(19) Евразийское патентное ведомство

(11) 042532 (13) B1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОМУ ПАТЕНТУ

(45) Дата публикации и выдачи патента
2023.02.22

(21) Номер заявки
202100143

(22) Дата подачи заявки
2020.03.17

(51) Int. Cl. ***G06Q 10/00*** (2012.01)
G06Q 10/06 (2012.01)
G06Q 50/10 (2012.01)

(54) РЕКОМЕНДАТЕЛЬНАЯ СИСТЕМА ПОДБОРА ПЕРСОНАЛА С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ И С ПОНИЖЕНИЕМ РАЗМЕРНОСТИ МНОГОМЕРНЫХ ДАННЫХ И СПОСОБ ПОДБОРА ПЕРСОНАЛА С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ И С ПОНИЖЕНИЕМ РАЗМЕРНОСТИ МНОГОМЕРНЫХ ДАННЫХ

(31) **2019107661**
(32) **2019.03.18**
(33) **RU**
(43) **2021.08.16**
(86) **PCT/RU2020/050048**
(87) **WO 2020/190176 2020.09.24**

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

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

(74) Представитель:
Киселев А.Е. (RU)

(56) US-A1-2018173803
US-A1-2019034793
US-A1-2017032035

(57) Изобретение относится к поисковым системам и может быть использовано для поиска вакансий и резюме в базах данных. Также изобретение может быть использовано для сортировки вакансий и резюме по профессиональным областям и другим заданным критериям сходства. Техническим результатом является сокращение трудозатрат на подготовку материалов, необходимых для машинного обучения, сокращение объемов машинных вычислений, ускорение получения результатов, сокращение времени на анализ результатов поисков. Для достижения технического результата в рекомендательной системе используют обучающую и рабочую базы данных, при этом обучающие последовательности формируются путем сортировки и обработки матриц обучающей базы данных на основании отзывов пользователей о степени результатов представленных им результатов тематике запросов.

042532 B1

042532 B1

Техническая область

Предложенное изобретение относится к поисковым системам и может быть использовано для поиска вакансий и резюме в рекомендательных системах подбора персонала с большим количеством записей в базах данных. Также система и способ могут быть использованы для сортировки вакансий и резюме по профессиональным областям и другим заданным критериям сходства.

Одной из проблем, возникающей при поиске резюме, соответствующих вакансиям, является большое количество документов, выдаваемых по поисковым запросам, что обусловлено 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 работодателям принять решение о выборе кандидата с минимальными трудозатратами, в связи с чем имеется потребность в экспертных системах, снижающих трудозатраты работодателей при выборе сотрудников. Обратная задача не является источником существенных проблем, поскольку вакансии, как правило, открываются при наличии потребности в сотруднике и их количество в базах данных, как правило, на один или два порядка меньше, чем количество резюме.

Одной из экспертных систем, наиболее близкой к предложенному изобретению, является система, описанная в патентной заявке США US2018173803 (A1), опубликованной 21.06.2018 г. 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.

