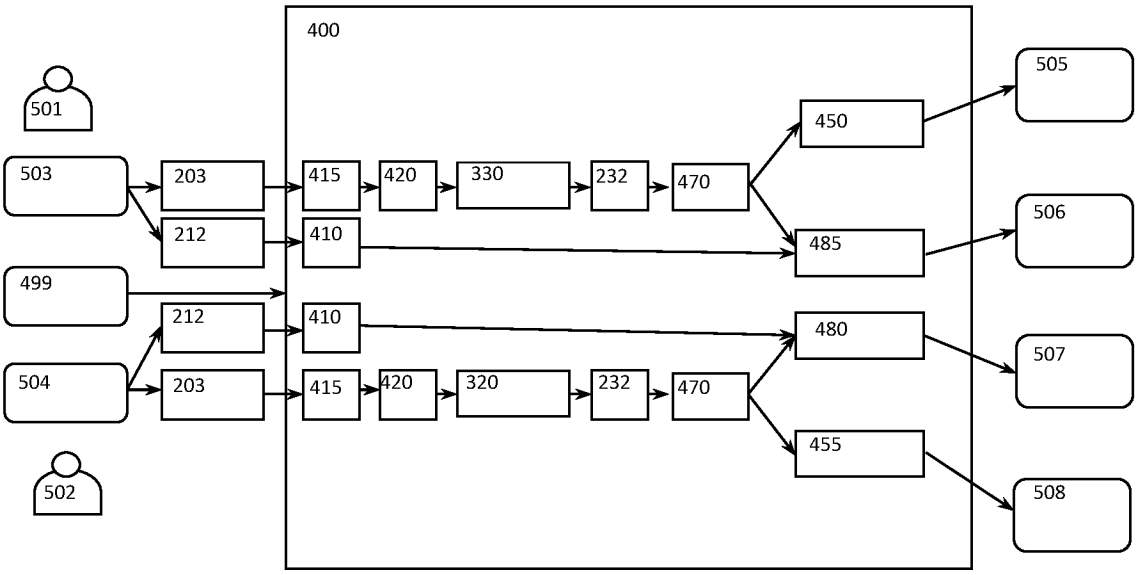
Фиг. 2


300
302
303
304
305
306
307
310
302
303
304
305
306
307
320
330

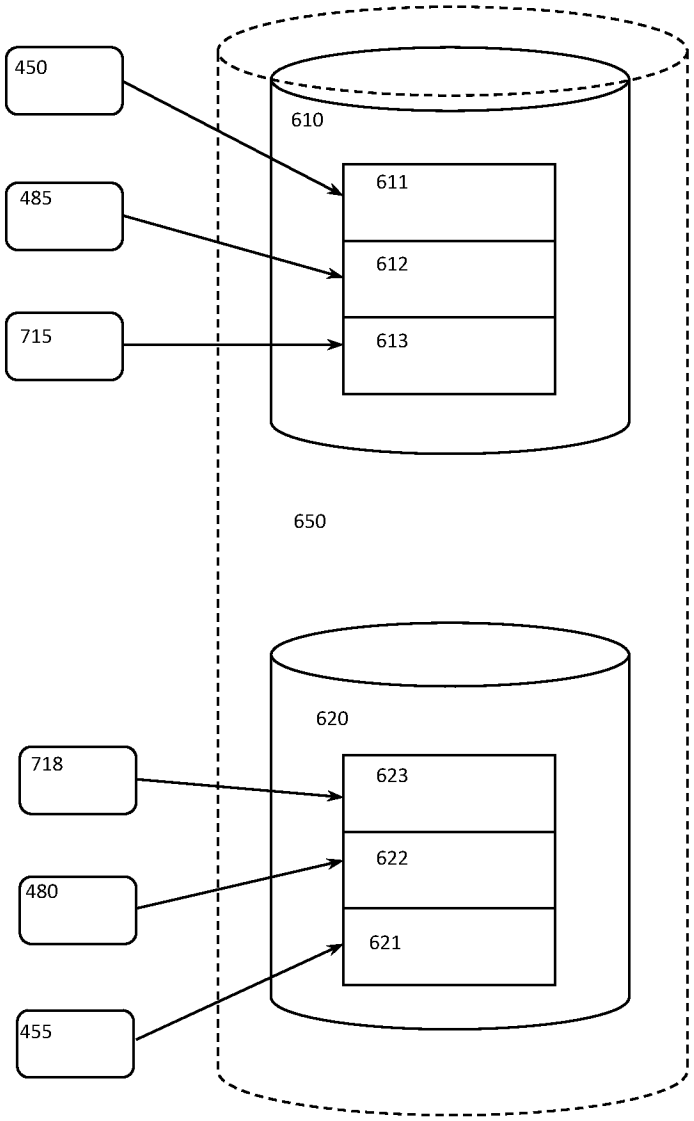
Фиг. 3


400
401
203
212
415
420
330
232
470
450
410
485
460
403
212
410
480
402
203
415
420
320
232
470
455

