**БЕЗОПАСНОСТЬ ЦИФРОВЫХ ТЕХНОЛОГИЙ. – 2022. – № 4 (107).** – 9–26

*ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И ТЕЛЕКОММУНИКАЦИИ*

УДК 004.056 DOI: 10.17212/2782-2230-2022-4-9-26

# АЛГОРИТМЫ, МЕТОДЫ И ПОДХОДЫ К ОБЕЗЛИЧИВАНИЮ И ОБОГАЩЕНИЮ ДАННЫХ, В ТОМ ЧИСЛЕ ПЕРСОНАЛЬНЫХ *

А.А. МАЛЯВКО[1], В.В. РЕУТОВ[2], И.В. КОРОТКИХ[3], В.К. ШПЕРЛИНГ[4]

[1] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, кандидат технических наук, доцент, доцент кафедры вычислительной техники. E-mail: a.malyavko@corp.nstu.ru*

[2] *630008, РФ, г. Новосибирск, ул. Бориса Богаткова, 63/1, ООО «Системы информационной безопасности», начальник коммерческого отдела. E-mail: rvv@sib-nsk.net*

[3] *109129, РФ, г. Москва, ул. 8-я Текстильщиков, 11/2, МОО «Ассоциация руководителей служб информационной безопасности», руководитель регионального отделения. E-mail: nsk@aciso.ru*

[4] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, магистрант кафедры вычислительной техники. E-mail: shperling.2017@stud.nstu.ru*

В настоящей статье представлены результаты исследования алгоритмов, методов и подходов к обезличиванию и обогащению данных, в том числе и персональных данных. Среди типов обогащения данных были рассмотрены обогащения демографических, географических и поведенческих данных, а также изучены структурный, статистический, семантический и прагматический алгоритмы обогащения данных. Также в работе было рассмотрено обогащение онтологий, а именно: выразительные онтологии, легкие онтологии, такие как таксономии, и т. д. Обогащение онтологий – это обширная область исследований, в которой можно выделить три категории работ, посвященных извлечению семантических знаний из разнородных данных. В результате проведенного анализа было выяснено, что процессы обогащения данных оптимизируют продажи и уменьшают затраты бизнеса. Были представлены преимущества и недостатки рассмотренных подходов и методов обогащения данных и онтологий. Основным преимуществом обогащения является повышение ценности и точности информации, помогающее компаниям принимать важные бизнес-решения. Основным недостатком является риск нарастания избыточных данных, что может привести к неправильной аналитике и, соответственно, к неправильным бизнес-решениям, что, в свою очередь, вредит бизнесу. Также пред-

* Статья получена 17 августа 2022 г.

ставлена значимость проведенного анализа: на основе проведенных исследований предполагается формирование технического предложения для создания базовой инфраструктуры проекта ЦК НТИ «Доверенная среда обмена информаций» для проведения дальнейших исследований на тему обогащения и обезличивания данных.

**Ключевые слова:** обогащение данных, обезличивание данных, персональные данные, большие данные, обезличивание, онтологии, обогащение онтологий, методы обогащения

## ВВЕДЕНИЕ

В литературе упоминается три измерения, характеризующие большие данные, которые стали отраслевым стандартом для определения больших данных [1, 2]. Во-первых, объем данных [3] (их количество); во-вторых, скорость изменения данных; в-третьих, разнообразие данных. Датчики, Интернет вещей (IoT), записи баз данных, видео и аудио имеют разные форматы и стандарты. Эти измерения, которые в основном носят технический характер, впоследствии были дополнены некоторыми дополнительными соображениями. Отсутствие необходимого управления и однородности определяется достоверностью. Валидность связана с правильностью и точностью, а волатильность относится к тому, как долго данные действительны и как долго они должны храниться [3–5].

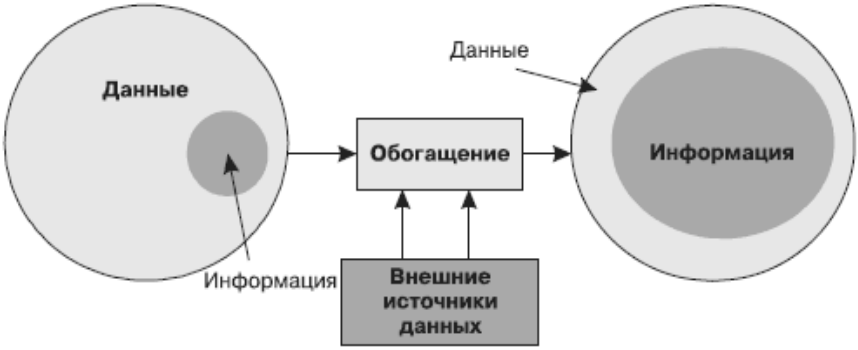
Объектом исследования в настоящей работе являются алгоритмы, методы и подходы к обезличиванию и обогащению данных, в том числе персональных.

Цель работы – исследование алгоритмов, методов и подходов к обезличиванию и обогащению данных с выявлением граничных условий применения и характеристик получаемого результата.

## 1. ОБОГАЩЕНИЕ ДАННЫХ

### 1.1. ТИПЫ ОБОГАЩЕНИЯ ДАННЫХ

Обогащение данных улучшает данные с помощью различных средств (см. рисунок ниже). Существует столько же типов обогащения данных, сколько и источников данных, но компании часто используют несколько распространенных видов. К ним можно отнести описанные далее виды.




