Министерство образования Республики Беларусь

Учреждение образования

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

УДК 368.022.15

Тимур Каримович АХМЕТОВ

ИНСТРУМЕНТЫ И МЕТОДЫ ЭКОНОМИЧЕСКОЙ АНАЛИТИКИ НА ОСНОВЕ БОЛЬШИХ ОБЪЕМОВ ДАННЫХ

Специальность - 1-40 80 07 Электронная экономика

Автореферат диссертации

Научный руководитель

доктор.эк. наук, доцент

Татьяна Николаевна БЕЛЯЦКАЯ

Минск, 2021

ВВЕДЕНИЕ

В настоящее время идеей цифровой трансформации охвачен весь мир и во многих странах цифровизация является стратегическим приоритетом развития.

На сегодняшний день более 15 стран мира реализуют национальные программы цифровизации: Дания, Норвегия, Великобритания, Канада, Германия, Саудовская Аравия, Индия, Россия, Китай, Южная Корея, Малайзия, Сингапур, Австралия, Новая Зеландия.

Одними из передовых стран по цифровизации национальных экономик являются Китай, Сингапур, Южная Корея.

Китай в своей программе «Интернет плюс» интегрирует цифровые индустрии с традиционными. Сингапур формирует «Умную экономику», Канада создает ИКТ-хаб в Торонто, драйвером которой становится ИКТ. А Южная Корея в программе «Креативная экономика» ориентируется на развитие человеческого капитала, предпринимательство и распространение достижений ИКТ, а Дания фокусируется на цифровизации госсектора.

Наиболее ярким примером подхода цифровой приватизации является Сингапур. Так, в 2014 году государство инициировало разработку концепции Smart Nation и пригласило бизнес и экспертное сообщество к сотрудничеству для ее уточнения и реализации. Так, одна из ключевых инициатив, определенных изначально, развитие национальной сенсорной сети для построения «умного города». Под каждую из задач государство организовывает тендер для выбора подрядчика на разработку технического решения. Участие в тендере открыто для всех участников, отвечающих требованиям брифинга: таким образом, государство обеспечивает фокус не только на крупный бизнес, но и на привлечение малого и среднего бизнеса. Примечательно, что в 2015-2016 гг. более половины контрактов были подписаны с малым и средним бизнесом.

Государство может обеспечить «цифровой скачок» в стране за счет ускоренного развития конкретных технологий. В таких случаях государство принимает на себя роль инвестора, определяющего ключевые, наиболее перспективные направления финансирования, исходя из оценки долгосрочного возврата на инвестиции, конкурентной позиции, трендов, а также вкладывается в фундаментальные условия успеха, такие как образование и переквалификация кадров.

Цель диссертационного исследования – развитие тенденций цифровизации на постсоветском пространстве, выработка рекомендаций и стратегий для стран, занимающихся цифровой трансформацией и развитием цифровой экономики, возможность применения методов аналитики больших данных для анализа.

Поставленная цель обусловила необходимость решения ряда следующих задач:

1 Изучить на примере пяти постсоветских стран (Эстония, Беларусь, Российская Федерация, Казахстан, Узбекистан) тенденции развития

цифровой экономики в период с 1995 по 2020 год, импорт и экспорт ИКТ товаров и ИКТ сервиса.

2 Опробовать методы кластеризации (метод К-средних) для анализа цифровых рынков и тенденций их развития.

3 Опробовать метод анализа временных рядов для анализа цифровых рынков и тенденций их развития.

4 Выработать рекомендации для дальнейшего развития цифровой экономики.

Объектом исследования является цифровая экономика на постсоветском пространстве.

Предмет исследования – применение методов больших данных для анализа тенденций развития цифрового рынка.

Методология исследования основывается на системном подходе к изучаемым проблемам, комплексном рассмотрении вопросов формирования цифровой экономики. При исследовании использованы следующие научные методы: анализ временных рядов, кластерный анализ, метод К-средних.

Результаты исследования были апробированы на конференции БГУИР. Также данные результаты обсуждались на Научном совете КазНИТУ имени Сатпаева для выработки рекомендаций развития Парка информационных технологий «Алатау».

В первой главе на основе литературных источников исследовано, что на анализе big data строят и развивают проекты, которые направлены на серьезное повышение эффективности процессов маркетинга и продаж, оптимизации производства, логистики, управления рисками, планирования, управления персоналом и другие рабочие процессы различных бизнесов.

В реализации используются как собственные технологии (например, NoSQL базы данных Tarantool), так и другие open source-решения (Apache Hadoop, Apache Spark). Для построения предиктивных математических моделей используются методы машинного обучения (Machine Learning), втом числе собственные разработки компании, например, алгоритм построения моделей Multiclass Look-alike, являющийся развитием метода PU Learning.