Краткое изложение изобретения

Для достижения технического результата предлагается способ автоматизированного поиска релевантных документов для рекомендательной системы подбора персонала с использованием машинного обучения и с понижением размерности многомерных данных, заключающийся в том, что формируют рабочую базу 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; обновляют рабочую базу данных для документов путем удаления устаревших документов и добавления новых документов; формируют обучающую базу данных для документов, содержащую документы, для которых получены отзывы пользователей, характеризующие релевантность документов поисковым пользовательским запросам; обновляют поисковую базу данных путем добавления документов, для которых получены новые отзывы пользователей о соответствии документов пользовательским запросам; для каждого из документов индексируют поля путем преобразования сведений индексируемого поля в индексную табличную строку, отдельные ячейки которой соответствуют одному из заранее заданных признаков, соответствующему положению ячейки в индексной табличной строке, а значения, занесенные в ячейку, соответствуют наличию в сведениях индексируемого поля термина естественного языка, соответствующего ячейке; формируют обучающую индексную матрицу, в качестве строк которой используют индексные строки, соответствующие документам обучающей базы данных для документов; с использованием заранее заданной процедуры сжатия матриц сжимают обучающую индексную матрицу с частичной потерей данных с понижением размерности с формированием сжатой индексной матрицы; группируют строки сжатой индексной матрицы по заранее заданным формальным признакам сходства, где документы, соответствующие сгруппированным строкам в одной группе, составляют обучающую группу документов; и формируют решающую систему, для обучения которой в качестве исходных данных используют индексные строки документов соответствующей обучающей группы документов, а в качестве ожидаемых результатов - отзывы пользователей для документов из обучающей группы; и для поискового выражения пользователя формируют индексную табличную строку, в которой отдельные ячейки соответствуют заранее заданному признаку, а значения, занесенные в ячейку, соответствуют наличию соответствующего признака в поисковом выражении; для индексной строки поискового выражения, определяют по крайней мере один из заранее заданных формальных признаков сходства, соответствующий поисковому выражению; предварительно отбирают документы из рабочей базы данных, индексным строкам которых соответствуют признаки сходства, соответствующие поисковому выражению, и представляют пользователю документы, отобранные решающей системой из документов, предварительно отобранных из базы данных.

В одном из частных вариантов реализации производят нормализацию семантических структур естественного языка по заранее заданному алгоритму, а латентно-семантическое индексирование производят с использованием с использованием нормализованных структур.

В одном из частных вариантов реализации нормализацию структур естественного языка производят путем лемматизации.

В одном из частных вариантов реализации нормализацию структур естественного языка производят путем стемминга.

В одном из частных вариантов реализации при нормализации для нескольких заранее заданных синонимов используют один заранее заданный общий термин.

В одном из частных вариантов реализации в качестве заранее заданной процедуры сжатия матриц используют процедуру понижения ранга матрицы с использованием сингулярного разложения матрицы.

В одном из частных вариантов реализации для хранения обучающей базы данных документов используют хранилище данных с повышенной надежностью.

В одном из частных вариантов реализации в качестве заранее заданного признака сходства используется значение хеш-функции, чувствительной к местоположению, в качестве аргументов которой используют строки матрицы.

Техническая задача

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.

Краткое описание чертежей

Чертеж показывает пример компьютерной системы общего назначения, которая может быть использована при создании и конфигурировании отдельных элементов системы, таких как сервера обработки и хранения данных, а также модулей системы.

Описание вариантов осуществления

Объекты и признаки настоящего изобретения, способы для достижения этих объектов и признаков станут очевидными посредством отсылки к примерным вариантам осуществления. Однако настоящее изобретение не ограничивается примерными вариантами осуществления, раскрытыми ниже, оно может воплощаться в различных видах, соответствующих сущности изобретения, раскрытого в формуле.

Используемые в настоящем описании изобретения термины "компонент", "элемент", "система", "модуль", "часть", в частности, "составная часть", "блок" и подобные используются для обозначения компьютерных сущностей (например, объектов, связанных с компьютером, вычислительных сущностей), которые могут являться аппаратным обеспечением, в частности, оборудованием (например, устройством, инструментом, аппаратом, аппаратурой, составной частью устройства, в частности, процессором, микропроцессором, печатной платой и т.д.), программным обеспечением (например, исполняемым программным кодом, скомпилированным приложением, программным модулем, частью программного обеспечения и/или кода и т.д.) или микропрограммой (прошивкой/firmware). Так, например, компонент может быть процессом, выполняющемся/исполняющимся на процессоре, процессором, объектом, исполняемым файлом, программой, функцией, методом, библиотекой, подпрограммой и/или вычислительным устройством (например, микрокомпьютером или компьютером) или комбинацией программного или аппаратного обеспечения. В качестве иллюстрации: как приложение, запущенное на сервере, в частности, на центральном сервере, может быть компонентом или модулем, так и сервер может быть компонентом или модулем. По крайней мере один компонент может находиться (располагаться) внутри процесса. Компонент может располагаться на одном вычислительном устройстве (например, компьютере) и/или может быть распределен между двумя и более вычислительными устройствами. Так, например, в частном случае приложение (компонент), может быть представлено серверным компонентом (серверной частью) и клиентским компонентом (клиентской частью). В частном случае, клиентский компонент устанавливается по крайней мере на одно вычислительное устройство, а серверный компонент устанавливается на второе вычислительное устройство, с которого, в частном случае, производится управление и/или настройка первого вычислительного устройства (и/или его составных компонентов/частей. В частном случае, модуль (и т.д.) может быть реализован (являться) одним или набором файлов, в том числе исполняемым файлом (файлами), который в свою очередь может быть связан по крайней мере с одной программной библиотекой, например, реализованной в виде dll-файла, являющегося скомпилированной формой библиотеки динамической компоновки (от англ. Dynamic Link Library), а также по крайней мере с одним файлом, например, содержащим служебные данные приложения, метаданные приложения, данные, необходимые для функционирования приложения, и/или сервисом (локальным и/или удаленным, например, веб-сервисом/веб-службой), включая приложения и сервисы, построенные на сервис-ориентированной архитектуре (от англ. service-oriented architecture/SOA), включая, но не ограничиваясь, технологии REST (от англ. Representational State Transfer -"передача состояния представления"), удаленный вызов процедур (от англ. Remote Procedure Call, RPC) и другие.

