(19) Евразийское патентное ведомство

(11) 041000 (13) B1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОМУ ПАТЕНТУ

(45) Дата публикации и выдачи патента
2022.08.29

(21) Номер заявки
202092854

(22) Дата подачи заявки
2020.12.23

(51) Int. Cl. *G06F 16/21* (2019.01)

(54) СПОСОБ И СИСТЕМА ФОРМИРОВАНИЯ СТРУКТУРЫ ДАННЫХ НА ОСНОВЕ МНОГОСЛОЙНОГО ГРАФА

(31) **2020115241**
(32) **2020.04.27**
(33) **RU**
(43) **2021.10.29**

(71)(73) Заявитель и патентовладелец:
ПУБЛИЧНОЕ АКЦИОНЕРНОЕ ОБЩЕСТВО "СБЕРБАНК РОССИИ" (ПАО СБЕРБАНК) (RU)

(72) Изобретатель:
Шарануца Виталий Алексеевич, Булавин Алексей Александрович (RU)

(74) Представитель:
Герасин Б.В. (RU)

(56) US-A1-20170221240
US-A1-20170236225
US-B1-7512612
RU-C2-2399948

(57) Заявленное техническое решение относится к области компьютерной техники, в частности к способам организации структур хранения данных с помощью многослойных графов для связей информационных сущностей. Техническим результатом является повышение эффективности хранения данных с помощью формирования структуры хранения на основании многослойного графа связей между информационными сущностями. Заявленный результат достигается за счет способа формирования структуры хранения данных, выполняемого с помощью процессора вычислительного устройства и содержащего этапы, на которых формируют многослойный граф на основании данных, получаемых из различных источников информации, в котором вершинами каждого слоя является информация, относящаяся к одной сущности, а ребрами - данные, связывающие сущности в каждом слое графа, причем каждый слой графа представляет собой связанный граф, сформированный по данным отдельного информационного источника; осуществляют сопоставление вершин в каждом слое графа для выявления ребер, связывающих сущности в разных слоях многослойного графа; определяют вершины, содержащие информацию об одной сущности в по меньшей мере двух слоях многослойного графа; и выполняют формирование суперузла с помощью объединения упомянутых вершин, причем упомянутый суперузел содержит связанную информацию об одной сущности из по меньшей мере двух информационных источников.

041000 B1

041000 B1

Область техники

Заявленное техническое решение относится к области компьютерной техники, в частности к способам организации структур хранения данных с помощью многослойных графов для связей информационных сущностей.

Уровень техники

Для анализа графов социальных отношений между людьми и организациями и применения графовых алгоритмов данные, хранящиеся в различных информационных системах, преобразуют в графовую структуру, где узлы - это объекты или субъекты, а ребра - это отношения между этими объектами или субъектами. Зачастую информация об одних и тех же объектах или субъектах хранится в разных информационных системах, но в каждой конкретной системе хранится информация, рассказывающая о своем типе взаимоотношений между людьми и организациями. Данные в различных системах имеют разную природу, качество, полноту.

Чем более полную картину связанности между людьми и организациями мы можем получить, тем более точно будут срабатывать гипотезы, графовые алгоритмы, предсказания. Возникает задача построения единого графа клиентских связей, объединяющего все возможные источники данных. Объединение данных из различных источников в одну золотую запись может вызывать проблемы, связанные с трудностями матчинга (сопоставления) сущностей (профилей людей и организаций), проблемы определения уникального идентификатора, проблемы дедубликации записей, противоречивости информации в разных источниках, проблемы организации доступа к данным при дальнейшем использовании, проблемы ресурсозатратности пересчета единого графа больших данных. Все эти проблемы в совокупности приводят к снижению эффективности хранения данных с помощью графовых структур и их применения для целей поиска информации.

Из уровня техники известен способ организации структуры хранения данных с помощью графовой модели (патент US 10062128 B2, 28.08.2018), который включает в себя объединение информации из различных источников данных, в которых информация объединяется по принципу наличия схожих идентификаторов информационных сущностей для обеспечения процесса обхода вершин графа для целей доступа к связанной информационной сущности. Данное решение позволяет сформировать информационную структуру, объединяющую информацию из различных информационных источников, однако имеет существенный недостаток в части отсутствия объединения сущностей по их связанности с целью формирования суперузлов, объединяющих связи между радом сущностей, что позволяет обеспечить более эффективный процесс использования графовой модели построения структуры данных.