Многие компаний все больше стремятся к работе в режиме реального времени. Для достижения высоких темпов движения данных на всех уровнях работы компании данные необходимо собрать и разработать специальные предикторы (системы анализа, позволяющие проецировать данные на возможное поведение пользователя в будущем) и анализировать дальнейшее поведение пользователя. Источниками Big Data в этом случае будут данные, которые передаются от компьютера к компьютеру, социальные медиа, данные о пользователях, транзакционные данные, прочие неструктурированные данные.

Информационно-аналитические системы объединяют, анализируют и хранят как единое целое информацию, извлекаемую как из учетных баз данных организации, так и из внешних источников. Входящие в состав информационно-аналитических систем хранилища данных обеспечивают

преобразование больших объемов сильно детализированных данных в обобщенную выверенную информацию, которая пригодна для принятия обоснованных решений. Для того, чтобы находить в накопленных данных скрытые закономерности и преобразовывать их в знания, пригодные для принятия решений, используются методы и алгоритмы, объединенные общим названием Data Mining или интеллектуальный анализ данных.

Аналитические методы можно разделить на три категории — описательные, предписательные и предиктивные. Многие организации, как частном, так и в государственном секторе, виртуозно оперируют инфографикой — схемами и диаграммами, иллюстрирующими различные показатели: число компаний, принявших участие в государственной программе кредитования научных исследований, их местонахождение, объем выделенных средств и др. Предиктивные методы наиболее развиты в использовании, их применяют во многих областях. Предиктивная аналитика основана на прогнозировании вероятного эффекта. Хотя прогнозы носят скорее субъективный характер, однако, эффективности использования данных в разработке стратегий улучшается.

Отдельные апологеты Big Data радикально говорят о “конце теории” с приходом эры больших данных. Когда доступны все большие массивы данных, предельные издержки по проверке гипотез сокращаются. Благодаря компьютерной революции снизилось и время на их проверку. Зачем строить подлежащую проверке регрессионную модель детерминант экономического роста страны, если можно оценить на компьютере два миллиона ее вариантов.

Также уже рассматриваются большие данные как часть ими разрабатываемой конвергентной информационно-аналитической платформы. Конвергентная аналитическая платформа представляет собой совокупность программно-аппаратных средств, взаимодействующих между собой, которые предназначены для автоматизации процессов сбора и обработки больших данных с использованием вычислительного кластера, облачных технологий и мобильных систем связи. Платформа включает: а) комплекс вычислительных средств центра обработки данных, б) комплекс средств сбора, обработки и загрузки данных в хранилище, облачное хранилище данных, в) прикладные программные комплексы для решения задач интеллектуального анализа и прогноза, г) экспертную подсистему для настройки прогнозных и аналитических моделей, д) систему удаленного доступа, е) систему администрирования информационной безопасности, ж) средства мониторинга и управления функционированием системы.

Термин конвергенция введен в 2002 г. М. Роко и У . Бейнбриджем для определения процесса сближения нано-, био-, информационных, когнитивных и социальных технологий . В I Т-сфере конвергенция связана с развитием информационных и телекоммуникационных технологий. Научно-технологическая конвергенция определяет процесс взаимопроникновения технологий и стирания границ между ними, когда результаты и инновации

появляются в междисциплинарной области знаний. Иногда процесс конвергенции рассматривается в качестве синонима целостного системного подхода, в основе которого лежит принцип интеграции и свойство является эмерджентности, когда новые качества у целостной системы появляются в результате соединения ее частей. Конвергентный подход, по нашему мнению, это результат синергетического взаимодействия и взаимовлияния когнитивных, социальных, информационных, телекоммуникационных, нейробиологических технологий в процессе синтеза инструментария получения новых знаний и инноваций. Приведем пример. Процесс конвергенции технологий и систем стационарной и мобильной телефонной связи привел к тому, что абонентам доступны практически идентичные услуги, а сами системы связи тесно взаимодействуют друг с другом, но это не означает, что они интегрируются.

Во второй главе рассмотрены два метода кластеризации больших данных-метод К-средних и метод DBSCAN.

Кластерный анализ (англ. cluster analysis) - многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы. Задача кластеризации относится к статистической обработке, а также к широкому классу задач обучения без учителя.

Метод k-средних (англ. k-means) — наиболее популярный метод кластеризации. Был изобретён в 1950-х годах математиком Гуго Штейнгаузом и почти одновременно Стюартом Ллойдом. Особую популярность приобрёл после работы Маккуина.