Одна из возможных схем обогащения данных

One of the possible data enrichment schemes

Обогащение демографических данных расширяет наборы данных о клиентах за счет применения демографической информации, такой как семейное положение, размер семьи, уровень дохода, кредитный рейтинг и многое другое. Этот тип информации обеспечивает большую персонализацию ваших критериев таргетинга, месседжей и креативов.

Обогащение географических данных включает добавление в профили клиентов такой информации, как почтовые индексы, картографические адреса, координаты и многое другое. Этот тип данных особенно полезен для мобильной рекламы и для определения местоположения новых магазинов. Его также можно использовать для определения локальных цен.

Обогащение поведенческих данных применяет модели поведения клиентов к их профилям, включая их прошлые покупки и поведение при просмотре. Это часто связано с отслеживанием покупательского пути пользователя, чтобы определить ключевые области интересов каждого покупателя. Поведенческие данные необходимы компаниям для определения того, какие рекламные компании работают лучше всего и какова будет рентабельность инвестиций каждой компании.

Каждый тип обогащения данных помогает компании достигать различных бизнес-целей. Прежде чем выбрать правильный метод обогащения данных для вашего бизнеса, определите, какая именно информация вам нужна.

Основное преимущество, обеспечиваемое обогащением данных, – это повышенная ценность и точность понимания клиентов компании. Компаниям нужны высококачественные данные, чтобы принимать важные бизнес-решения и делать ценные выводы.

Оценка и анализ видов данных помогает отделам продаж расставить приоритеты в своих усилиях. Только это практически невозможно сделать, когда у вас есть неполные профили клиентов. Обогащение данных может улучшить их профили с помощью качественных данных, обеспечивая надежную и значимую оценку. Качество и глубина данных также могут позволить автоматизировать оценку лидов, исключить предположения и позволить вашей команде продаж сосредоточиться на своих целях.

Процессы обогащения данных обеспечивают соблюдение вашей компанией нормативных требований, касающихся конфиденциальности данных. Многие страны на законодательном уровне устанавливают ограничения на тип клиентских данных, которые вы можете хранить, и на время их сбережения. Есть также необходимость регулярно вести «do-not-call lists» или списки людей, которые не желают, чтобы их данные как-то использовались. Если у вашей компании нет механизма, обеспечивающего соблюдение нормативных требований, вы можете столкнуться с дорогостоящими штрафами.

С другой стороны, вы можете настроить data enrichment процессы для регулярной потоковой очистки базы данных. При этом можно сохранить ценность данных и соответствие справочным требованиям.

Неточные данные могут привести к потере рекламных бюджетов, недовольству клиентов и неправильной аналитике, что дорого обходится компаниям. Многие работают с избыточными данными, потому что либо не знают о их существовании, либо не знают, какие данные следует удалить.

Инструменты обогащения данных могут устранить избыточные и неточные данные. Это происходит за счет автоматического анализа информации, объединения избыточностей и исправления ошибок при сохранении обновленных профилей. Этот метод повышает качество данных компании.

Обогащение данных снижает затраты и оптимизирует продажи. Процессы обогащения данных экономят компании деньги за счет управления существующей информацией – ее объемом и хранением. Этот процесс также снижает затраты за счет минимизации штрафов из-за несоответствия данных. Одновременно с этим обогащение данных обеспечивает максимальную прибыль за счет увеличения продаж. Это происходит благодаря более эффективному маркетингу и клиентскому менеджменту. Data enrichment может определять возможности перекрестных и дополнительных продаж, продвигая при этом конструктивные отношения с клиентами.

Для уменьшения избыточности используется процесс обогащения информации. Далее рассмотрим методы обогащения информации.

## 1.2. МЕТОДЫ ОБОГАЩЕНИЯ ДАННЫХ

Структурное обогащение предполагает изменение параметров сообщения, отображающего информацию, в зависимости от частотного спектра исследуемого процесса, скорости обслуживания источников информации и требуемой точности.

При статистическом обогащении происходит накопление статистических данных и обработка выборок из общего набора накопленных данных.

Семантическое обогащение предполагает минимизацию логической формы, исчисления и высказываний, выделение и классификацию понятий и содержания информации, переход от частных понятий к более общим.

Прагматическое обогащение предполагает выделение из полученной информации наиболее ценной, наиболее соответствующей целям и задачам пользователя.

В настоящее время, в условиях становления информационного общества, публикация данных различных предприятий и организаций в открытом доступе становится всё более масштабной и значимой. Одним из инструментов для этого является технология связанных открытых данных, развивающаяся в рамках концепции семантической паутины. Связанные открытые данные (Linked Open Data, LOD) – это опубликованные структурированные данные, каждый элемент которых имеет собственный URI, представлен в виде структуры описания ресурсов (RDF) и имеет связь с другими данными. Технология Semantic Web послужила основой для создания сети данных, в которой узлы соответствуют интересующим ресурсам в предметной области, а ребра соответствуют связям между ними. Поскольку все ресурсы представлены URI, создается огромная распределенная сеть наборов данных. Приложения могут динамически обнаруживать эти наборы данных, получать к ним доступ, интерпретировать их с использованием связанных метаданных, представленных в виде онтологий, и интегрировать их в свои операции. Инициатива Linked Open Data (LOD), основанная на стандартах Semantic Web, привела к созданию огромного веб-корпуса наборов данных в различных предметных областях (доменах). Большинство этих данных относится к типу конкретных объектов (например, Москва – столица России). Существует острая необходимость в дополнении наборов данных утверждениями, связывающими понятия более высокого уровня, как показано выше. Добавление утверждений такого рода является частью задачи по обогащению наборов данных LOD, называемой «обогащением онтологий» [6].