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.

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. 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 те (или только те) 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 осуществляться 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.

В рамках реализации настоящего изобретения осуществляется формирование обучающей базы данных для документов (обучающей базы данных документов), содержащей документы, для которых получены отзывы пользователей. 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 распределенное хранения в соответствии с признаком "расположение", соответствующему расположению вакансии или желаемому месту работы, указанному в резюме.

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 пользователей о соответствии 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, поля документов баз данных индексируют 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, используемого при индексации. То есть все признаки, используемые для индексации, распределены по номерам ячеек, а значения, занесенные в ячейки табличных строк, соответствуют наличию признаков, например, терминов естественного языка, 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.

Формализация и нормализация признаков производится, 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.

Для совокупностей вакансий и резюме или для всей совокупности данных, формируется исходная матрица данных, каждая из строк которых является полем вакансии или резюме. В частном случае, формируются матрицы отдельные для резюме и вакансий. В другом частном случае, поля имеют заранее заданные свойства, и матрицы формируются для полей с общими свойствами. Например, поля, относящиеся к учебным заведениям, в которых было получено образование, а также регион, в котором может быть предоставлено место работы, обрабатываются отдельно. Исходя из сущности последующих операций, для сокращения объема вычислений, группировка полей по свойствам предпочтительна перед обработкой матриц. В частном случае, матрицы географических мест работы имеют низкий ранг и могут использоваться при предоставлении результатов запроса без дополнительной обработки. Например, если соискатель указывает географическую локацию, в которой он может осуществлять профессиональную деятельность, другие локации не принимаются во внимание. В данном случае используется определение точного соответствия, не требующее сложной обработки данных. Аналогичным образом может быть осуществлена проверка на требование получения образования в определенном учебном заведении.

При проведении латентно-семантического анализа и латентно-сематической индексации могут использоваться операции формирования и использования решающих деревьев, в том числе, с построением систем решающих деревьев с использованием градиентного бустинга. Для формирования решающих деревьев с использованием градиентного бустинга могут быть применены способы обработки данных, описанные и использованные в библиотеке программного обеспечения XGBoost.

Перед проведением латентно-семантического индексирования может быть произведено исправление опечаток и очевидных ошибок, связанных, например, с ошибками при автозамене слов на компьютере пользователя.

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. В связи с этим, при реализации изобретения используются процедуры формальной обработки данных, обеспечивающие сокращение количества одновременно обрабатываемых записей при проведении латентно-семантического анализа.

Для целей последующей обработки совокупности строк полей представляются в виде двумерного массива данных, аналогичного матрице, где строки являются строками полей, имеющих уникальный идентификатор, а столбцы формируются сведениями, относящимися к одинаковым признакам строк полей. В дальнейшем, для упрощения понимания текста описания двумерные массивы данных будут именоваться "матрицами".

