УДК 338.24

ИНСТРУМЕНТАРИЙ АЛГОРИТМИЗАЦИИ ОТКРЫТЫХ И BIG DATA – ДАННЫХ В ЦЕЛЯХ РАЗВИТИЯ МЕЖДУНАРОДНОГО БИЗНЕСА КОМПАНИЙ

С.А. Затонский
Аспирант ЮФУ, руководитель отдела маркетинга BODIUM,
stanislav.zatonsky@gmail.com

С.А. Затонский
Окончил Южный Федеральный Университет (ЮФУ). Аспирант ЮФУ, факульет Менеджмента. Работает в инновационном пространстве BODIUM в должности руководителя отдела маркетинга. Проводит научные исследования на тему цифровой трансформации международного бизнеса организации.

Аннотация. Благодаря интернету и распространению компьютерных технологий, международнй бизнес может анализировать открытые и большие данные в целях наиболее эффективной адаптации полученной информации, влияющей на получение коммерческой прибыли. В данном исследовании описаны преимущества использования данных в целях развития бизнеса. Проведен пошаговый анализ алгоритма работы с данными. Определены ключевые этапы работы по исследованию данных: поиск и сбор информации, основная очистка и обработка данных, приведение данных к одному виду кодировки. Определены наиболее популярные инструменты, позволяющие работать с данными на всех представленных этапах.

Ключевые слова: международный бизнес, большие данные, массивы данных, исследование данных, алгоритмизация данных

Введение.

Сегодня исследование данных является востребованным направлением по всему миру. Международный бизнес может использовать большие данные в повседневной работе для множества целей, таких как прогнозирование тенденций и рыночной конъюнктуры, пптимизация процессов производства и логистики, управление рисками, а также улучшения обслуживания клиентов и модернизации маркетинговых стратегий.

Аналитика больших данных имеет потенциал преобразовать способ взаимодействия компаний с международным бизнесом, позволяя им лучше понимать и реагировать на сложную и динамичную глобальную среду [1]. По всему миру большие данные все чаще используются в исследованиях международного бизнеса для изучения широкого спектра тем, включая глобальные цепочки добавленной стоимости, интернационализацию и кросс-культурный менеджмент.

Одной из главных польз использования Big Data для международного бизнеса является возможность получения более точной и своевременной информации о клиентах и рынке [2]. Например, анализ данных социальных сетей может помочь компаниям лучше понимать потребности и предпочтения своей ЦА, а анализ данных покупок и поведения клиентов может помочь компаниям оптимизировать свои продукты и услуги. Еще одним преимуществом использования Big Data для международного бизнеса является возможность оптимизации бизнес-процессов и сокращения расходов [3]. Например, анализ данных о производственных процессах может помочь компаниям выявить узкие места и улучшить производительность, а анализ данных о расходах может помочь компаниям сократить затраты на определенные процессы (табл. 1).

Таблица 1. Соотношение задач МБ и типов больших данных

Задачи	Типы больших данных
Прогнозирование спроса	Исторические данные о продажах, данные о клиентах, данные о производственных циклах
Улучшение маркетинговых кампаний	Данные о клиентах, социальные медиа-данные, данные о транзакциях
Оптимизация логистики	Данные о транспортировке, данные о трафике, данные GPS
Анализ рынка	Данные об экономике, данные о конкурентной среде, данные о клиентах
Улучшение качества продукции	Данные о производственных процессах, данные о качестве продукции, данные о сбоях и проблемах
Управление рисками	Данные о рисках, данные о финансовых рынках, данные о клиентах
Исследование новых рынков	Данные о рынке, данные о конкурентной среде, данные о клиентах
Развитие инноваций	Данные о рынке, данные о клиентах, данные о технологиях и инновациях

Источник: составлено автором по материалам собственного исследования

Компании малого и среднего производственного бизнеса могут встречать проблемы в работе с большими данными, так как у них может не быть достаточных ресурсов для разработки и реализации программ и технологий, способных обрабатывать и анализировать большие объемы данных.