Известно решение по нахождению оптимального маршрута обхода графа для доступа к требуемой информационной сущности (патент US 7512612 B1, 31.03.2009), которое включает в себя принцип анализа релевантности узлов и соединяющих их ребер для построения маршрута обхода множества связанных вершин. Данное решение также не предусматривает объединение вершин графа в суперузел для формирования более эффективного принципа формирования структуры хранения данных и работы с ней для связей информационных сущностей.

Объединение данных из нескольких источников в один граф часто проблематично, потому что нет единого ID сущностей для двух источников. Данные в источнике могут иметь дубликаты, намеренно ложные записи, ошибки ввода, ошибки транспортирования и конвертации данных, потерянные записи и т.п.

При этом сущность в том или ином источнике может не иметь своего ID и быть представлена совокупностью записей, относящихся к совершенно другой сущности. При матчинге (сопоставлении) сущностей нескольких источников в "золотую запись", если нет явного ключа, то составной ключ складывается из большого числа атрибутов (жесткий ключ), чтобы не получить ошибок первого рода, когда сопоставляются сущности, которые не должны сопоставляться. При матчинге по жесткому ключу ошибки первого рода уменьшаются, но ошибки второго рода, которые возникают в том случае, когда не происходит матчинг сущностей, которые подлежат матчингу, увеличиваются и результат количественно мал.

Если в нескольких системах, сливаемых в "золотую запись", есть один и тот же атрибут сущности, то возникают коллизии выбора информации, так как эта информация может оказаться разной. Некоторые атрибуты могут быть использованы в качестве критерия связанности; таким образом, при слиянии можно получить, например, два противоположных взаимоисключающих друг друга факта о связи сущностей, формирующих узлы графа. Информация о сущностях со временем может пополняться дополнительными данными в источниках. При слиянии источников в "золотую запись" изменение информации может приводить к изменениям матчинга и объединению двух ранее раздельных "золотых записей" и, наоборот, разделению объединенных. После слияния источников в "золотую запись" затрудняется поиск первопричины ошибки в конечных данных, поскольку связь с исходной информацией можно восстановить только по логам операций слияния и дедубликации. При этом данные после слияния из нескольких источников становятся новым источником и управлять доступом к ним исходя из природы данных этих источников уже невозможно.

Раскрытие изобретения

Для решения существующих технических проблем в данной области предлагается новый эффек-

тивный способ организации структуры хранения данных с помощью графов с объединением схожих информационных сущностей в суперузлы.

Техническим результатом является повышение эффективности хранения данных с помощью формирования структуры хранения на основании многослойного графа связей между информационными сущностями.

Дополнительным эффектом является повышение скорости поиска данных с помощью объединения связанных сущностей в суперузлы связанных записей.

Заявленный результат достигается за счет способа формирования структуры хранения данных, выполняемого с помощью процессора вычислительного устройства и содержащего этапы, на которых

формируют многослойный граф на основании данных, получаемых из различных источников информации, в котором вершинами каждого слоя является информация, относящаяся к одной сущности, а ребрами - данные, связывающие сущности в каждом слое графа, причем каждый слой графа представляет собой связанный граф, сформированный по данным отдельного информационного источника;

осуществляют сопоставление вершин в каждом слое графа для выявления ребер, связывающих сущности в разных слоях многослойного графа;

определяют вершины, содержащие информацию об одной сущности в по меньшей мере двух слоях многослойного графа; и

выполняют формирование суперузла с помощью объединения упомянутых вершин, причем упомянутый суперузел содержит связанную информацию об одной сущности из по меньшей мере двух информационных источников.

В одном из частных вариантов осуществления способа сущность содержит сведения о физическом и/или юридическом лице.

В другом частном варианте осуществления способа вершины графа содержат тематическую информацию, относящуюся к сущности.

Заявленный результат достигается также за счет системы формирования структуры хранения данных, содержащей по меньшей мере один процессор и по меньшей мере одно средство памяти, содержащее машиночитаемые инструкции, которые при их выполнении процессором осуществляют вышеуказанный способ.