В частности, для сокращения количества данных, подлежащих обработке, на первом этапе, понижаются ранги матриц, используемые в системе при реализации способа. Понижение ранга матриц реализует сжатие данных с частичной потерей данных, но без существенного ущерба для результатов поиска, осуществляемого в дальнейшем.

В частном случае реализации изобретения для матриц предварительно задают ранг, до которого осуществляют понижение ранга. Ранг может быть задан эмпирически, например, путем проверки качества поиска до и после понижения ранга матриц. Ранг матрицы с пониженным рангом может быть определен по результатам машинного обучения, например, путем определения качества поисков после применения сингулярного разложения для различных заданных значений ранга матрицы с пониженным рангом.

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

В частном случае реализации изобретения, для понижения ранга применяется технология понижения ранга матрицы с использованием сингулярного разложения или "rank-k SVD".

По результатам понижения ранга с использованием сингулярного разложения, значения признаков, указанные в полях, изменяются. При этом, если начальные значения равнялись нулям и единицами, то после понижения ранга значения признаков могут становиться произвольным и выраженными рациональными числами. В связи с этим, задача латентно-семантического анализа, для выявления полей, соответствующих запросу, существенно усложняется. Несмотря на существенное сокращение объема обрабатываемых данных, упрощение процесса поиска не происходит, в связи с перехода от дискретных значений к непрерывным.

Исходно признаки или термы имеют два дискретных значения. Для функции понижения ранга может задаваться степень дискретности, вплоть до непрерывной функции, что также существенно улучшает результаты последующего вычисления LSH.

В другом частном случае происходит округление результатов сингулярного разложения с сохранением дискретности значений признаков, вплоть до сохранения значений "0" и "1".

Дополнительно, из матрицы с пониженным рангом удаляют столбцы, в которых для всех полей записано одно и то же значение, либо столбцы, не влияющие существенно на результаты поиска. Столбцы, которые могут быть удалены при понижении ранга, также могут быть определены с использованием технологии понижения ранга матрицы с использованием сингулярного разложения. В результате формируется матрица или формируются матрицы с пониженной размерностью. Кроме этого может осуществляться консолидация сходных полей, строки признаков которых принимаются тождественными друг другу. В таких матрицах каждая из совокупностей тождественных строк представлена одной строкой, которой приведены в соответствие исходные строки, преобразованные в соответствующие тождественные строки.

На следующем этапе для полей, характеризующих матрицу с пониженным рангом, осуществляется группировка полей по значениям координат векторов, являющихся векторной суммой значений векторов отдельных признаков. Для целей упрощения последующего изложения, векторная сумма значений векторов отдельных признаков будет обозначаться "вектором поля".

Группировка полей 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.

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. Предпочтительно, чтобы исчерпы-

вающий набор значений набор значений хэш-функции составлял 1/100-1/50 от значения ранга матрицы с пониженным рангом.

В частном случае реализации хэш-функция может быть задана в табличном виде.

Для вычисления LSH также могут использоваться алгоритмы MinHash, SimHash, Метод случайных проекций.

Снижение размерности матрицы признаков уменьшает объем 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, по сравнению с тем, как если бы использовались исходные, разреженные строки или поля признаков.

При определении LSH для строк матрицы, свернутых с помощью SVD, тождественные значения хэшей LSH точнее характеризуют похожесть исходных текстов для полей, чем если бы хэши LSH считались по не свернутым векторам термов. Таким образом, последовательная группировка признаков сначала с использованием SVD, а затем с использованием LSH дает более высокую релевантность результатов группировки сходных полей. Преобразование данных в обратной последовательности, то есть, при первоначальном применении LSH, а затем SVD повышает ресурсоемкость вычислений и снижает релевантность результатов. В частности, хеширование разреженных данных не обеспечивает точное определение строк, релевантных друг к другу, а понижение ранга и размерности матрицы для вычисленных хэшей еще более снижает релевантность преобразований.

Таким образом, после вычисления значений хэшей LSH и группировки полей документов по значениям хэшей формируются обучающие последовательности, каждая из которых характеризуется одним из значений хэш-функции, чувствительной к местоположению. При необходимости, например, при малом количестве полей в одной группе, первоначально определенные группы могут быть объединены.