Вот некоторые из задач, которые могут возникнуть в этой области и которые менеджмент может решить с помощью больших данных:

1. Определение целей и преимуществ использования больших данных. Менеджеры могут помочь определить, какие данные могут быть наиболее полезными для компании и как они могут использоваться для увеличения эффективности производственных процессов, оптимизации затрат и улучшения качества продукции.
2. Разработка уникальных инструментов и методов для работы с данными в целях компании. Например, использование статистических методов, машинного обучения, искусственного интеллекта и других технологий [4].
3. Оценка стоимости внедрения систем обработки данных: в рамках развитияв международного бизнеса могут быть внедрены системы обработки данных, а также оценена потенциальная коммерческая выгоду от использования этих систем.
4. Анализ данных и выявление тенденций: топ-менджемент корпораций, используя методы работы с большими данными, можем проанализировать информаций, собраннуюкомпанией, и выявить определенные тенденции и закономерности, которые могут помочь принимать более эффективные решения по управлению бизнесом.
5. Разработка стратегии безопасности данных. Крупный международный бизнес может быть заинтересован в разработке собственной стратегии «безопасности данных», которая будет обеспечивать сохранность и конфиденциальность данных компании, включая защиту от внешних угроз и внутренних нарушений безопасности.

Тем не менее, реализация потенциала использования больших данных в международном бизнесе влечет за собой ряд технических, организационных и даже иногда этических вызовов для компаний по всему миру.

Следует отметить, что поскольку суть использования информации, основанной на больших и открытых данных, заключается в познании и поиске, то изначально зафиксировать единый статичный план первоначальной идеи исследования, необходимой для бизнеса, как правило, невозможно. Однако, какой бы ни была сумбурной динамика работы с большими и открытыми данными, существуют общепринятые алгоритмы работы, подходящие для различных бизнес-целей. Так, благодаря команде дата-исследователей из ‘The Guardian’, основавшей значимый для развития исследования данных ‘Datablog’, на данный момент установлен общий план выполнения работ по исследованию данных (рисунок 1).

Рисунок 1. Визуализированный производственный процесс Guardian Datablog, 2019 [5]

Вне зависимости от целей и сектора международного бизнеса, первым этапом плана по работе с данными всегда является их поиск и сбор. Как правило, источники дифференцированы. Ими могут являться научные, медицинские или государственные базы данных, а также поисковых сетей, новостные сводки, социологические опросы, открытая информация и разнообразные исследования. Так, на данном этапе необходимо создать основную базу данных

с актуальной для бизнеса информацией, напрямую связанный с определенной актуальной бизнес-задачей.

Следующий шаг корректной работы с данными – анализ полученной базы на предмет ее корректности, целостности и упорядочиваемости. На этом этапе, как правило, проводится первичная очистка характеризующаяся отсортировкой и удалением «битых» файлов с данными. В последствии необходимо провести анализ базы на предмет информационной заполненности. Результат будет влияет на последующее базы в последствии и другими массивами данных. Вдбоавок к логическому анализу, исследователю необходимо также определить потенциально возможные взаимосвязи и корреляции данных в рамках как уже собранной базы данных, так и при ее расширении в дальнейшем.

Далее необходимо перейти к этапу работы с данными, который характеризуется очисткой основных данных и их обработкой. На данном этапе необходимо выявить и удалить лишние элементы таблиц, таких как дубликаты, ненужые столбцы, объединенные ячейки, посторонние символы и т.д. Данная работа должна выполняться постепенно в момент построения базы. Если информация в базе данных является зашифрованной, то исследователю также потребуется полная дешифровка и чистка данных. В случае необходимости, если необходимо отдельно восстанавливать и расшифровывать значения каждой ячейки в таблице базы, на данном этапе может быть привлечен отдельный дешифровшик-специалист.

При составлении общей таблицы из исследуемых источников, для объединения данных и продолжения дальнейшей работы, необходимо кодировать данные к одному виду. Так, следует провести работу и с исходными форматами данных – в случае, если данные не являются машиночитаемыми (допустим, изображениями или PDF-форматам), то необходимо использовть специализированное программное обеспечение (ПО). Как правило, подходящее под данную задачу ПО не рассчитано на работу с данными в формате таблицы – как результат, это может привести к появлению недопустимых ошибок. Следовательно, вероятно, что полученные данные в таком случае придется переносить вручную или с помощью сторонних программ, использующих технологии искуственного интелекта.

Таким образом, алгоритм работы с данными состоит из четко структурированных пошаговых этапов. Менеджмент компании, анализируя big data – данные, обязан следовать данному алгоритму, т.к., в противном случае, любая полученная информация может быть ошибочный и привести компанию к потерям и убыткам. Несмотря на близость работы с данными к деятельности программистов, в международном бизнесе навыки исследователя-данных играют первостепенную роль, ведь в центре задачи всегда стоит определенная бизнес-цель, а не просто ряд цифр.