## 2. ТИПЫ ОБОГАЩЕНИЯ ОНТОЛОГИЙ

Обогащение онтологии – это широкая область исследований, которую можно разделить на три категории работ, посвященных извлечению семантических знаний из разнородных данных. Это могут быть структурированные данные (данные в базах данных) или неструктурированные данные (тексты на естественном языке), а также полуструктурированные данные (документы HTML) [7, 8].

Первая категория касается выразительных онтологий и генерации определений понятий. Дело в том, что большинство методов создания онтологий направлено на создание довольно невыразительных онтологий (таксономий и отношений), но многие приложения в различных областях требуют гораздо более сложной аксиоматизации. Существует несколько подходов к автоматической генерации таких выразительных онтологий. Некоторые подходы работают с текстами, описывающими понятия. Например, Lexo [9] применяет правила преобразования синтаксиса к определениям естественного языка для создания аксиом в логике описания (DL). В работе [10] используется подход извлечения отношений, который основывается на введении формальных ограничений для обеспечения качества результирующих определений [11]. Другие подходы строятся на индуктивном логическом программировании [12] для поиска новых логических описаний понятий из утверждений онтологии.

Вторая категория работ посвящена созданию легких онтологий, таких как таксономии. Они позволяют извлекать различные онтологические элементы из текстовых ресурсов [13]. Для извлечения понятия ключевым шагом является извлечение соответствующей терминологии предметной области [14]. Затем применяются методы классификации для обнаружения синонимов, и для каждой группы похожих терминов может быть получен соответствующий онтологический класс.

К третьей категории относятся работы, в которых рассуждения частично заменяют традиционные методы извлечения знаний. В них понятия делятся на примитивные и составные, причем последние определяются из первых. Примитивные понятия формируются с помощью стандартных средств извлечения знаний. Составные понятия формулируются на основе извлеченных свойств и примеров примитивных понятий.

Таким образом, современное состояние показывает, что ни один из подходов, взятых отдельно, не является решением общей проблемы обогащения онтологий. Одним из наиболее интересных и прогрессивных с этой точки зрения является комбинированный подход для обогащения онтологий из текстовых и открытых данных [15]. Особенность этого подхода заключается в том,

что он решает тройную задачу: 1) понятия, используемые для разметки, не имеют прямой терминологии в документах; 2) их формальные определения изначально неизвестны; 3) информация, полезная для разметки документов, необязательно упоминается в них. Для решения этих проблем используется существующая онтология предметной области, которая обогащается определениями понятий, используемых для последующей разметки. Для построения этих определений создается и затем используется набор документов с ручной разметкой, используемых в качестве примеров. Онтология заполнена информацией, извлеченной из этих документов, и информацией, поступающей из внешних ресурсов (связанных открытых данных). Определения, которые необходимо получить, могут затем быть сформированы на основе этой заполненной онтологии и набора помеченных документов. Эти определения затем добавляются к онтологии (обогащение онтологии). Следовательно, всякий раз, когда новые документы той же предметной области должны быть размечены, онтология может заполняться одинаково и применяются определения, позволяющие помечать новые документы. Этот подход, получивший название SAUPODOC, является новым подходом к заполнению и обогащению онтологий, использующим основы семантической сети, методы анализа текста, извлечения связанных открытых данных, машинного обучения и инструментов логического вывода.

Многие методы исследования баз данных о различных болезнях, в частности о раке, основаны на алгоритмах построения «пропозиционально подобных» моделей. Существенным ограничением ряда алгоритмов анализа является отсутствие или серьезная непонятность таких моделей. Для улучшения восприятия и производительности алгоритмов анализа и прогнозирования развития болезней в работе [16] построены понятные модели с использованием алгоритмов Graph Mining и Inductive Logic Programming. Они основаны на интеллектуальном анализе данных – процессах сбора и выявления закономерностей в информации, извлекаемой из кучи необработанных данных. Когда закономерности установлены, можно определить различные отношения между наборами данных, и они могут быть представлены в обобщенном формате (деперсонализация), который помогает при проведении различных видов статистического анализа. Среди других структур данных графы широко используются при моделировании сложных структур и шаблонов. В интеллектуальном анализе данных графы используются для поиска шаблонов для различения, классификации, кластеризации, обогащения данных и т. д. Графы используются в сетевом анализе. Они могут содержать обширные данные о сетевых коммуникациях, веб- и компьютерных сетях, социальных сетях и т. д. В мультиреляционном анализе данных графы или сети используются для отображения различных взаимосвязанных отношений между наборами дан-

ных в реляционной базе данных. Если необработанные данные представлены в виде некоторого множества графов, то к ним применимы методы графического майнинга или интеллектуального анализа [17]. Майнинг графов – это процесс, в котором методы интеллектуального анализа используются для поиска закономерностей или взаимосвязей в данном реальном наборе графов. Изучая граф, можно выявить часто встречающиеся подструктуры и взаимосвязи, что помогает группировать наборы графов, находить взаимосвязь между наборами графов или различать или характеризовать графы. Прогнозирование этих тенденций формирования шаблонов может помочь в построении моделей для улучшения любого приложения, которое используется в режиме реального времени. Основной задачей майнинга графов можно считать обнаружение часто встречающихся подграфов.