Краткое описание чертежей

Фиг. 1 иллюстрирует общий вид графовой структуры многослойного хранения данных с примерами информационных сущностей.

Фиг. 2 иллюстрирует пример графовой структуры с матчингом информационных сущностей.

Фиг. 3 иллюстрирует пример формирования суперузла между слоями графа.

Фиг. 4 иллюстрирует пример форм суперузлов.

Фиг. 5 иллюстрирует пример исполнения вычислительного устройства.

Осуществление изобретения

В настоящих материалах заявки будут использованы следующие термины и определения.

Узел - минимальная совокупность информации, однозначно представляющая сущность (персону, организацию и др.) в виде записи в информационном источнике.

Ребро - минимальная совокупность информации, однозначно представляющая факт и характер связанности сущностей по одному или нескольким принципам в информационном источнике.

Атрибуты узла - совокупность тематической информации, однозначно относящаяся к конкретной сущности в информационном источнике, раскрывающая подробности по ней.

Атрибуты ребра - совокупность информации, относящаяся к конкретному ребру, раскрывающая подробности по этой связи.

Слой графа - генеральная совокупность узлов и ребер, образующих связанный граф на данных одного источника.

Матчинг слоев - ребро особого рода, соединяющее узлы в разных слоях графа при условии, что узлы относятся к одной и той же сущности реального мира.

Суперузел - подмножество узлов в многослойном графе, относящихся к одной и той же сущности реального мира, объединенных матчингами.

Многослойный граф - граф связей, состоящий из нескольких слоев, в котором узлы из разных слоев объединены в суперузлы при помощи матчингов.

На фиг. 1 представлен пример структуры хранения данных, организованной с помощью многослойного графа (10). Каждый слой (110)-(160) графа (10) представляет собой источник данных и представляет собой отдельный граф со связанными сущностями. Каждый из слоев графа (10) содержит, как правило, информацию, тематически связанную с различными сущностями, например персонами и/или организациями, а также связи между ними. Сущности представляют собой узлы (вершины) в каждом слое графа (10), ребра графа представляют собой связи между узлами.

Как представлено на примере фиг. 1, слой (110) может представлять собой данные из источника, содержащего информацию о персонах, осуществляющих взаимодействие между собой, например транзакции или иную финансовую активность. Слой (120) может содержать информацию о территориальном

размещении персон. Слой (130) может представлять данные, отображающие родственные связи между физическими лицами и т.п.

Предлагаемый подход в формировании структуры хранения данных направлен на создание единой графовой модели для поиска и управления информацией о связях физических и/или юридических лиц между собой при использовании данных из множества раздельных источников (110)-(160). В каждом слое графа (10) сущности, относящиеся к персоне или организации, имеют один или несколько атрибутов, позволяющих осуществить их полную или частичную идентификацию. В качестве таких атрибутов могут выступать: данные банковского счета, ФИО персон, наименование организаций, ИНН, адрес, паспортные данные, ОГРН/ОГРНИП и т.п.

Каждая вершина в каждом слое (110)-(160) графа (10) анализируется на предмет наличия связанной информации о по меньшей мере одном отношении в другом одном или нескольких слоях графа. Для этого может применяться один из алгоритмов матчинга сущностей в графовой структуре (10) (см., например, https://habr.com/m/company/sberbank/blog/354564/).

С помощью формирования графовой структуры (10) из нескольких отдельных графов реализуется возможность выполнения графовых операций над любым слоем (110)-(160) из единой точки входа. Для того чтобы графовые операции обхода графа можно было выполнять над единым графом связанных сущностей, а не только над отдельными слоями, в графовую базу хранения единого графа добавляются дополнительные ребра-матчинги, которые соединяют слои (110)-(160) посредством связанных узлов, относящихся к одной и той же сущности реального мира.