Также еще применяется основанная на плотности пространственная кластеризация для приложений с шумами (англ. Density-based spatial clustering of applications with noise, DBSCAN). DBSCAN-это алгоритм кластеризации данных, который предложили Маритин Эстер, Ганс-Петер Кригель, Ёрг Сандер и Сяовэй Су в 1996. Это алгоритм кластеризации, основанной на плотности — если дан набор точек в некотором пространстве, алгоритм группирует вместе точки, которые тесно расположены (точки со многими близкими соседями), помечая как выбросы точки, которые находятся одиноко в областях с малой плотностью (ближайшие соседи которых лежат далеко). DBSCAN является одним из наиболее часто используемых алгоритмов кластеризации, и наиболее часто упоминается в научной литературе.

Алгоритм DBSCAN требует двух параметров:

1 eps : он определяет окрестность вокруг точки данных, т. е. если расстояние между двумя точками меньше или равно «eps», то они считаются соседями. Если значение eps выбрано слишком маленьким, большая часть данных будет считаться выбросами. Если оно выбрано очень большим, кластеры будут объединены, и большинство точек данных будут находиться

в одних и тех же кластерах. Один из способов найти значение eps основан на графике k-расстояния .

2 MinPts : минимальное количество соседей (точек данных) в радиусе eps. Чем больше набор данных, тем большее значение MinPts должно быть выбрано. Как правило, минимальные значения MinPts могут быть получены из числа измерений D в наборе данных как MinPts >= D+1 . Минимальное значение MinPts должно быть выбрано не менее 3.

Действие алгоритма таково, что он стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров.

По аналогии с методом главных компонент центры кластеров называются также главными точками, а сам метод называется методом главных точек и включается в общую теорию главных объектов, обеспечивающих наилучшую аппроксимацию данных .

Алгоритм представляет собой версию EM-алгоритма, применяемого также для разделения смеси гауссиан. Он разбивает множество элементов векторного пространства на заранее известное число кластеров k.

Основная идея заключается в том, что на каждой итерации перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике.

Алгоритм завершается, когда на какой-то итерации не происходит изменения внутрикластерного расстояния. Это происходит за конечное число итераций, так как количество возможных разбиений конечного множества конечно, а на каждом шаге суммарное квадратичное отклонение V уменьшается, поэтому зацикливание невозможно.

В алгоритмах глубокого обучения метод k-средних иногда применяют не по прямому назначению (классификация разбивкой на кластеры), а для создания так называемых фильтров (ядер свёртки, словарей). Например, для распознавания изображений в алгоритм k-средних подают небольшие случайные кусочки изображений обучающей выборки, допустим, размером 16x16 в виде линейного вектора, каждый элемент которого кодирует яркость своей точки. Количество кластеров k задается большим, например 256. Обученный метод k-средних при определенных условиях вырабатывает при этом центры кластеров (центроиды), которые представляют собой удобные базисы, на которые можно разложить любое входное изображение. Такие "обученные" центроиды в дальнейшем используют в качестве фильтров, например для свёрточной нейронной сети в качестве ядер свёртки или других аналогичных систем машинного зрения. Таким образом осуществляется обучение без учителя при помощи метода k-средних.

В третьей главе была построена математическая модель, где были рассмотрены 5 стран постсоветского пространства (Эстония, Российская Федерация, Беларусь, Узбекистан и Казахстан). Данные были взяты с сайта

Всемирного банка. Был рассмотрен вклад этих стран в процессы цифровой экономики, процент импорта и экспорта хайтек товаров этих стран. Были рассмотрены три временных промежутка 1995-2005, 2005-2015 и 2015-2019. Также при кластеризации этих стран по показателям участия цифровой экономики были выделены как показатели три года-1998, 2008, 2018.

При построении математической модели в программе Python использовались два метода для анализа вклада этих пяти стран в цифровую экономику:

1 Анализ временных рядов.

2 Кластерный анализ.

При анализе временных рядов использовались три временных промежутка: 1995-2005, 2005-2015, 2015-2019.

Таким образом, на защиту выносятся следующие положения:

1 Пример Беларуси, Эстонии и Российской Федерации и их быстрое включение в систему мировой цифровой экономики говорит о том, что научно-техническая база, оставшаяся от СССР, позволила этим странам развить ИКТ-сферу.

2 Пример Казахстана показывает, что достаточно большая доля ИКТ в структуре экспорта и быстрый рост возможен при сочетании двух факторов: иностранных инвестиций и специальных госпрограмм.

3 Пример Узбекистана и взрывной рост его доли в мировой цифровой экономике в 2016 году показывают, что либеральные экономические реформы и открытость мировому рынку дают такие результаты.

4 В целом же, данную программу можно использовать также для кластеризации не только целых стран, но и различных объектов рынка (бизнесов, городов, областей, месторождений).

Результаты исследования были апробированы на 57-ой научной конференции аспирантов, магистрантов и студентов БГУИР «Проблемы экономики и информационных технологий». Также данные результаты обсуждались на Научном совете КазНИТУ имени Сатпаева для выработки рекомендаций развития Парка информационных технологий «Алатау».