Сбор и первичная обработка данных.

Наиболее важным и первым этапом в работе над дата-исследованием является поиск и сбор данных. Среди работающих с данными журналистами принято считать, что в зависимости от уровня технической подготовки и профессиональных навыков, исследователь имеет возможность выбрать наиболее удобный инструмент для сбора данных [6].

Традиционно первичный поиск осуществляется через наиболее попуялрные системы поиска, такие как Yandex или Google. Такие современные гибкие системы поиска позволяют дата-исследователю, используя поиск по ключевым фразам, цитатам, датам, ограничивая поиск по определенным ресурсам или форматам данных, получить возможность найти не только необходимые базы данных, но и другую важную информацию по изучаемой теме.

Сегодня многие государства поддержали концепцию открытых данных и свободы распространения официальных данных тех или иных структур. К сожалению, правительственные и научные структуры, в силу отсутствия дополнительных человеческих ресурсов или постановления других приоритетных задач, не всегда имеют возможность контролировать качества загружаемых в открытые базы данных, а также степень адаптации старых, плохо обработанных баз данных, под современное программное обеспечение

компьютеров. Таким образом, исследователь, работающий с базами данных с открытых источников такого рода, часто может столкнутся с нечитабельным материалом, непригодным для дальнейшего компьютерного анализа.

Помимо официальных государственных сервисов, таких как вышеупомянутые порталы открытых данных разных стран как data.gov.ru, data.gov или, допустим, data.gov.uk, исследователь может также использовать и агрегированные международные системы, занимающиеся составлением тематических баз данных. Ими могут являются следующие порталы: Open Data Network, Google Public Data Explorer, The Data Hub, Datamarket, The World Factbook, BuzzData, UNData и множество других. Данные системы открывают доступ к мировой коллекции данных из многообразия разных интернет-источников. Следует учитывать, что данные, попадающие в эти системы, могут поступать не только из официальных источников, но также и от интернет-пользователей: при работе с такими данным исследователю необходимо с особым вниманием проверять загруженные базы данных.

Не стоит забывать и о наиболее классическом и распространенном методе получения информации, котором часто пользуются исследователи: официальном запросе или обращении к экспертам [7]. Многие организации и ведомства не загружают все имеющиеся данные в собственные базы данных, однако по запросу могут предоставить необходимую для исследователя информацию. В таком случае, при обращения к организациям и ведомствам, исследователь данных обязан определить заранее какую именно информацию он ставит цель получить, так как подобная специализированная информация от «экспертов» может быть ограничена в распространении законом. Распространено мнение, что в случаях, если информация может представлять собой государственную или коммерческую тайну, исследователю надо до публикации проконсультироваться с юристом.

В дополнение к вышеописааным классическим способам получения необходимой дата-информации, исследователь может привести поиск по темам и запросам на узкоспециализированных форумах. Сегодня интернет объединяет большое количество исследователь данных, которые, являясь активными пользователями сети, создают специализированные форумы и порталы. Здесь диджитал-исследователи могут делиться опытом, получать рекомендации по поиску и обработке данных, а также запрашивать необходимую информацию у своих коллег. Примерами могут являться специализированные порталы Get The Data, Data Mos, AZSecure-daa, посвященные журналистики данных сабреддита портала Reddit и менее узкоспециализированный сервис Quora. При желании, в целях получение постоянной информации о новых доступных базах данных, дата-исследователь может стать членом интернет-сообществ, связанных с открытыми и большими данными, а также AI-тенденциями в исследовании и обработки таких данных.

В случае, если исследователь не может использовать описанные выше способы, то для получения необходимой информации существуют и иные эффективные методы. Один из таких потенциально возможных, но более сложных методов, является получение API-доступа к данным [8].

Интерфейс прикладного программирования (application programming interface) является набором методов, процедур и функций, предоставляемым сервисом для использования профессиональными разработчиками и сторонним ПО. Так, API классифицируются по видам операционных и аутентификационных систем, а также звуковых или графических интерфейсов.