Пусть имеется граф h с множеством ребер E(h) и множеством вершин V(h) и множество графов данных H, для которого вводится понятие поддержки s(h), означающей процент графов в H, подграфом которых является граф h. Частый граф (паттерн) имеет поддержку, которая будет не меньше минимального заданного порога поддержки. Обозначим этот порог как min_support. Для нахождения частых графов нужно выполнить следующие шаги:

– создание частых кандидатов в подструктуры (частое основание);

– вычисление поддержки каждого кандидата.

Необходимо оптимизировать первый шаг, потому что второй шаг представляет собой NP-полное множество, где вычислительная сложность очень высока.

Существует два метода разработки частого основания.

Подход, основанный на поиске частых графов, который начинается с формирования некоторого известного графа небольшого размера. Алгоритм продвигается снизу вверх, создавая новых кандидатов путем добавления вершин или ребер и проверяя величину их поддержки. Затем найденные частые графы используются для создания следующих кандидатов путем их объединения. Этот шаг для создания частых кандидатов в графовом виде является сложным процессом в отличие от создания кандидатов в неструктурированном наборе элементов. Для структурированных отображений в виде графов существует не один метод соединения двух подструктур. Чаще всего используется так называемый метод BFS (поиск в ширину), основанный на построении изоморфных подграфов при объединении в них кандидатов в таковые, а затем добавлении к ним оставшихся неизморфных частей.

Подход с ростом шаблона: этот подход может использовать как BFS, так и DFS (поиск в глубину), однако для него предпочтительнее использовать именно DFS, обеспечивающий меньшее потребление памяти. Суть метода состоит в том, что из графа h можно построить новый граф, добавив ребро.

При этом может быть добавлена и вершина, но это необязательно. Процесс роста подграфа прост, но не очень эффективен, потому что всегда есть возможность создания аналогичного уже созданному подграфу, что требует дополнительных трудоемких проверок. Сгенерированные повторяющиеся графы нужно удалять. Чтобы избежать создания повторяющихся графов, частые графы следует вводить очень осторожно и консервативно, что вызывает потребность в других алгоритмах.

В работе [18] рассматривается детально задача выявления часто встречающихся подсетей, известная как майнинг частых подграфов для обработки сетей коэкспрессии над набором генов. Максимально частые подграфы – это репрезентативный набор частых подграфов. Частый подграф максимален, если он не имеет суперграфа, который является частым. Максимально частые подграфы можно использовать для обнаружения важных сетевых свойств, объясняющих сложные взаимодействия между генами. Эти взаимодействия «регистрируются» на краях частых подсетей. Дальнейшее изучение частых подсетей коэкспрессии улучшает обнаружение биологических модулей и биологических сигнатур для экспрессии генов и классификации болезней. Предлагается алгоритм обратного поиска под названием RASMA для извлечения часто встречающихся данных и максимально частых подграфов в данном наборе графов. Ключевым нововведением в RASMA является перечислитель связанных подграфов, который использует стратегию обратного поиска для перечисления связанных подграфов неориентированного графа. Используя эту стратегию перечисления, RASMA очень эффективно получает все максимально частые подграфы. Чтобы преодолеть трудоемкую вычислительную задачу перечисления всех часто встречающихся подграфов при поиске максимально часто встречающихся подграфов, RASMA использует несколько стратегий сокращения, которые существенно улучшают общую производительность во время выполнения. Экспериментальные результаты показывают, что в больших сетях коэкспрессии генов предложенный алгоритм эффективно извлекает биологически релевантные максимально частые подграфы. Извлечение повторяющихся подсетей коэкспрессии генов из нескольких экспериментов по экспрессии генов позволяет обнаруживать функциональные модули и биомаркеры подсетей. Анализ обогащения извлеченных максимально частых подсетей показывает, что частые подсети сильно обогащены известными биологическими онтологиями.

Далее с использованием графического майнинга можно решать множество задач, в том числе:

- извлечение биохимических структур;
- поиск биологически консервативных подсетей;

- поиск функциональных модулей;
- анализ потока управления программой;
- сетевой анализ вторжений;
- анализ структур сетей связи;
- обнаружение аномалий;
- анализ XML-структур.

Графы знаний RDF (или наборы данных) содержат ценную информацию, которую можно использовать для решения множества реальных задач. Однако из-за огромного размера доступных наборов данных RDF трудно найти наиболее ценные наборы данных для конкретной задачи. Для улучшения возможности обнаружения, связывания и повторного использования наборов данных существует тенденция к построению систем поиска наборов данных. Такие системы в основном основаны на метаданных и игнорируют содержимое, однако в задачах, связанных с интеграцией и обогащением данных, приходится учитывать содержимое наборов данных. Это важно для интеграции данных, а также для обогащения данных. Например, часто владельцы наборов данных хотят обогатить их содержимое. Для этого выбирают дополнительный набор, который предоставляет полную информацию для исходного набора данных. Вышеупомянутые задачи требуют метрик объединения и дополнения на основе контента между любым подмножеством наборов данных, однако часто такие подходы не приводят к желаемому результату. Чтобы сделать возможным вычисление таких метрик в очень больших масштабах, в работе [19] предлагается подход, основанный на а) наборе предварительно созданных (и периодически обновляемых) семантических индексов и б) инкрементных алгоритмах.