На фиг. 2 представлен пример формируемых ребер-матчингов, которые связывают сущности в различных слоях графа (10). Между слоями (140)-(150) графа формируется первое ребро-матчинг, которое связывает узлы (142)-(153), содержащие сведения об одной сущности. Второе ребро-матчинг формируется между узлами (153)-(163). При этом все узлы (142), (153), (163) связаны с одной сущностью, например персоной или юридическим лицом. При выполнении графовых операций обхода графа (10) необходимо учитывать особое свойство ребра-матчинга, считать его длиной 0, которое по сути является технической связкой многослойного графа (10). Как было указано выше, в каждом слое графа (10) сущности, относящиеся к персоне или организации, имеют один или несколько атрибутов, позволяющих осуществить их полную или частичную идентификацию. Данные атрибуты применяются для матчинга узлов в слоях графа. Таким атрибутами могут выступать данные банковского счета, ФИО персон, наименования организаций, ИНН, адрес, паспортные данные, ОГРН/ОГРНИП и другая информация, по которой возможно осуществить полную или частичную идентификацию информационной сущности.

Как представлено на фиг. 3, несколько матчингов, объединяющих узлы из разных слоев, относящиеся к одной и той же сущности реального мира, образуют подграф суперузла (20), который обеспечивает построение единого объекта графа (10), организованного связями соответствующих узлов, относящихся к одной сущности. Многослойный граф (10) позволяет получить значительно более богатый по составу информации граф связей, описывающий взаимодействие субъектов или объектов, чем любой граф отдельного источника.

Например, связь показанная на фиг. 2 и образующая суперузел (20) на фиг. 3, может представлять собой следующее: узел (142) слоя (140) с данными, содержащими значения в полях по клиенту А, по которому имеется информация в вершине слоя графа типа <фамилия><имя><отчество><дата рождения><название города проживания>, будет соединен связью матчинг 1 с узлом (153) другого слоя (150), в котором содержатся данные с такими же значениями в сходных полях: <фамилия><имя><отчество><дата рождения><название города проживания>, а узел (153) источника (150), в свою очередь, будет соединен связью типа матчинг 2 с узлом (163) источника (160) при точном совпадении данных других полей, например <имя><отчество><серия и номер паспорта> и т.п.

Данные полей перед сопоставлением необходимо нормализовать, т.е. привести в унифицированный вид. К таким полям могут относиться, например, ФИО, пол, телефон, адрес, дата, email, серийный номер, регистрационный номер, реквизиты компании.

Приведенный пример не является исчерпывающим и может иметь иную структуру полей нормализации или других полей для описания параметров информационных сущностей.

За счет применения механизма формирования суперузлов (20) появляется возможность объединения равнозначных по отношению друг к другу узлов в одном или различных слоях графа (10) без необходимости выделенного явного центра. Объединение узлов происходит при помощи добавления ребер связей матчингов к другим связям, что позволяет использовать естественные механизмы графовых хранилищ для извлечения подграфа суперузла (20). При использовании любого графового алгоритма на графах с суперузлами (20) обработке будет подлежать атомарный шаг хода такого алгоритма с учетом того, что ребро-матчинг - это ребро длиной 0. В качестве графовых алгоритмов могут использоваться, например, кратчайший путь (англ. shortest path), алгоритм выделения кластеров или сообществ (англ. label propagation), алгоритм ссылочного ранжирования (например, pagerank), степень посредничества (англ. betweenness centrality), степень близости (англ. closeness centrality) и т.п.

Как показано на фиг. 4, подграф матчингов суперузла (20) может быть как в форме клики (связь все со всеми), так и любой другой формы, например кольца, двойного кольца и т.п. Форма клики является

наиболее предпочтительной, так как разрыв любого матчинга на разрушает суперузел (20), но в случае если слоев много, то матчингов в суперузле (20) с кликой будет (n×(n-1))/2. Матчингов может оказаться слишком много, что, в свою очередь, может повлиять на скорость обработки графа; в этом случае предпочтительнее использовать суперузел с матчингами в форме одинарного или двойного кольца.

Связь матчинга в суперузле (20) можно разорвать, например, в следствие изменения, уточнения данных или ограничения прав доступа к данным. Если количество связей матчингов в суперузле (20) изначально мало, например, из-за невысокого качества данных источников или невозможности получить требуемые составные ключи, то разрыв определенного матчинга может приводить к разрушению суперузла (20) на два отдельных суперузла-осколка.