Возможно использование интерфесов прикладного программирования, таких как: OpenAl, GDI, Motif, Pam Qt, Amiga ROM, Kemel, Cocoa, OS/2 API, SDL, X11 и Zune. Используя интерфейсы прикладного программирования, исследователи могут взаимодействовать с другими программами и получать доступ к непубличным данным социальных медиа, библиотек и сервисов. Безусловно, обработка данных, полученных таким образом, запрещена, так как изначально их стороннее использование третьими лицами нарушает законодательство.

Метод получения данных напрямую с интернет-страницы может быть использован, если для доступа к закрытым данным на сайте не предусмотрен API. Как правило, использование данного метода требует дополнительных технических навыков: более глубокий знаний в программировании, а также базовых знаний CSS и HTML.

Данный способ используется в случае, если информация встроена в страницу и ее невозможно скопировать. Безусловно, исследователь может попробовать провести провести аналогичный сбор информации: тем не менее, если информация распределена по различным страницам или даже отдельным сайтам, то данный сбор информации требует, как правило, большого количества времени. Таким образом, специалисты с навыками программирования используют готовые утилиты или создают собственные программы, автоматизирующие процесс сбора данных.

Так, сегодня в распоряжении исследователя-данных могут находится интсрументы, облегчающие работу по извлечению больших объемов данных с веб-страниц. Например, ими могут являться браузерные расширения, профессиональные веб-сервисы и отдельное программное обеспечение. Например, для поиска в браузере исследователь может использовать «Альтернативный поиск Google», «Контекстный поиск», «Chrome Scraper extension» для Chrome, DownThemAll для Firefox и расширение FireBug доступное сегодня для большинства браузеров: Firefox, Chrome, Internet Explorer и Safari. Интернет-сервис QuickCod являющийся примером сервиса для коллективной разработки программного, необходим для анализа и излвлечения публичных данных. Благодаря QuickCod, исследователю открывается доступ к работе с кодами программ, написанных на языках программирования, таких как как Ruby, Python и PHP.

Процесс поиска и извлечения данных, как и любые другие практические способы и методы работы, имеет свои ограничения. Ими могут являться: неправильно структурированный код веб-страниц, нарушающий алгоритмы поиска, аутентификационные барьеры, запрещающие автоматический доступ, а также блокировка широкого доступа со стороны администрации сервера.

Выше было упомянуто, что исследователю, помимо технических ограничений, никогда не стоит забывать и об ограничения правового рода. Сегодня извлечение данных из веб-страниц-первоисточников не является официальным способом получения информации. Таким образом, данные, полученные в результате таких методов, имеют ограничения как на использование, так и на публикацию. В случае дальнейшего распространения данных, извлечение которых даже могло быть и не запрещено официально, могут возникнуть юридические конфликты.

Так, внезависимости от используемоего метода получения данных, основной бизнес-целью является получение эффективной машиночитаемой базы. Основная задача, которую необходимо выполнить – собрать данные и затем подготовить их к дальнейшей компьютерной обработке. Среди исследователей данных наиболее удобными и популярными форматами для этого являются общепринятый Excel, а также XML, CSV и JSON [9].

Вторичная очистка, обработка и формирование базы данных исследователя.

Следующий этап обработки является упорядоченной последовательностью действий исследователя, направленных на выявление и исправление некорректных элементов баз данных. Цель данного этапа состоит в том, чтобы привести информацию к структурированному и упорядоченному состоянию, что является важным шагом в проведении последующего анализа. Сегодня исллеователи выделяют три предварительных этапа обработки данных: получение входных данных, непосредственный процесс обработки и анализ результатов. Для базовой очистки данных автор может использовать бесплатные инструменты, широко представленные в сети Интернет. Ими являются: Potter’s Wheel ABC, Wranglerm, OpenRefine и другие. Однако наиболее популярным и удобным на сегодняшний день инструментом является именно специализированная программа OpenRefine, являющаяся особым автономным настольным приложением. Программа OpenRefine имеет так называемый «открытый исходный код», что позволяет исследователю очищать и преобразовать данные в другие форматы, а также работать

с таблицами (например, программа может совершать автоматическое исправление ячеек и данных с ошибками), анализировать распределение значений по всему набору данных т.д. (рисунок 2) Исследователь имеет возможность определять критерии фильтрации. Особенным отличием от прочих электронных таблиц является возможность выполнения большинства операций над всеми видимыми строками.