В концепции сетевого анализа отношения между единицами называются связями в графе. С точки зрения интеллектуального анализа данных это называется анализом ссылок. Сеть представляет собой диверсионный набор данных с многореляционной концепцией в виде графа. Граф обычно очень велик, с узлами в виде объектов и ребрами в виде связей, которые, в свою очередь, обозначают взаимосвязь между узлами или объектами. Системы телефонных сетей, WWW (Всемирная паутина), социальные сети [20] являются очень хорошими примерами подобных графов и дают представление о масштабе проблем, возникающих при анализе их поведения. Методы, предлагаемые в [20, 21], позволяют фильтровать наборы данных и предоставлять наиболее предпочтительные для клиентов услуги. Изучение и извлечение полезной информации из графов, представляющих структуру и поведение разнообразных сетей, может помочь в организации их эффективной работы.

Динамические графы обычно используются для описания структуры связей, которая изменяется со временем. Несмотря на то что было проведено большое количество исследовательских работ по пониманию динамических сетей с использованием прогнозирования ссылок, классификации узлов и обнаружения сообществ, существует очень мало работ, специально посвященных решению проблемы обработки динамических сетей большого размера. В работе [21] изучается постоянно возникающая сложная проблема укрупнения или огрубления графов связей в динамических сетях. Огрубление сети относится к классу сетевых операций «уменьшения масштаба», при выполнении которых некоторые пары узлов и/или ребер графа группируются вместе для эффективного анализа больших сетей. Известные ранее подходы к укрупнению сети могли работать только со статическими сетями, где веса сетевой структуры были предварительно определены до расчета укрупнения. Принимая во внимание, что большие сети очень динамичны и естественным образом меняются со временем, в этой работе рассматривается возможность встраивания в исследуемые графы данных о распространении информации, которые отражают динамику сетей для укрупнения сети. В частности, предлагается новый подход Semi-NetCoarsen, который совместно максимизирует вероятность обнаружения данных о распространении информации и минимизирует регуляризацию графа по отношению к предопределенным структурным данным сети. Функция обучения является выпуклой, используется алгоритм ускоренного проксимального градиента для получения глобального оптимального решения.

Существует ряд традиционных методов машинного обучения, в которых берутся однородные объекты из одного отношения. Но в больших сетях такой подход неприменим из-за их многореляционной гетерогенной природы [22]. Ссылки – это отношения между узлами в сети. Интеллектуальный анализ ссылок появился как новая область исследований и представляет собой конвергенцию исследований, проведенных в области анализа графов, сетей, гипертекстов, логического программирования, прогнозного анализа и моделирования. Она включает в себя следующее.

*Прогнозирование типа ссылки*. В соответствии с ресурсами задействованного объекта система анализа предсказывает мотив этой связи. В организациях это помогает предлагать интерактивные сеансы общения между сотрудниками, если это необходимо.

*Прогнозирование типа объекта*. Здесь предсказание основывается на типе задействованного объекта, его атрибутах и свойствах, связях и признаках объекта, связанного с ним. Например, в домене ресторана аналогичный метод используется, чтобы предсказать: предпочитает ли клиент заказывать еду на дом или непосредственно посещать ресторан? Это также помогает в прогно-

зировании способа общения, который предпочитает клиент (по телефону или по почте).

*Оценка кардинальности ссылки*. В этой задаче выделяют два вида оценки. Во-первых, это прогнозирование количества ссылок, связанных с объектом. Например, процент авторитета веб-страницы можно рассчитать, найдя количество ссылок, связанных с ней, которое называется in-links. Веб-страницы, которые выступают в качестве концентратора (это означает, что набор веб-страниц имеет другие ссылки, которые попадают под ту же тему), могут быть идентифицированы с помощью внешних ссылок.

*Прогнозирование существования связи*. Здесь система предсказывает, существует ли связь между двумя объектами. Например, эта задача используется для прогнозирования наличия связи между двумя веб-страницами.

*Сверка объектов*. В этом методе функция состоит в том, чтобы предсказать, являются ли любые два объекта одинаковыми на основе их атрибутов, признаков или связей. Этот метод также называется неопределенностью идентичности или связыванием записей. Эта задача имеет ту же процедуру в сопоставлении цитирования, извлечении деталей, избавлении от дубликатов, консолидации объектов.

Одним из примеров отечественных исследований в области применения интеллектуального анализа графов является работа [23]. В ней на основе использования библиотеки networkx для Python-программ изучаются способы эффективного выявления взаимосвязей между клиентами банка с целью улучшения процессов их обслуживания.

## ЗАКЛЮЧЕНИЕ

В работе представлены преимущества и недостатки отдельных подходов и методов к обогащению данных. Приведены примеры, в том числе оригинальные, по использованию свойств данных как инструментов обогащения. Показаны структуры организации данных, позволяющие обеспечивать обогащение данных путем «майнинга» знаний. На текущий момент интеллектуальный анализ данных является самым перспективным направлением для обогащения данных путем генерации новых знаний через выявление скрытых связей между различными базами данных.