Управление такой структурой графа (10) может обеспечиваться доступом к данным на уровне слоев, относящихся к источнику, что является более эффективным в части управления процессом доступа к структуре данных, а также повышает скорость выявления требуемой информации по той или иной сущности. При регулярном перестроении многослойный граф (10) позволяет более эффективно расходовать ресурсы, так как каждый слой графа может обновляться отдельно по своему расписанию, а также нет затрат на создание и поддержание "золотой записи".

На фиг. 5 представлен пример общего вида вычислительного устройства (200), на базе которого может быть реализована одна или несколько автоматизированных систем, обеспечивающих реализацию заявленного способа формирования структуры хранения данных. В общем случае устройство (200) содержит объединенные общей шиной информационного обмена (210) один или несколько процессоров (201), средства памяти, такие как ОЗУ (202) и ПЗУ (203), интерфейсы ввода/вывода (204), устройства ввода/вывода (205) и устройство для сетевого взаимодействия (206). В общем случае устройство (200) может представлять собой сервер, серверный кластер, мейнфрейм, суперкомпьютер или иной тип пригодного вычислительного устройства. Процессор (201) (или несколько процессоров, многоядерный процессор и т.п.) может выбираться из ассортимента устройств, широко применяемых в настоящее время, например, таких производителей, как Intel™, AMD™, Apple™, Samsung Exynos™, MediaTEK™, Qualcomm Snapdragon™ и т.п. Под процессором или одним из используемых процессоров в устройстве (200) также необходимо учитывать графический процессор, например GPU NVIDIA или Graphcore, тип которых также является пригодным для полного или частичного выполнения заявленного способа, а также может применяться для обучения и применения моделей машинного обучения в различных информационных системах.

ОЗУ (202) представляет собой оперативную память и предназначено для хранения исполняемых процессором (201) машиночитаемых инструкций для выполнение необходимых операций по логической обработке данных. ОЗУ (202), как правило, содержит исполняемые инструкции операционной системы и соответствующих программных компонент (приложения, программные модули и т.п.). При этом в качестве ОЗУ (202) может выступать доступный объем памяти графической карты или графического процессора.

ПЗУ (203) представляет собой одно или более устройств постоянного хранения данных, например жесткий диск (HDD), твердотельный накопитель данных (SSD), флэш-память (EEPROM, NAND и т.п.), оптические носители информации (CD-R/RW, DVD-R/RW, BlueRay Disc, MD) и др.

Для организации работы компонентов устройства (200) и организации работы внешних подключаемых устройств применяются различные виды интерфейсов В/В (204). Выбор соответствующих интерфейсов зависит от конкретного исполнения вычислительного устройства, которое может представлять собой, не ограничиваясь, PCI, AGP, PS/2, IrDa, FireWire, LPT, COM, SATA, IDE, Lightning, USB (2.0, 3.0, 3.1, micro, mini, type C), TRS/Audio jack (2.5, 3.5, 6.35), HDMI, DVI, VGA, Display Port, RJ45, RS232 и т.п. Для обеспечения взаимодействия пользователя с устройством (200) применяются различные средства (205) В/В информации, например клавиатура, дисплей (монитор), сенсорный дисплей, тачпад, джойстик, манипулятор, мышь, световое перо, стилус, сенсорная панель, трекбол, динамики, микрофон, средства дополненной реальности, оптические сенсоры, планшет, световые индикаторы, проектор, камера, средства биометрической идентификации (сканер сетчатки глаза, сканер отпечатков пальцев, модуль распознавания голоса) и т.п.

Средство сетевого взаимодействия (206) обеспечивает передачу данных посредством внутренней или внешней вычислительной сети, например Интранет, Интернет, ЛВС и т.п. В качестве одного или более средств (206) может использоваться без ограничения Ethernet карта, GSM модем, GPRS модем, LTE модем, 5G модем, модуль спутниковой связи, NFC модуль, Bluetooth и/или BLE модуль, Wi-Fi модуль и др. Представленные материалы заявки раскрывают предпочтительные примеры реализации технического решения и не должны трактоваться как ограничивающие иные, частные примеры его воплощения, не выходящие за пределы испрашиваемой правовой охраны, которые являются очевидными для специалистов соответствующей области техники.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Способ формирования структуры хранения данных, выполняемый с помощью процессора вычислительного устройства и содержащий этапы, на которых