В качестве обучающих данных для каждого из полей используется реакция пользователей на предоставленные им результаты работы поисковой системы. 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, то трудоустройства кандидата, разместившего резюме, с промежуточными градациями "резюме просмотрено" и кандидат приглашен на собеседование. 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 или несколько решающих деревьев, сформированных путем градиентного бустинга. Сформированные деревья могут использоваться совместно при поиске вакансии, даже при анализе группы вакансий, не связанной с группой вакансий, соответствующей обучающей выборке.

Поля рабочей базы данных также преобразуются 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 LSH, 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.

Чертеж показывает 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 съемный магнитный диск 29, и накопитель 30 на оптическом диске для чтения с или записи на съемный 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. 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, съемный магнитный диск 29 и съемный 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 жестком 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 часто посредством интерфейса 46 последовательного порта, который связан с системной шиной, но могут быть соединены посредством других интерфейсов, таких как параллельный порт, игровой порт или универсальная последовательная шина (УПШ). 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.

В заключение следует отметить, что приведенные в описании сведения являются примерами, которые не ограничивают объем настоящего изобретения, определенного формулой. Специалисту в данной области становится понятным, что могут существовать и другие варианты осуществления настоящего изобретения, согласующиеся с сущностью и объемом настоящего изобретения.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Способ автоматизированного поиска релевантных документов для рекомендательной системы подбора персонала с использованием машинного обучения и с понижением размерности многомерных данных, заключающийся в том, что

формируют рабочую базу 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;

формируют обучающую базу данных для документов, содержащую документы, для которых получены отзывы пользователей, характеризующие релевантность документов поисковым пользовательским запросам;

обновляют обучающую базу данных путем добавления документов, для которых получены новые отзывы пользователей о соответствии документов пользовательским запросам;

для каждого из документов индексируют поля путем преобразования сведений индексируемого поля в индексную табличную строку, отдельные ячейки которой соответствуют одному из заранее заданных признаков, соответствующему положению ячейки в индексной табличной строке, а значения, занесенные в ячейку, соответствуют наличию в сведениях индексируемого поля термина естественного языка, соответствующего ячейке;

формируют обучающую индексную матрицу, в качестве строк которой используют индексные строки, соответствующие документам обучающей базы данных для документов;

с использованием заранее заданной процедуры сжатия матриц сжимают обучающую индексную матрицу с частичной потерей данных с понижением размерности с формированием сжатой индексной матрицы, где в качестве заранее заданной процедуры сжатия матриц используют процедуру понижения ранга матрицы с использованием сингулярного разложения матрицы;

группируют строки сжатой индексной матрицы по заранее заданным формальным признакам сходства, где документы, соответствующие сгруппированным строкам в одной группе, составляют обучающую группу документов, а в качестве заранее заданного признака сходства используется значение хеш-функции, чувствительной к местоположению, в качестве аргументов которой используют строки матрицы;

формируют решающую систему, для обучения которой в качестве исходных данных используют индексные строки документов соответствующей обучающей группы документов, а в качестве ожидаемых результатов - отзывы пользователей для документов из обучающей группы, с формированием решающих деревьев, сформированных путем градиентного бустинга; и

для поискового выражения пользователя формируют индексную табличную строку, в которой отдельные ячейки соответствуют заранее заданному признаку, а значения, занесенные в ячейку, соответствуют наличию соответствующего признака в поисковом выражении;

для индексной строки поискового выражения определяют по крайней мере один из заранее заданных формальных признаков сходства, соответствующий поисковому выражению;

предварительно отбирают документы из рабочей базы данных, индексным строкам которых соответствуют признаки сходства, соответствующие поисковому выражению; и

представляют пользователю документы, отобранные решающей системой, с использованием сформированных, при обучении, решающих деревьев, из документов, предварительно отобранных из базы данных.

2. Способ по п.1, отличающийся тем, что производят нормализацию семантических структур естественного языка по заранее заданному алгоритму, а латентно-семантическое индексирование производят с использованием нормализованных структур.

3. Способ по п.2, отличающийся тем, что нормализацию структур естественного языка производят путем лемматизации.