Дальнейший этап работы заключается в формировании требований к базовой инфраструктуре проекта, которая послужит для проведения исследований по обогащению данных.

## СПИСОК ЛИТЕРАТУРЫ

1. *Laney D.* 3D data management: controlling data volume, velocity, and variety // Application Delivery Strategies. – Meta Group, 2001. – Fail 949. – P. 1–4.

2. *Supriya M., Chattu V.K.* A review of artificial intelligence, big data, and blockchain technology applications in medicine and global health // Big Data and Cognitive Computing. – 2021. – Vol. 5, iss. 41. – P. 41. – DOI: 10.3390/bdcc5030041.

3. *Khan M.A.-u.-d., Uddin M.F., Gupta N.* Seven V's of Big Data understanding Big Data to extract value // Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education. – IEEE, 2014. – DOI: 10.1109/ASEEZone1.2014.6820689.

4. *Zhang L.* A framework to model big data driven complex cyber physical control systems // 20th International Conference on Automation and Computing. – IEEE, 2014. – P. 283–288. – DOI: 10.1109/IConAC.2014.6935501.

5. *Provost F, Fawcett T.* Data science and its relationship to big data and data-driven decision making // Big Data. – 2013. – Vol. 1 (1). – P. 51–59. – DOI: 10.1089/big.2013.1508.

6. *Романов С.В., Сытник А.А., Шульга Т.Э.* О возможностях использования коммуникативных грамматик и LSPL-шаблонов для автоматического построения онтологий // Известия Самарского научного центра Российской академии наук. – 2015. – Т. 17, № 2 (5). – С. 1104–1108.

7. Разработка методов дискретного анализа семантики слабоструктурированных систем: отчет о НИР / Саратовский государственный технический университет имени Гагарина Ю.А.; А.А. Сытник, Н.И. Вагарина, Н.И. Мельникова и др. – № ГР 01201459267. – Саратов, 2015.

8. Semantic marking method for non-text documents of website based on their context in hypertext clustering / S. Papshev, A. Sytnik, N. Melnikova, A. Bogomolov // Recent Research in Control Engineering and Decision Making. ICIT 2019. – Cham: Springer, 2019. – P. 313–323. – (Studies in Systems, Decision and Control; vol. 199). – DOI: 10.1007/978-3-030-12072-6_26.

9. *Völker J., Hitzler P., Cimiano P.* Acquisition of OWL DL Axioms from Lexical Resources // 4th European Semantic Web Conference (ESWC), Innsbruck, Austria. – Springer, 2007. – P. 670–685.

10. *Ma Y., Distel F.* Concept adjustment for description logics // 7th International Conference on Knowledge Capture, K-CAP' 13, Banff, Canada. – ACM, 2013. – P. 65–72.

11. *Ma Y., Distel F.* Learning formal definitions for snomed CT from text // Proceedings of Artificial Intelligence in Medicine (AIME), Murcia, Spain. – Springer, 2013. – P. 73–77.

12. *Chitsaz M.* Enriching ontologies through data // Doctoral Consortium co-located with International Semantic Web Conference (ISWC). – Sydney, Australia, 2013. – P. 1–8.

13. *Cimiano P.* Ontology learning and population from text: algorithms, evaluation and applications. – New York: Springer, 2006. – 347 p.

14. *Cimiano P., Völker J., Studer R.* Ontologies on demand? – A description of the state-of- the-art, applications, challenges and trends for ontology learning from text // Information, Wissenschaft und Praxis. – 2006. – Vol. 57 (6–7). – P. 315–320.

15. *Alec C., Reynaud-Delaître C., Safara B.* A combined approach for ontology enrichment from textual and open data // Advances in Knowledge Discovery and Management. – Cham: Springer, 2018. – P. 1–21. – (Studies in Computational Intelligence; vol. 732).

16. *Ferreira P., Ladeiras J., Camacho R.* Assessing the impact of data set enrichment to improve drug sensitivity in cancer // Practical Applications of Computational Biology and Bioinformatics, 15th International Conference (PACBB 2021). – Cham: Springer, 2021. – P. 74–84. – DOI: 10.1007/978-3-030-86258-9_8.

17. *Borgwardt K., Stegle O.* An introduction to Graph Mining. – URL: https://ethz.ch/content/dam/ethz/special-interest/bsse/borgwardt-lab/documents/slides/CA10_GraphMining.pdf (accessed: 07.12.2022).

18. *Saeed S., Alokshiya M., Hasan M.A.* RASMA: a reverse search algorithm for mining maximal frequent subgraphs // BioData Mining. – 2021. – Vol. 14. – Art. 19. – DOI: 10.1186/s13040-021-00250-1.

19. *Mountantonakis M., Tzitzikas Y.* Content-based union and complement metrics for dataset search over RDF knowledge graphs // Journal of Data and Information Quality. – 2020. – Vol. 12 (2). – Art. 10. – P. 1–31. – DOI: 10.1145/3372750.

20. *Tang L., Liu H.* Graph mining applications to social network analysis // Managing and Mining Graph Data. – Boston, MA: Springer, 2010. – P. 487–513. – DOI: 10.1007/978-1-4419-6045-0_16.

21. Towards embedding information diffusion data for understanding big dynamic networks / H. Yang, P. Zhang, H. Wang, C. Zhou, Z. Li, L. Gao, Q. Tan // Neurocomputing. – 2021. – Vol. 466. – P. 265–284. – DOI: 10.1016/j.neucom.2021.09.024.