формируют многослойный граф на основании данных, получаемых из различных источников информации, в котором вершинами каждого слоя является информация, относящаяся к одной сущности, а ребрами - данные, связывающие сущности в каждом слое графа, причем каждый слой графа представляет собой связанный граф, сформированный по данным отдельного информационного источника;

осуществляют сопоставление вершин в каждом слое графа для выявления ребер, связывающих сущности в разных слоях многослойного графа;

определяют вершины, содержащие информацию об одной сущности в по меньшей мере двух слоях многослойного графа; и

выполняют формирование суперузла с помощью объединения упомянутых вершин, причем упомянутый суперузел содержит связанную информацию об одной сущности из по меньшей мере двух информационных источников.

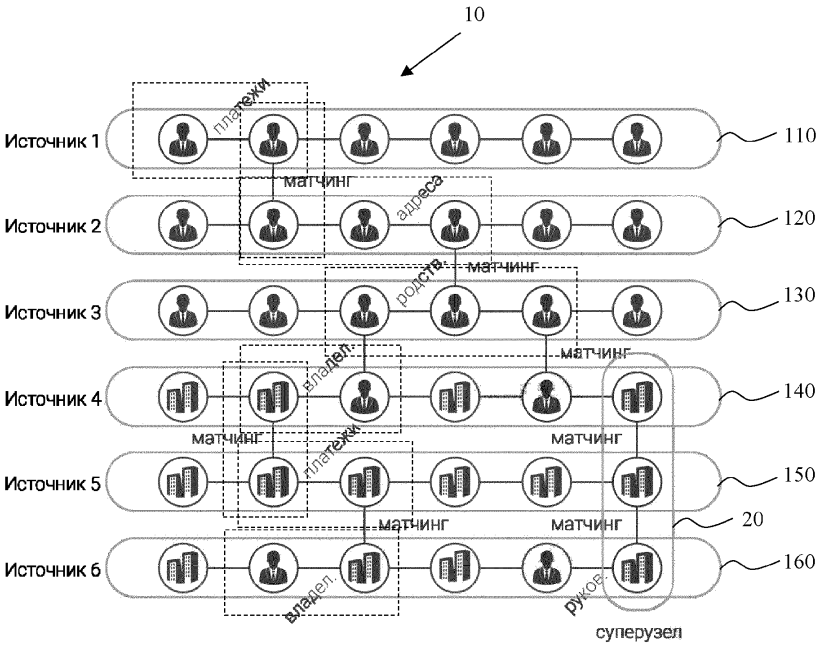
2. Способ по п.1, характеризующийся тем, что сущность содержит сведения о физическом и/или юридическом лице.

3. Способ по п.1, характеризующийся тем, что вершины графа содержат тематическую информацию, относящуюся к сущности.

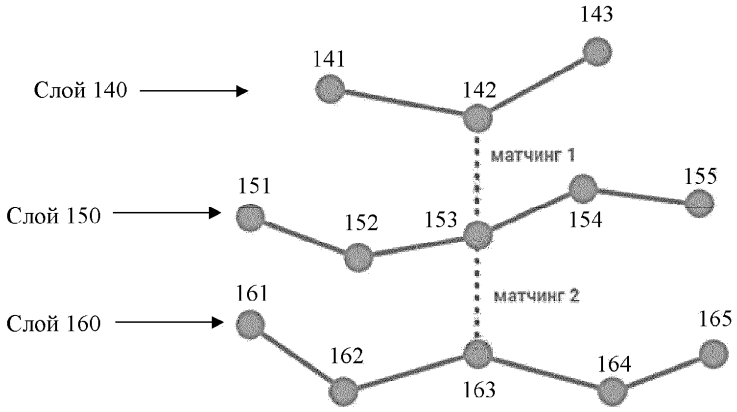
4. Способ по п.1, характеризующийся тем, что суперузлы формируются на основе ребер-матчингов, связывающих вершины графа в одном или различных слоях графа.

5. Способ по п.1, характеризующийся тем, что форма суперузла может представлять собой по меньшей мере одно из клика, кольца или двойного кольца.