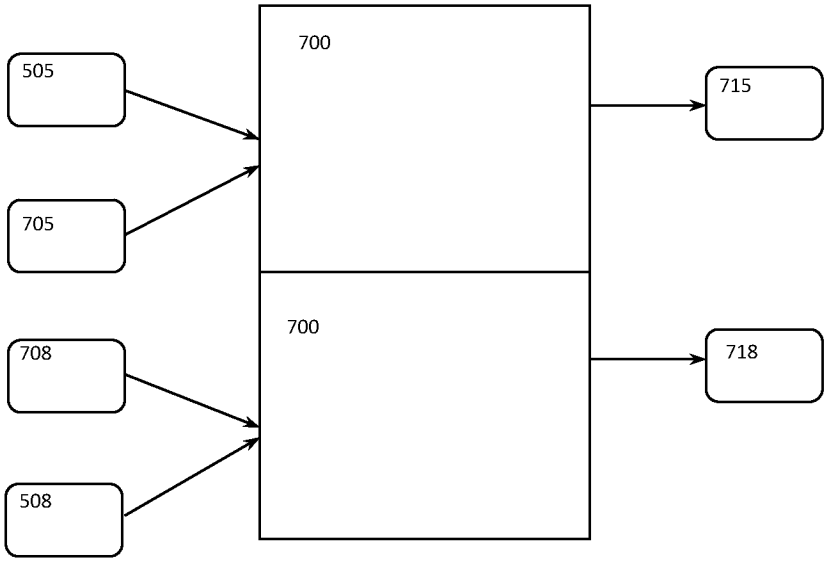
Фиг. 4


400
501
503
499
504
502
203
212
212
203
415
420
330
232
470
410
410
415
420
320
232
470
450
485
480
455
505
506
507
508