22. Link mining: models, algorithms, and applications / P.S. Yu, J. Han, C. Faloutsos, eds. – New York: Springer, 2010. – 586 p. – DOI: 10.1007/978-1-4419-6515-8.

23. *Носов Р., Курносов А.* Применение технологии graph mining в аудите банковских транзакций. – URL: https://newtechaudit.ru/graph-mining-audit/? (дата обращения: 08.12.2022).

***Малявко Александр Антонович***, кандидат технических наук, доцент, доцент кафедры вычислительной техники Новосибирского государственного технического университета. Область научных интересов – параллельные высокопроизводительные вычисления, нейронные сети. E-mail: a.malyavko@corp.nstu.ru

***Реутов Владимир Владимирович***, начальник коммерческого отдела ООО «Системы информационной безопасности», общественный эксперт НТИ рынков «Хелснет», «Сейфнет», эксперт в области информационной безопасности. E-mail: rvv@sib-nsk.net

***Коротких Игорь Валерьевич***, руководитель регионального отделения МОО «Ассоциация руководителей служб информационной безопасности». E-mail: nsk@aciso.net

***Шперлинг Владимир Константинович***, магистрант кафедры вычислительной техники Новосибирского государственного технического университета. Область научных интересов – мобильная робототехника, защита информации в мобильных системах. E-mail: shperling.2017@stud.nstu.ru

DOI: 10.17212/2782-2230-2022-4-9-26

## Algorithms, methods and approaches to details and enrich data, including personal data*

**A.A. Malyavko [1], V.V. Reutov[2], I.V. Korotkikh[3], V.K. Shperling[4]**

[1] *630073, Russian Federation, Novosibirsk, Karl Marx Prospekt, 20, Novosibirsk State Technical University, PhD in Technology, Associate Professor of the Department of Computer Science. E-mail: a.malyavko@corp.nstu.ru*

[2] *630008, Russian Federation, Novosibirsk, Street Borisa Bogatkova, 63/1, Information Security Systems LLC, Head of the Commercial Department. E-mail: rvv@sib-nsk.net*

[3] *109129, Russian Federation, Moscow, Street 8th Tekstilshchikov, 11/2, International Public Organization "Association of Heads of Information Security Services", Head of the Regional Branch. E-mail: nsk@aciso.ru*

[4] *630073, Russian Federation, Novosibirsk, Karl Marx Prospekt, 20, Novosibirsk State Technical University, master student of the Department of Computer Science. E-mail: shperling.2017@stud.nstu.ru*

This article presents the results of a study of algorithms, methods and approaches to depersonalization and enrichment of data, including personal data. Among the types of data enrichment, demographic, geographic and behavioral data enrichments were considered, as well as statistical, semantic and pragmatic data enrichment algorithms were studied. In addition to

* *Received 17 August 2022.*

data enrichment, categories of ontology enrichment were considered, namely expressive ontologies, lightweight ontologies such as taxonomies, and a category that includes works that use reasoning to partially replace traditional methods of knowledge extraction. Ontology enrichment is a broad area of research that can be divided into three categories of work devoted to extracting semantic knowledge from heterogeneous data. As a result of the analysis, it was found that data enrichment processes optimize sales, as well as reduce business costs, by saving finances through information management. The advantages and disadvantages of the considered approaches and methods of data enrichment and ontologies were presented. The main benefit of fortification is the increased value and accuracy of information that helps companies make important business decisions. The main disadvantage is the risk of growing redundant data, which can lead to incorrect analytics and, accordingly, to wrong business decisions, which in turn harms the business. The significance of the analysis is also presented – on the basis of the studies carried out, it is planned to form a technical proposal for creating the basic infrastructure of the project of the NTI Central Committee "Trusted Information Exchange Environment" for further research on the topic of data enrichment and depersonalization.

**Keywords**: data enrichment, data depersonalization, personal data, big data, depersonalization, ontologies, ontology enrichment, enrichment methods

## REFERENCES

1. Laney D. 3D data management: controlling data volume, velocity, and variety. *Application Delivery Strategies*. Meta Group, 2001, fail 949, pp. 1–4.

2. Supriya M., Chattu V.K. A review of artificial intelligence, big data, and blockchain technology applications in medicine and global health. *Big Data and Cognitive Computing*, 2021, vol. 5, iss. 41, p. 41. DOI: 10.3390/bdcc5030041.

3. Khan M.A.-u.-d., Uddin M.F., Gupta N. Seven V's of Big Data understanding Big Data to extract value. *Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education*. IEEE, 2014. DOI: 10.1109/ASEEZone1.2014.6820689.

4. Zhang L. A framework to model big data driven complex cyber physical control systems. *20th International Conference on Automation and Computing*. IEEE, 2014, pp. 283–288. DOI: 10.1109/IConAC.2014.6935501.

5. Provost F, Fawcett T. Data science and its relationship to big data and data-driven decision making. *Big Data*, 2013, vol. 1 (1), pp. 51–59. DOI: 10.1089/big.2013.1508.