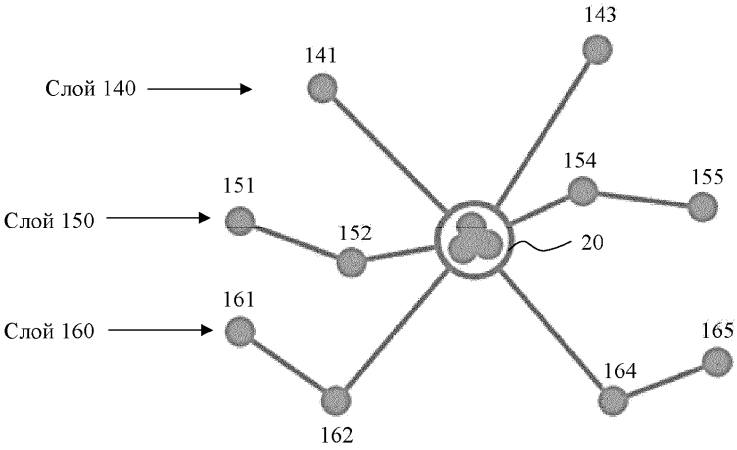
6. Система формирования структуры хранения данных, содержащая по меньшей мере один процессор и по меньшей мере одно средство памяти, содержащее машиночитаемые инструкции, которые при их выполнении процессором осуществляют способ по любому из пп.1-5.



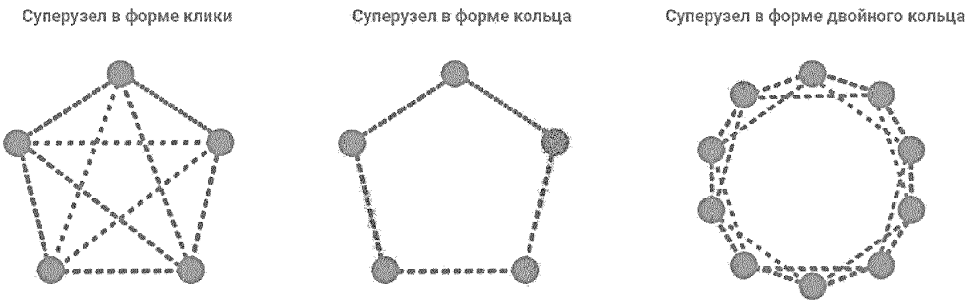
Фиг. 1



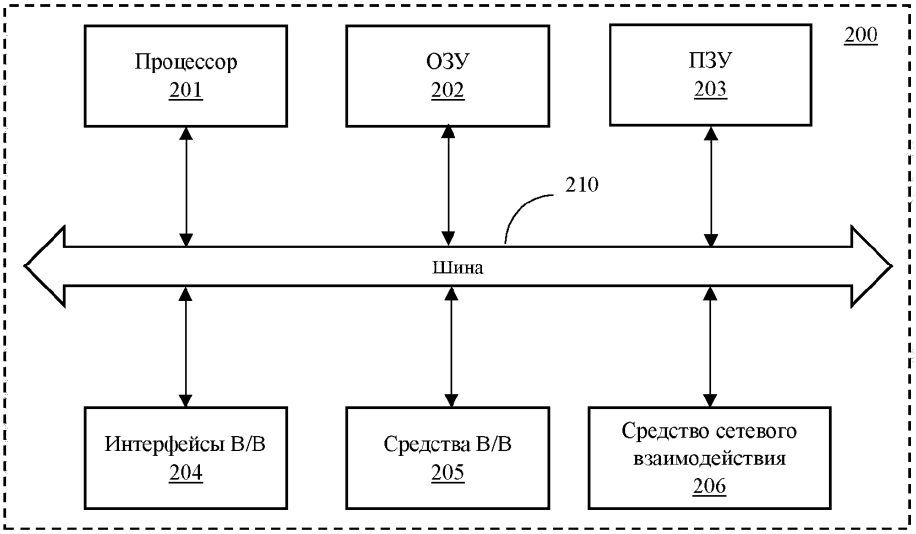
Фиг. 2



Фиг. 3



Фиг. 4


Процессор
201
ОЗУ
202
ПЗУ
203
200
210
Шина
Интерфейсы В/В
204
Средства В/В
205
Средство сетевого взаимодействия
206

Фиг. 5