4. Способ по п.2, отличающийся тем, что нормализацию структур естественного языка производят путем стемминга.

5. Способ по п.2, отличающийся тем, что при нормализации для нескольких заранее заданных синонимов используют один заранее заданный общий термин.

6. Способ по п.1, отличающийся тем, что для хранения обучающей базы данных для документов используют хранилище данных с повышенной надежностью.

7. Рекомендательная система подбора персонала с использованием машинного обучения и с понижением размерности многомерных данных, содержащая

блок предоставления пользователям документов, соответствующих пользовательским запросам;

блок сбора отзывов пользователей в отношении документов, предоставленных блоком предоставления пользователям документов;

рабочую базу 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, где рабочая база данных для документов выполнена с возможностью обновления путем удаления устаревших документов и добавления новых документов;

обучающую базу данных для документов, содержащую документы, для которых получены отзывы пользователей, характеризующие релевантность документов поисковым пользовательским запросам, где обучающая база данных выполнена с возможностью добавления документов, для которых получены новые отзывы пользователей о соответствии документов пользовательским запросам;

систему индексации, выполненную индексирующей поля, формирующей для каждого из документов путем преобразования сведений индексируемого поля в индексную табличную строку, отдельные ячейки которой соответствуют одному из заранее заданных признаков, соответствующему положению ячейки в индексной табличной строке, а значения, занесенные в ячейку, соответствуют наличию в сведениях индексируемого поля термина естественного языка, соответствующего ячейке;

подсистему подготовки данных для решающей системы, выполненную

формирующей обучающую индексную матрицу, в качестве строк которой используют индексные строки, соответствующие документам обучающей базы данных для документов;

с использованием заранее заданной процедуры сжатия матриц, сжимающей обучающую индексную матрицу с частичной потерей данных с понижением размерности с формированием сжатой индексной матрицы, где в качестве заранее заданной процедуры сжатия матриц используют процедуру понижения ранга матрицы с использованием сингулярного разложения матрицы;

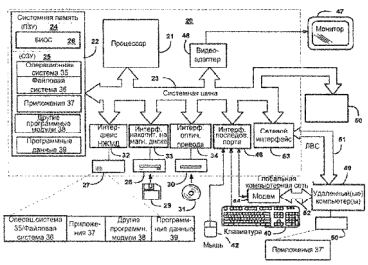
группирующей строки сжатой индексной матрицы по заранее заданным формальным признакам сходства, с формированием обучающей группы документов, содержащей документы, соответствующие сгруппированным строкам в одной группе, с формированием решающих деревьев, сформированных путем градиентного бустинга; и

обучаемую решающую систему, выполненную использующей для обучения индексные строки документов соответствующей обучающей группы документов в качестве исходных данных, и отзывы пользователей для документов из обучающей группы в качестве ожидаемых результатов; и

блок обработки поискового выражения, формирующий для поискового выражения пользователя индексную табличную строку, в которой отдельные ячейки соответствуют заранее заданному признаку, а значения, занесенные в ячейку, соответствуют наличию соответствующего признака в поисковом выражении;

блок определения признака сходства, выполненный определяющим, для индексной строки поискового выражения, по крайней мере один из заранее заданных формальных признаков сходства, соответствующий поисковому выражению, а в качестве заранее заданного признака сходства используется значение хеш-функции, чувствительной к местоположению, в качестве аргументов которой используют строки матрицы;

блок предварительного отбора документов из рабочей базы данных, индексным строкам которых соответствуют признаки сходства, соответствующие поисковому выражению, а блок предоставления пользователям документов выполнен представляющим пользователю документы, отобранные решающей системой, с использованием сформированных, при обучении, решающих деревьев, из документов, предварительно отобранных из базы данных.

8. Рекомендательная система по п.7, отличающаяся тем, что выполнена с возможностью нормализации семантических структур естественного языка по заранее заданному алгоритму и также выполнена с возможностью проведения латентно-семантического индексирования с использованием нормализованных структур.

9. Рекомендательная система по п.7, содержащая хранилище данных с повышенной надежностью, предназначенное для хранения обучающей базы данных для документов.