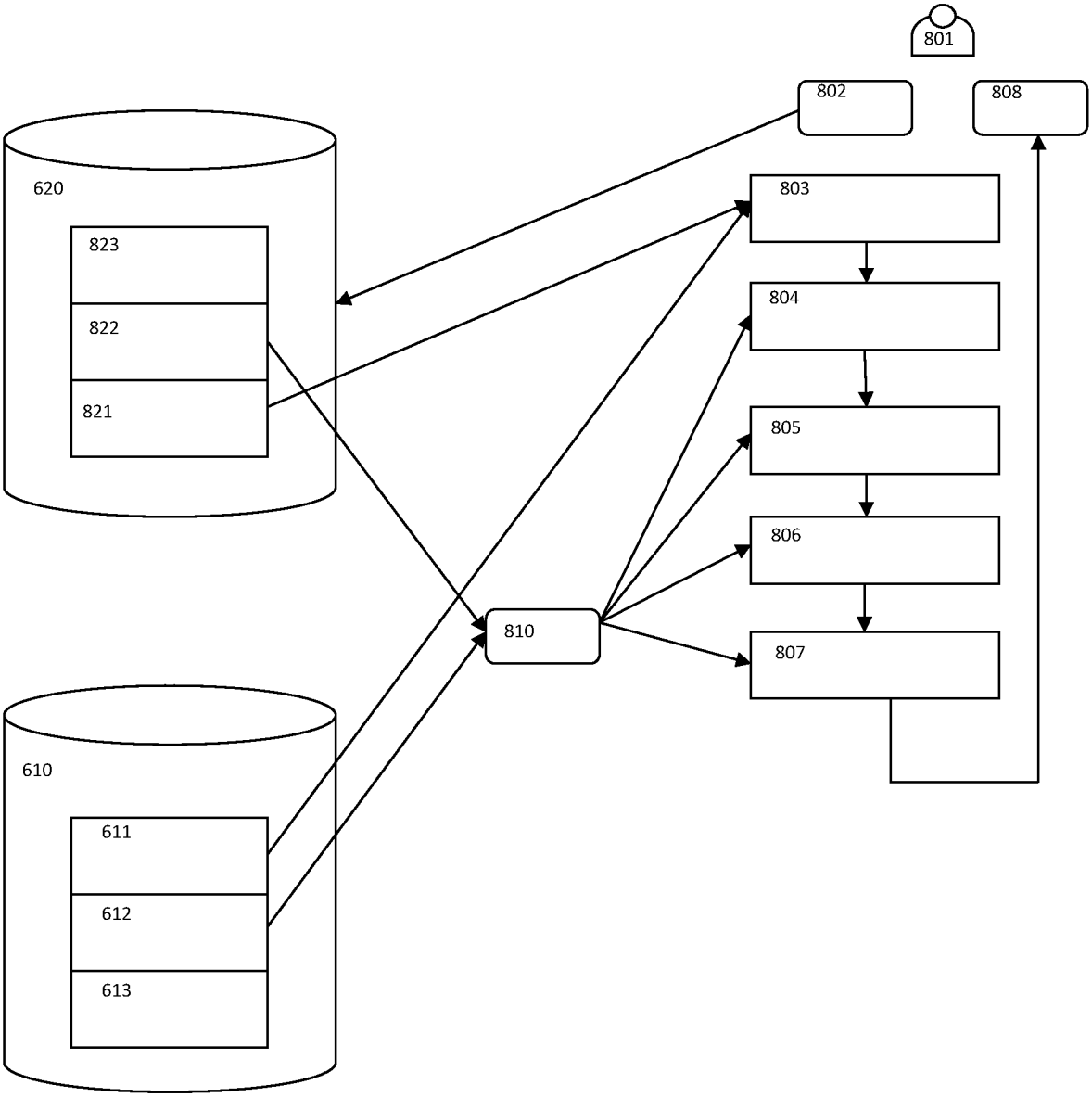
Фиг. 5


450
485
715
610
611
612
613
650
620
718
480
455
623
622
621

Фиг. 6



Фиг. 7


801
802
808
620
823
822
821
803
804
805
806
807
810
610
611
612
613

Фиг. 8

ОТЧЕТ О ПАТЕНТНОМ ПОИСКЕ

(статья 15(3) ЕАПК и правило 42 Патентной инструкции к ЕАПК)

Номер евразийской заявки:

202193231

А. КЛАССИФИКАЦИЯ ПРЕДМЕТА ИЗОБРЕТЕНИЯ:

G06F 16/20 (2019.01)
G06F 16/33 (2019.01)
G06F 40/253 (2020.01)
G06N 3/02 (2006.01)
G06N 20/00 (2019.01)

Согласно Международной патентной классификации (МПК)

Б. ОБЛАСТЬ ПОИСКА:

Просмотренная документация (система классификации и индексы МПК)
G06F 16/00-16/33, 17/00-17/30, 40/00-40/253, G06N 3/00-3/02, 20/00, G06Q 10/00-10/00

Электронная база данных, использовавшаяся при поиске (название базы и, если, возможно, используемые поисковые термины)
ESP@CENET, K-PION, PAJ, USPTO, WIPO, GOOGLE, ИС «ПОИСКОВАЯ ПЛАТФОРМА» (РОСПАТЕНТ)

В. ДОКУМЕНТЫ, СЧИТАЮЩИЕСЯ РЕЛЕВАНТНЫМИ

Категория*	Ссылки на документы с указанием, где это возможно, релевантных частей	Относится к пункту №
X	CHIRAG DARYANI et al., «AN AUTOMATED RESUME SCREENING SYSTEM USING NATURAL LANGUAGE PROCESSING AND SIMILARITY», ETHICS AND INFORMATION TECHNOLOGY, January 2020, 5 л., [онлайн] [найдено 02.06.2022]. Найдено в < https://www.researchgate.n et/publication/347633 082_AN_ AUTOMATED_RESUM E_SCREENING_SY STEM_USING_NATURAL _LANGUAGE _PROCESSING_AN D_SIMILARITY> Реферат , разделы 1, 3.1.2, 3.1.3, 3.2.1, 3.2.3, 4, 5, 6, фиг. 1, 2	1 – 5
X	US10,691,732 B1, (ZHANG G.), 23.06.2020 реферат, колонка 1, строка 63 – колонка 2, строка 12, колонка 3, строки 18 –25, колонка 7, строки 50 – 61, колонка 10, строки 17 – 28, колонка 17, строки 17 – 24	1 – 5
A	US2019/0114593 A1, (EXPERTHIRING, LLC), 18.04.2019	1 – 5
A	US2019/0164132 A1, (MICROSOFT TECHNOLOGY LICENSING, LLC), 30.05.2019	1 – 5

☐ последующие документы указаны в продолжении

* Особые категории ссылочных документов:
«A» - документ, определяющий общий уровень техники
«D» - документ, приведенный в евразийской заявке
«E» - более ранний документ, но опубликованный на дату подачи евразийской заявки или после нее
«O» - документ, относящийся к устному раскрытию, экспонированию и т.д.
"P" - документ, опубликованный до даты подачи евразийской заявки, но после даты испрашиваемого приоритета"

«T» - более поздний документ, опубликованный после даты приоритета и приведенный для понимания изобретения
«X» - документ, имеющий наиболее близкое отношение к предмету поиска, порочащий новизну или изобретательский уровень, взятый в отдельности
«Y» - документ, имеющий наиболее близкое отношение к предмету поиска, порочащий изобретательский уровень в сочетании с другими документами той же категории
«&» - документ, являющийся патентом-аналогом
«L» - документ, приведенный в других целях

Дата проведения патентного поиска: **02/06/2022**

Уполномоченное лицо:
Начальника отдела механики,
физики и электротехники

Д.Ф. Крылов