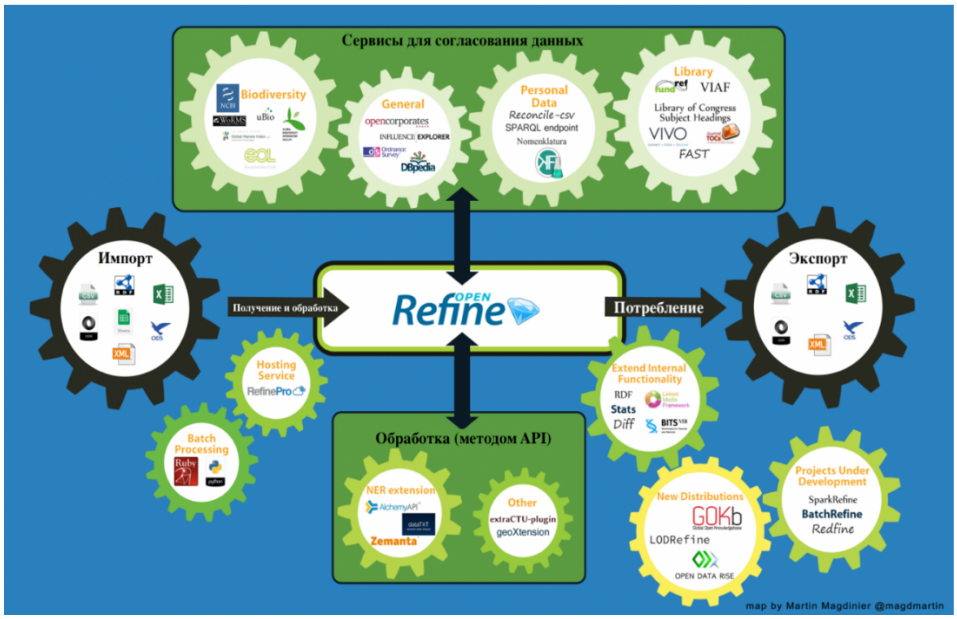

Рисунок 2. Процесс работы с данными в программе OpenRefine, 2019 г. [10]

Для работы с этим приложением исследователю хватит классических навыков работы с таблицами. Тем не менее, следует отметить, что программа OpenRefine позволяет выполнять и профессиональные задачи - данный инструмент повсеместно используют в организациях, чья работа связана с открытыми данными официальных государственных порталов.

Имея определённые навыки программирования (например, владение языками программирования Python и R), исследователь, как и в случае с процессом извлечения данных, его обработки и очистки, может автоматизировать процесс работы и в OpenRefine. Тем не менее, несмотря на то, что данный вариант работы является наиболее сложным, он позволяет использовать более широкий спектр возможностей для быстрой автоматической очистки.

Несмотря на системную необходимость обработки данных, следует исходить и из особенностей организаций международного бизнеса: требования и форматы оформления и структурирования данных могут кардинально отличаться друг от друга. Тем не менее, общепринятых требований и официальных стандартов к оформлению открытых данных на данный момент не существует. Следовательно, каждая организация фомрирует их на свое усмотрение в соответствии с нуждами бизнеса и его целями. Однако, это может приводить к возникновению неточностей и нестыковок в базах данных, что затрудняет выявление в таблице закономерностей между полученными данными.

В целях решения этой задачи, исследователь данных может изучить полученный материал и, устранив очевидные ошибки, провести сортировку данные. При работе с базой, созданной путем слияния других баз, потребуется процедура форматирования, позволяющая привести данные к одному формату. Кроме того, иногда при работе с данными могут быть использованы различные методы кодирования, затрудняющие обработку данных. В данном случае задачу решает создание специализированных словарей с переченем ключевых значений, используемых

в таблицах. В таких словарях описаны все виды кодировок, позволяющих, как правило, избегать вольного трактования в работе с той или иной базой данных.

Таким образом, при глубокой обработке, очистке и формировании рабочей базы данных, менеджмент должен учитывать все перечисленные выше факторы, использовать специализированные программы и постоянно анализировать полученные результаты под цели компании. Именно на этапе формирования логической структуры и происходит формирование базы, а, следовательно, неверная интерпретация данных может привести к ложным выводам и доказательствам, что полностью подрывает ценность использования больших данных в международном бизнесе.

Результаты.

В целях развития международного бизнеса компаний, менеджмент может использовать большие данные как для оптимизации своей деятельности, так и для получения новой уникальной информации, позволяющей принимать решения на основе данных и доказательств.