6. Romanov S.V., Sytnik A.A., Shulga T.E. O vozmozhnostyakh ispol'zovaniya kom-munikativnykh grammatik i LSPL-shablonov dlya avtomaticheskogo postroeniya ontologii [About using communicative grammar and lexico-syntactic patterns for automatic ontology building]. *Izvestiya Samarskogo nauchnogo tsentra Rossiiskoi akademii nauk = Proceedings of the Samara Scientific Center of the Russian Academy of Sciences*, 2015, vol. 17, no. 2 (5), pp. 1104–1108.

7. Sytnik A.A., Vagarina N.I., Mel'nikova N.I., et all. *Razrabotka metodov diskretnogo analiza semantiki slabostrukturirovannykh sistem*: otchet [Development of methods for discrete analysis of the semantics of semi-structured systems]. Yuri Gagarin State Technical University of Saratov. No. 01201459267, 2015.

8. Papshev S., Sytnik A., Melnikova N., Bogomolov A. Semantic marking method for non-text documents of website based on their context in hypertext clustering. *Recent Research in Control Engineering and Decision Making. ICIT 2019*. Cham, Springer, 2019, pp. 313–323. DOI: 10.1007/978-3-030-12072-6_26.

9. Völker J., Hitzler P., Cimiano P. Acquisition of OWL DL Axioms from Lexical Resources. *4th European Semantic Web Conference* (ESWC), Innsbruck, Austria. Springer, 2007, pp. 670–685.

10. Ma Y., Distel F. Concept Adjustment for Description Logics. *7th International Conference on Knowledge Capture, K-CAP' 13*, Banff, Canada. ACM, 2013, pp. 65–72.

11. Ma Y., Distel F. Learning formal definitions for snomed CT from text. *Proceedings of Artificial Intelligence in Medicine (AIME)*, Murcia, Spain. Springer, 2013, pp. 73–77.

12. Chitsaz M. Enriching ontologies through data. *Doctoral Consortium colocated with International Semantic Web Conference (ISWC)*, Sydney, Australia, 2013, pp. 1–8.

13. Cimiano P. *Ontology learning and population from text: algorithms, evaluation and applications*. New York, Springer, 2006. 347 p.

14. Cimiano P., Völker J., Studer R. Ontologies on demand? – A description of the state-of- the-art, applications, challenges and trends for ontology learning from text. *Information, Wissenschaft und Praxis*, 2006, vol. 57 (6–7), pp. 315–320.

15. Alec C., Reynaud-Delaître C., Safara B. A combined approach for ontology enrichment from textual and open data. *Advances in Knowledge Discovery and Management*. Cham, Springer, 2018, pp. 1–21.

16. Ferreira P., Ladeiras J., Camacho R. Assessing the impact of data set enrichment to improve drug sensitivity in cancer. *Practical Applications of Computational Biology and Bioinformatics, 15th International Conference (PACBB 2021)*. Cham, Springer, 2021, pp. 74–84. DOI: 10.1007/978-3-030-86258-9_8.

17. Borgwardt K., Stegle O. *An introduction to Graph Mining*. Available at: https://ethz.ch/content/dam/ethz/special-interest/bsse/borgwardt-lab/documents/slides/CA10_GraphMining.pdf (accessed 07.12.2022).

18. Saeed S., Alokshiya M., Hasan M.A. RASMA: a reverse search algorithm for mining maximal frequent subgraphs. *BioData Mining*, 2021, vol. 14, art. 19. DOI: 10.1186/s13040-021-00250-1.

19. Mountantonakis M., Tzitzikas Y. Content-based union and complement metrics for dataset search over RDF knowledge graphs. *Journal of Data and Information Quality*, 2020, vol. 12 (2), art. 10, pp. 1–31. DOI: 10.1145/3372750.

20. Tang L., Liu H. Graph mining applications to social network analysis. *Managing and Mining Graph Data*. Boston, MA, Springer, 2010, pp. 487–513. DOI: 10.1007/978-1-4419-6045-0_16.

21. Yang H., Zhang P., Wang H., Zhou C., Li Z., Gao L., Tan Q. Towards embedding information diffusion data for understanding big dynamic networks. *Neurocomputing*, 2021, vol. 466, pp. 265–284. DOI: 10.1016/j.neucom.2021.09.024.

22. Yu P.S., Han J., Faloutsos C., eds. Link *Mining: models, algorithms, and applications*. New York, Springer, 2010. 586 p. DOI: 10.1007/978-1-4419-6515-8.

23. Nosov R., Kurnosov A. *Primenenie tekhnologii graph mining v audite bankovskikh tranzaktsii* [Application of graph mining technology in the audit of banking transactions]. Available at: https://newtechaudit.ru/graph-mining-audit/? (accessed 08.12.2022).

Для цитирования:

Алгоритмы, методы и походы к обезличиванию и обогащению данных, в том числе персональных / А.А. Малявко, В.В. Реутов, И.В. Коротких, В.К. Шперлинг // Безопасность цифровых технологий. – 2022. – № 4 (107). – С. 9–26. – DOI: 10.17212/2782-2230-2022-4-9-26.

For citation:

Malyavko A.A., Reutov V.V., Korotkikh I.V., Shperling V.K. Algoritmy, metody i pokhody k obezlichivaniyu i obogashcheniyu dannykh, v tom chisle personal'nykh [Algorithms, methods and approaches to details and enrich data, including personal data]. *Bezopasnost' tsifrovykh tekhnologii = Digital Technology Security*, 2022, no. 4 (107), pp. 9–26. DOI: 10.17212/2782-2230-2022-4-9-26.