Основными сервисами и инструментами для работы с данными являются:

- На этапе сбора и первичной обработки: государственные порталы открытых данных data.gov., порталы The Data Hub, Open Data Network, Google Public Data Explorer, Datamarket, а также метод получение доступа к данным через API через такие сервисы как miga ROM, Kemel, Cocoa, OS/2 API, OpenAl, GDI, Motif, Pam Qt, SDL и др.
- На этапе глубокой обработки, отчистки и формирования рабочей базы исследователь данных может прибегнуть к использованию общедоступных в сети Интернет инструментов, таких как Potter's Wheel ABC, Wranglerm, OpenRefine или, имея определённые навыки программирования, модернизировать данные инструменты методом автоматизации процессов и путем определения критериев фильтрации данных.

Заключение.

Аналитика больших данных может предоставить компаниям ценные инсайты в международные рынки, позволяя им выявлять паттерны, прогнозировать спрос и предсказывать тенденции. Однако использование больших данных в международном бизнесе также создает значительные вызовы, связанные с навыками работы с данными, а также их конфиденциальностью, безопасностью и соответствием регулятивным требованиям. Предложенный инструментарий алгоритмизации открытых и big data – данных позволяет международному бизнесу наиболее эффективно получить новую информацию, в конечном счете влияющую на управление компаний на всех этапах: от анализа ЦА для получения уникального конкуретного преимущества и до управления рисками, логистикой, качества продукции и интернационализацией бизнеса.

Список литературы

[1] A Systematic Review of Big Data: Research Approaches and Future Prospects / C. Cobanoglu, A. Terrah, M.-J. Hsu, V.D. Corte, G.D. Gaudio // Journal of Smart Tourism, 2(1), 2022, p. 21–31.

[2] The Role of Big Data in International Business Strategy / M.H. Jensen // Twenty-Seventh European Conference on Information Systems, Stockholm-Uppsala, Sweden, 2019, p. 5-11.

[3] Аксенова О. Н. Журналистика данных: проблемы и перспективы // Научный вестник Воронежского государственного архитектурно- строительного университета. Серия: Социально-гуманитарные науки No.3, 2015 г, с. 41-44.

[4] Using machine learning to create and capture value in the business models of small and medium-sized enterprises / R-C. Climent, D. Haftor, M. W. Staniewsk // International Journal of Information Management, Swansea, United Kingdom, 2023, p.5-19.

[5] Big Data (2023) // The Guardian URL: (https://www.theguardian.com/data) Дата обращения 12.03.2023

[6] Арбатская Е.О. Открытые данные как ресурс региональной журналистики // Вестник Челябинского государственного университета, No: 5 (360) - Челябинск - 2015, г с. 52-58

[7] Василика М.А. Основы теории коммуникации: учебник // М.: Гардарики, 2003 г., с. 610-615

[8] Видовский Л.А., Янаева М.В., Мурлин А.Г, Мурлина В.А., Гвозденко А.А. Анализ возможности использования технологии обработки больших данных в системах для территориально-распределенных комплексов // Научный журнал КубГАУ - Scientific Journal of KubSAU. 2017. No132, с.1-11

[9] Нил К., Шатт Р. Data Science. Инсайдерская информация для новичков. Включая язык R // Издательский дом «Питер», Санкт- Петербург, 2018, с. 273-290

[10] Mapping OpenRefine Ecosystem / M. Magdinier // Open Refine Blog URL: (https://openrefine.org/blog/2015/01/26/Mapping-OpenRefine-ecosystem), 2015. Дата обращения: 15.03.2023.

ALGORITHMIZATION OF OPEN AND BIG DATA FOR INTERNATIONAL BUSINESS DEVELOPMENT

S.A. Zatonkii
Postgraduate student at South Federal University, Head of BODIUM Marketing Department

BODIUM, Marcel LLC, Rostov-on-Don, Russia
Southern Federal University, Rostov-on-Don, Russia
E-mail: stanislav.zatonsky@gmail.com

Abstract. International business can analyze open and big data in order to effectively adapt the information that affects commercial profits. This study describes the benefits of using data for business development. A step-by-step analysis of the algorithm for working with data is conducted. The key stages of data research process are defined: searching and collecting information, basic data cleaning and processing, as well as reducing data to one type of coding. The most popular tools to work with data at all presented stages are identified.

Keywords: international business, big data, data arrays, data research, data algorithmization